Examining the Potential for Enhancing User Experience in Exploratory Search using Conversational Agent Support

Abhishek Kaushik

B.Tech. (Hons.), and M.Sc. in Computer Engineering and Information Technology

Supervisor: Prof. Dr. Gareth J. F. Jones

A thesis presented for the degree of Doctor of Philosophy

School of Computing

Dublin City University

May 2021

Dedication

This dissertation is dedicated to my dad and granny (Krishna). However, they are not present in this immortal world, but I have always felt their presence to motivate me.

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: 

(Candidate) ID No.: 17211733

Date: 7th May, 2021

Acknowledgements

First and foremost, I want to thank the Mother of Universe "Maa Shakti" and Almighty "Prabhu Shri Ram".

Most importantly, I would also like to thank my supervisor Gareth J. F. Jones. The synergies and conversation with him helped me improve, discover and develop myself as a better researcher and person, leading to my overall development. Additionally, I am thankful to the ADAPT centre and my colleagues, which provided an excellent opportunity to pursue my PhD. This work is supported by Science Foundation Ireland as part of the ADAPT Centre at DCU (Grant No.13/RC/2106).

I must thank my wife for her support in my thick and thin. Finally, I would like to express my gratitude towards my parents (Mr. Krishan Dutt Kaushik and Mrs. Durga Sharma), brother, sister, and extended family, who have constantly been encouraging and supportive.

A quote from *Bhagavad Gita* has always motivated and inspired me:

> *This knowledge is the king of education, the most secret of all secrets. It is the purest knowledge, and because it gives direct perception of the self by realization, it is the perfection of religion. It is everlasting, and it is joyfully performed.*
> *Adhyay -9 – Shloka -2*

Contents

List of Figures

List of Tables

Examining the Potential for Enhancing User Experience in Exploratory Search using Conversational Agent Support

Abhishek Kaushik

Abstract

Traditional information retrieval applications require users to develop a well-formed query describing their information need. This places a cognitive burden on the searcher who must expend effort in attempting to select words that succinctly describe their information need, with the implicit assumption that they know enough to describe their target. A more natural mode of human enquiry is via dialogues which enable incremental development of topical understanding and corresponding effective queries. The conversational search seeks to enable the next generation of more natural and efficient search applications that should be easy to use while less cognitively demanding on searchers. Since conversation is a natural means of human information inquiry, framing the information retrieval process within dialogue is hypothesized to make the search process more natural for the user in terms of query entry, interaction to locate relevant content, and engaging with system output.

This PhD research project seeks to make progress toward realizing the vision of conversational search systems. In this project, we investigate the opportunities to integrate the exploratory search process within a conversational setting. We propose a conceptual framework for dialogue-based exploratory search applications combining a standard search tool with an interactive agent in an integrated user interface. Additionally, we introduce an implicit evaluation framework for conversational search in exploratory search setting, including multiple dimensions: search experience, knowledge gain, software usability, cognitive load and user experience, based on studies of conversational systems and information retrieval.

We examine the behaviour of current conversational assistants to support complex information seeking tasks and propose and evaluate extensions to improve their effectiveness. Using implicit evaluation we examine the user-experience using our conversational search interface using a conversational agent taking a rule-based approach and a machine learning approach. Our findings show that users respond intuitively and positively to the introduction of conversational support into their interactive search experience.

Chapter 1

Introduction

Online search tools have become a ubiquitous technology enabling users to gain immediate access to content to support their information needs at home and at work. However, current established search tools offer only a basic interactive interface. Using these tools often places a demanding load on the user, in which they are required to create and enter a keyword based query describing the information that they need in its entirety in a single-shot search, and then to read and navigate among a set of retrieved items which may contain the information that they need. These interfaces generally offer little support to the user and can be cognitively demanding on the user, particularly when undertaking a complex search or examining an unfamiliar topical area. Users looking for information may lack sufficient knowledge to create an effective query on their topic of interest, or may need to iterate their way towards an effective query through multiple search cycles query, as they learn about the subject using less informed queries, or they may even abandon their attempt to locate relevant information if they cannot form a query able to obtain useful retrieved items.

The objective of the research described in this thesis is to examine moving beyond this user driven basic mode of interaction to the use of dialogue processes between the user and the search tool to support and facilitate the search process. Adopting the metaphor of an advisory librarian, this research adopts the concept of an artificial agent working with the user to support the satisfaction of their information

need. Rather than seeking to use of an individual query to retrieve the necessary information in a single shot, we propose to adopt conversational interaction in the manner of a searcher consulting a librarian for advice. The assistant agent will observe and act on information entered by the user, including query words and details of retrieved information which the user indicates to be relevant or useful to their interactions, to seek to progress in the search task. Effective conversational interaction with search tools has the potential to reduce the cognitive cost of search for the user, while providing support to enable them to satisfy their information needs more successfully by using a natural means of interaction with the search tool.

The key aspects of this research project are the exploration of the potential for conversational agents to support complex exploratory information seeking processes, and to find ways to reduce user the cognitive load while searching. The other dimension we investigated in this PhD project is knowledge expansion while using CS agents and its effects on cognitive load and cognitive engagement. Cognitive load is defined as stress, pressure and mental workload during the search process and cognitive engagement refers to the cumulative efforts by the user in searching, reading and satisfying the information needs without any cognitive load. Cognitive load can demotivate users or increase their frustration while searching, which could be a vital factor for the user in not completing their search task. In contrast high, user engagement can allow the user to gain more knowledge and extend the search process as per the user's information needs. Both of these cognitive aspects depend on multiple factors such as the topic of search, time spent during the search, background knowledge of the user, etc. It is very important to understand that cognitive factors are critical points to study in a dialogue-based information seeking model. CS is more exploratory as it includes the concepts of user revelement, system revelement, mixed-initiative, etc. These concepts are discussed in detail in the study conducted by Radlinski et al. [129]. In this PhD, we investigate the potential of the conversational agents in exploratory search to support complex information seeking and to promote knowledge expansion.

Our objective is to make search more accessible to people in general via a more natural and supportive process of engagement and, we hope, lead to a more informed society.

1.1 Overview of Information Retrieval and Conventional Search Interaction

In order to understand the potential for the conversational interaction to support the search process, we begin with an overview of Information Retrieval (IR). IR is concerned with the identification of information within archives with the potential to satisfy user information needs. In this section we introduce the concepts of information needs and the features of conventional search engines and interaction with them.

1.1.1 Information Need

A user approaches an IR system with a deficiency information, generally referred to as *information need* [83]. The information need can be specified with respect to two dimensions: task [106] and cognitive scenario [19, 32]. A summary of the relevant features of these dimensions is as follows:

1.1.1.1 Search Task Type

There are a number of different types of search task with which a user can approach a search engine [90, 142].

1. *Known Item*: Search for a previously seen relevant item.
2. *Known Subjects*: Open ended need without a known relevant item on a subject with which the searcher is familiar.
3. *Interpretation*: Learning oriented task where user interprets facts taken from retrieved relevant items, often involving inference of information.

4. *Exploratory*: Learning oriented search where the user explores a topic without a specific focused information need. For this PhD Project, we focus on exploratory search.

1.1.1.2 Cognitive Scenario

There are three standard types of information needs: (a) verificative information need; (b) conscious topical information need; and (c) muddled topical information need [19, 32].

1. *Verificative*: The user is seeking verification of information which they believe will address their information need. This type of information need is straightforward and the user clearly understands their information target.

2. *Conscious*: The user is seeking information about a topic of which they are knowledgeable.

3. *Muddled*: The user is seeking information about an unknown topic with an information need which is consequentially ill-defined.

1.1.2 Functional Overview

The operation of a conventional IR system is shown in Figure 1.1 [4, 35]. The user first enters a text query describing their information need. The IR system then identifies and returns a list of potentially relevant items which may be able to address the user's information need, ranked in order of their estimated likelihood of relevance [152]. The user can then select individual retrieved items to investigate further. The IR system can be divided into two key elements:

1. IR System: The system in the search process seeks potentially relevant items from an indexed collection based on a retrieval model.

2. User : This is concerned with the contribution of the user or searcher to the search process. This includes query formulation to describe the information

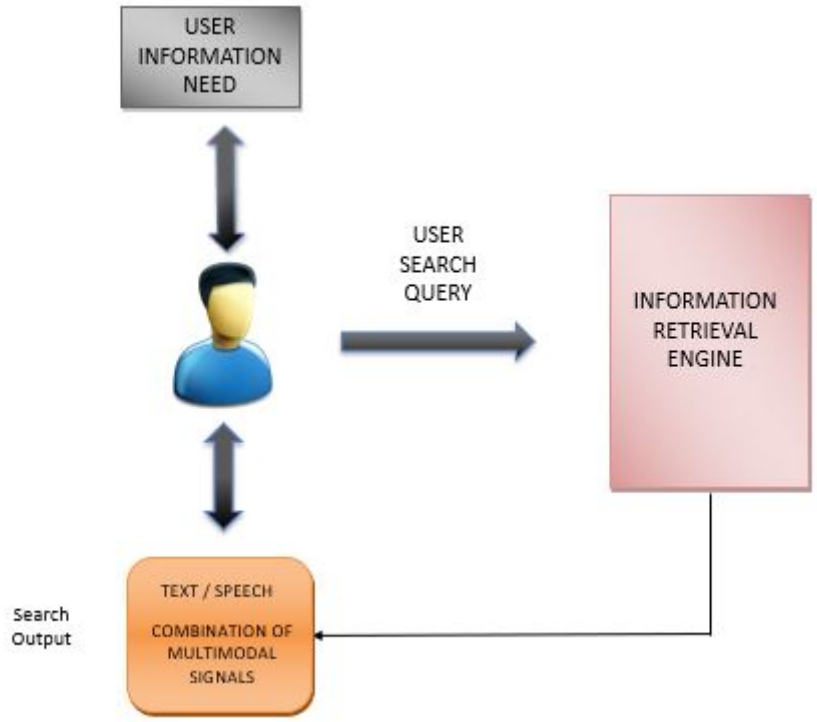


Figure 1.1: Traditional information retrieval system.

need and feedback indicating the usefulness of the search results, which can be used to enhance the effectiveness of subsequent searches. It has been observed that users are often not interested in providing feedback on the relevance of their current search results [83]. Therefore, indirect methods are often adopted to estimate relevance of retrieved items to the user [70].

1.2 Limitations of Current IR Systems

Often the user's lack of knowledge of the subject of their information need means that when using a standard IR system of the sort introduced in Section 1.1.2 the searcher is often unable to fully describe their information need. The query that they enter may be ambiguous or its vocabulary may fail to match well with the content of relevant items composed by more knowledgeable writers. These situations can result in reducing the rank of the relevant items or failure to retrieve them at all. If any of these situations arise while seeking information from a human intermediary, the natural response would be to enter into a dialogue to resolve the problem [152].

In the following subsections we examine the challenges of search as carried out using current search systems in more detail [46, 41, 25, 34, 21, 36].

1. **The user must completely describe their information need in a single query.**

   The user must include all necessary information in a single query to provide a sufficient description of their information need in order for an IR System to identify potentially relevant documents. This can be a very challenging task if the user has little or no background knowledge about the search topic. The user may not clearly understand their information need, or not know or be able to correctly use the vocabulary to describe it. A poorly formed query can result in failure to retrieve relevant results leading to user frustration [10, 63].

2. **High cognitive load on the user in forming a query.**

   The process of creating an effective single shot query can place a significant cognitive load on the user. This arises since the user needs to consider carefully their information need and to decide on a suitable way to describe it [46, 41, 36].

3. **The IR system is asked to return all relevant content available in the collection searched in a single pass.**

   Identifying relevant or useful results in one single shot search operation is often challenging for the search engine itself since the query may not properly or sufficiently describe the information need. As such it may not be possible for the search system to identify relevant context or to differentiate it from non relevant content [40, 47].

4. **Query mismatching from collections.**

   Even a well articulated query may suffer from word mismatch with some or all of the relevant documents. That is, the authors of relevant documents and the query may use different words to describe the same thing. This may lead to relevant documents being returned at lower rank, or not returned at all [34].

5. **The user must inspect returned content to identity relevant information.**

   The user must attempt to identity relevant returned items based on snippets of their content in the ranked results list. This relies on the snippets being sufficiently well formed to represent the documents with respect to the information needed to be represented in the query, in order for the user to be able to identify relevant items. The user must then engage with the full content of retrieved documents to find the information necessary to satisfy their information need. This can be challenging in the case of the complex queries or documents, or for any query that is poorly formed or where the user's knowledge of the subject is poor. Inspection of multiple snippets (and potentially multiple documents) can be time consuming and frustrating for the searcher, again potentially with high cognitive load [21, 25].

   Having outlined conventional IR system, in the next sections we introduce conversational engagement with information systems and introduce our vision of conversational search.

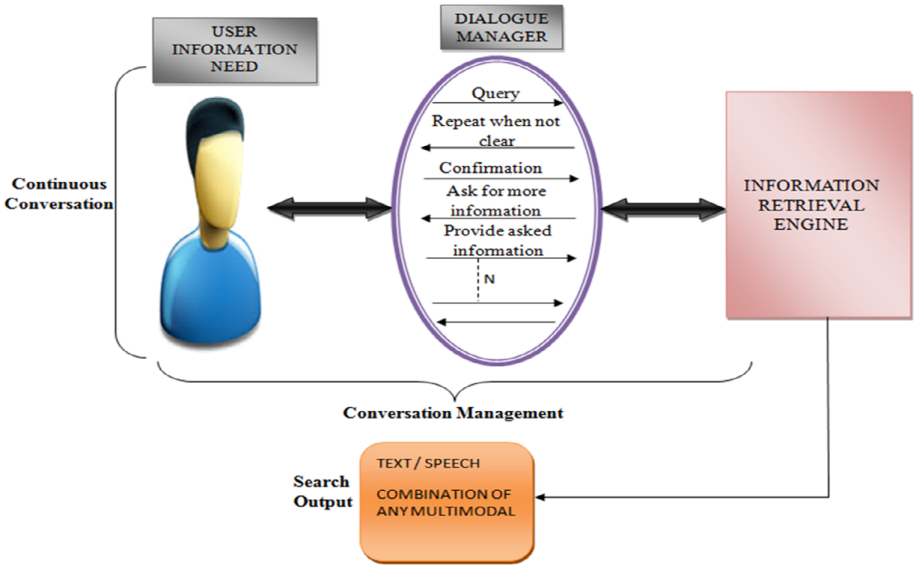


Figure 1.2: Concept of dialogue based multimodal IR.

1.3 Conversational Engagement with Information Systems

Conversational interfaces mediated by automated agents offer the ability to provide access to information services much more economically than using human agents. Advances in natural language processing and machine learning techniques have enabled the development of task-oriented conversational agents to interact with users via text or speech based dialogues [55]. These applications respond to customers requests either with definite answers or by engaging support from human assistants. These advances in conversational agents allow us to examine their incorporation in search applications. In this section we introduce the key issues relating to the use of conversational agents.

1.3.1 Challenges

The development of effective conversational systems poses a number of challenges.

1.3.1.1 Design

Design challenges of conversational agents relate mainly to interacting with them via natural language and speech recognition technologies, and determining task related actions by the agent. The major challenges fall into three categories: interpretation, determination and performance.

- Interpretation: The natural language input needs to be interpreted.
- Determination: The appropriate action in response to the user input must be determined.
- Performance: The agent must successfully perform the selected operations

1.3.1.2 Usability

Acceptance of agent technologies by users is determined by the their usability. Usability challenges agent technologies are associated with behavioural, design, linguistic and logical aspects of the agent system [148]. These challenges include:

1. Behavioural: Understanding user engagement enables identification of the user's of expectations of the agent and to respond appropriately [78].

2. Design: Design of conversational agents can be classified into two sub-classes: a) types, b) interface.

   (a) Types: The two most common approaches to agent implementation: Rule Based and Artificial Intelligence Based [165].

      i. Rule Based: This is a simple approach. The agent uses predefined rules to guide its behaviour.

      ii. AI based: The agents actions are primarily based on machine learning from observed interactions which the agent is required to support.

   (b) Interface: The user design of the interface is a critical aspect of a conversational agent [113, 148]. Conversational systems should focus on the specific requirements related to the way the information is input and presented in the interface. Additionally, speed of response, handling of human errors and etc. are crucial aspects in terms of system effectiveness related to the user interface.

3. Logical Flow: The logical flow of the agent is crucial to enable conversational agents to gain acceptance by users [72, 148, 184]. For example, users expect a conversational system to remember the information exchanged so far.

4. Linguistic: Natural language communication is a major challenge for any information system. It can be divided into three linguistic fundamentals.

   - syntax the structure of the language,

- semantics relating to the meaning of the words,
- pragmatics relating to the context of the content [182].

1.4 A Vision of Conversational Search

Natural human-to-human enquiry to satisfy information needs is typically based on conversational engagement. The vision of conversational search seeks to incorporate this process of conversation into the search process. The concept of conversational search is beginning to attract broad interest in the IR community [126], and has generally been defined as seeking to assist users in search with multi-model dialogue interaction [129]. A conversational search system should allow the searcher to make their query in natural language and support their engagement with the search process. The system could for example ask for clarification of a query to revise it and improve the output of the retrieval stage. A conversational search system might also allow the user to give feedback about the search results, and to provide suggestions for a subsequent search operation.

Conversational search presents opportunities to support users in their search activities to improve the effectiveness and efficiency of search while reducing cognitive load. Since conversation is a natural means of human information inquiry, framing the IR process within a dialogue is expected to make the search process more natural for the user, in terms of query entry, interaction to locate relevant content, and engaging with the system output.

1.5 Research Questions Addressed in this Thesis

Our review of the relevant state of the art research in conversational search, leads us to the following research questions for this PhD research: The research questions investigate user search behaviour, user interaction behaviour with conversational agents, challenges and possible areas to support the user in conversational search, investigating and developing dialogue strategies and the evaluation of conversational

search. In this thesis we address the following research questions:

1.5.1 RQ1: Based on analysis of user search behaviour with an existing search system, what are the challenges and opportunities for conversational search paradigms in exploratory search settings?

RQ 1 focuses on current state-of-the-art search in exploratory search setting. Our first investigations study user search behaviour using a current web search engine. Chapter 5 includes the investigation of this RQ. In these studies we examine the following sub questions:

1. What are the current search behaviours of the user?
2. What are the possibilities to support current user search behaviours in a conversational setting?

1.5.2 RQ2: Based on analysis of user search behaviour with existing conversational applications, what are the challenges and opportunities for conversational search paradigms in exploratory search settings?

RQ 2 focuses on current state-of-the-art conversational systems in exploratory search settings. Our second investigations study user search behaviour using a Smart Speakers. Chapter 6 includes the investigation of this RQ. In these studies we examine the following sub questions:

1. What are the opportunities and challenges for incorporating interactive search capabilities into a current conversational system?
2. What are the characteristics which prevent existing conversational systems from becoming effective tools for complex information seeking?

3. What are the main user expectations of conversational search systems?

1.5.3 RQ3: How might dialogues be embedded in search settings for improved satisfaction of user information needs?

Chapter 6 and Chapter 7 covers the investigation of this RQ. This RQ is divided into the following sub questions.

1. Dialogue strategy to support content engagement

   (a) Can we develop dialogue strategies which engage users in information seeking activities?

   (b) What factors affect dialogue strategies for user information seeking?

   (c) How is the user's search gain in topical knowledge affected by the use of dialogue strategies in the search process?

1.5.4 RQ4: Can enhanced content engagement support improve satisfaction of user information needs?

This RQ is divided into the following sub question. Chapter 7 and Chapter 8 covers the investigation of this RQ.

- Can highlighting potentially relevant content of important sentences within long documents help in the search process by reducing the work of the user in examining documents to satisfy their information needs?

1.5.5 RQ5: Can we demonstrate improved search effectiveness by using a conversational search tool in exploratory search task?

Chapter 7 and Chapter 8 covers the investigation of this RQ. This RQ is divided into the following sub question.

1. How does a conversational search interface compare to a comparable conventional search system and why?

1.6 Thesis Structure

Following the introduction to the topic of conversational search in this chapter, the remainder of this thesis is organized as follows:

Chapter 2 introduces topics relating to conversational search including search models, interactive information retrieval and evaluation of information retrieval and interactive information retrieval (IIR).

Chapter 3 reviews existing work in conversational agents and conversational search. This overviews the approaches taken to conversational search to date including information seeking models in dialogue settings, highlights the challenges of conversational agents, and also discusses current work on evaluation methods for conversational agents and conversational search.

Chapter 4 introduces our experimental methodology, experimental setup and implicit evaluation of conversational search interface (IECSI) framework. This is based on five different dimensions: knowledge expansion, search as learning, user interactive experience, cognitive load (Nasa task load) and software usability (IBM).

Chapter 5 investigates the potential for conversational search in current search systems. This investigation seeks to identify potential opportunities for conversational engagement to support the searcher to satisfy their information need. This chapter describe a user study conducted to investigate to user search behaviour in

a current search setting. The chapter concludes by outlining the opportunities for conversation search in a current search setting.

Chapter 6 focuses on dialogue strategies to support exploratory search processes. This section also examines the search functionality of a state-of-the-art conversational (smart speaker) system, and introduces a novel prototype extension to the existing commercial application.

Chapter 7 examines the requirements for conversational search interface following a requirements analysis. A multi-view conversational interface based on a rule-based agent is introduced and evaluated with a user study.

Chapter 8 investigates a machine learning agent approach for our conversational search interface prototype. The system is again evaluated by a user study.

Chapter 9 presents a summary of our work and draws conclusions of this study. We summarise our findings and the contributions of our work, identify open challenges in conversational search that remain to be explored in future research.

Appendix A includes retrieval models and evaluation techniques of IR and IIR.

Appendix B includes questionnaires and interview questions used in study discussed in Chapter 4, 5, 7, 8.

Appendix C presents a list of publications from our investigations carried out as part of this thesis and questionnaire.

Chapter 2

Overview of Conventional Information Retrieval

2.1 Overview

This chapter provides the overview an current information retrieval (IR) methods, including retrieval models and query reformulation. The chapter also introduces information-seeking models and interactive information retrieval (IIR), which consider user engagement, learning and knowledge, the challenges of measuring learning and knowledge and methodologies to measure learning. The chapter also addresses overviews evaluation methods for IR and IIR.

2.2 Information Retrieval

Chapter 1 introduced the workflow of a conventional IR systems. This takes in a user search query expressed in free text and enumerates the likely relevance to the user's information need of the items in the available document archive.

2.2.1 The Information Retrieval Process

Early work on IR focused on Boolean models based on identifying documents which precisely match the query [73, 83]. Queries to Boolean systems can use complex

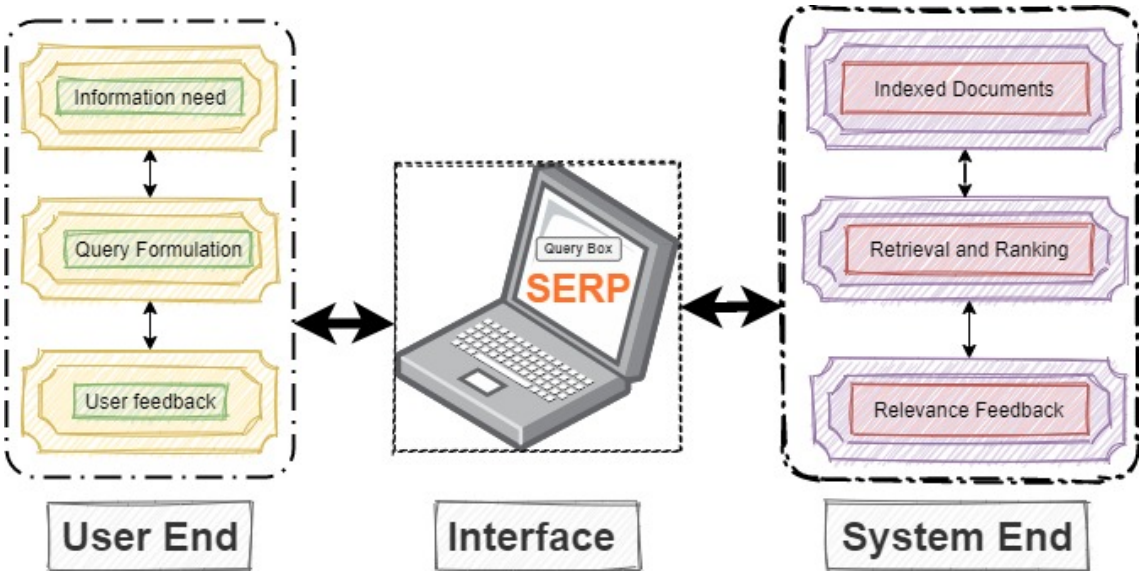


Figure 2.1: Illustrations of information retrieval process.

structures and the output has no ranking. Of more relevance to this PhD research are best-match ranked retrieval methods which seek to compute a statistical relevance coefficient between a query and each document.

Early work on ranked retrieval focused on the vector space model which represents the query and documents as vectors and uses vector similarity as the measure of potential relevance [83]. This was largely superseded by the probabilistic IR model, in particular the popular BM25 model. The probabilistic model seeks to directly compute a likelihood of relevance of each document [79]. This was followed by the language modelling approach to IR, where documents or queries are treated as statistical language models and the documents are ranked based on their relationship to the query in terms of the documents or query language model generating the query on the document respectively [83]. Other ranked IR models have been proposed, but they are less popular than the models outlined here. While the underlying models are very different, all of these approaches use the same single-shot query framework. The effectiveness of these IR models can be improved using methods such as relevance feedback [83], but this does not address the challenges of single-shot search identified earlier in Chapter 1.

As shown in Figure 2.1, the information-seeking process have three major components: user end, interface and system end. Each component is connected by the

search sub-processes such as query formulation, user feedback, relevance feedback, retrieval and ranking. These components are explained briefly:

- User End: This component is responsible for allowing the user to express their information need as an effective query in the due process supported by the sub-processes known as query formulation by user feedback based on pre-existing knowledge or knowledge gain. This component is a critical component where a non-effective or misleading query can result in non-relevant search results, which may increase the user's cognitive load and frustrations.

- Interface: The second components are responsible for the presentation of the search results and having a query box. The presentation mode and style can ease the user seeking the information required to satisfy the required information. Poor presentation style can lead to user frustration and can result in complexity in the information seeking process.

- System End: The third component deals with ranking and retrieval models, which are responsible for the algorithmic model to seek relevant documents from the document collection, known as an index, and rank them appropriately. As outlined earlier, there are different retrieval models which follow different documents ranking methods.

All three components and their related processes are explained in detail later section in this chapter. In this PhD, the prime focus of the work is on two components: user end and interface. This research motivation for this PhD research is to understand user expectations and behaviour in a conversational search setting, which could help in user feedback and query formulation. Another aspect is exploring the possibility of a new interface that allows the user to search in a conversational search setting conveniently. The third component, retrieval and ranking (discussed in Appendix A.1), is out of the scope of this PhD.

2.2.2 Query Formulation

A key element of effective IR is the query issued by the user to express their information need. Often, this query does not completely represent the information need in an exploratory search setting, since the user may not know the correct terminology this information need or does not know how to describe their information need correctly [17]. Attempting to construct an effective query can lead to high cognitive load and frustration to the user when a poor query fails to retrieve relevant content [46, 166].

Query formulation is a complex process including many of difficulties for users in exploratory search settings [58, 80, 110]:

- User lacks understanding of information seeking task: The users may not understand the search task and what they require to solve it. This is a pervasive problem with users who are new to the search task or search system.

- Difficulty in expressing information need into correct keywords: It is challenging for the users to find keywords that are specific in expressing their information need. Also, they are required to make sure that these keywords appear in the documents in the search index. Generally, this case arises when users have no background knowledge about the search topic or are looking for a specific goal-oriented answer. Unsuitable keywords concerning the search task will often provide non-relevant results.

- Trade off between long query and short query [62]: Current search systems are suitable for shorter queries that may not be sufficient to describe the information need. In contrast, if users explain their information need in longer keywords, risk of a misleading or vague queries may arise [76]. So users need to make a balance while querying formulation that it should not be too long or too short [71].

Query reformulation methods are designed to enable users to construct more effective queries. The most popular approach to query reformulation is relevance

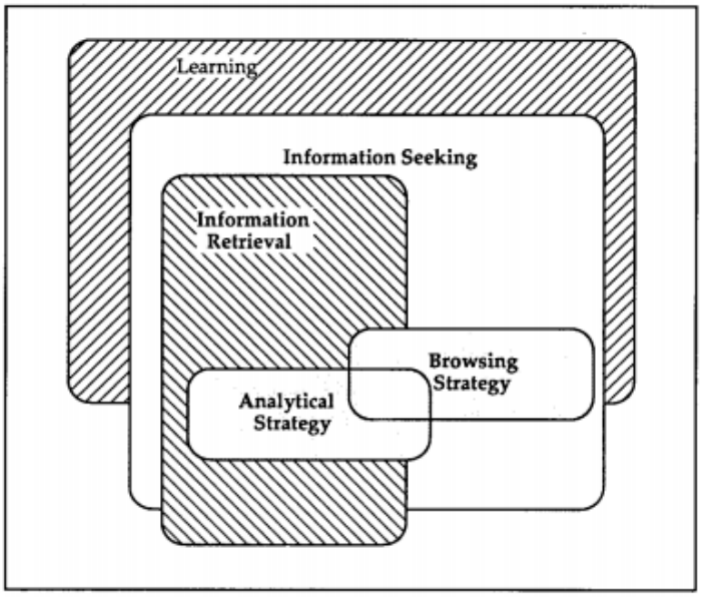


Figure 2.2: Different search processes and their relationships [29].

feedback in which the query is modified based on the results of search using the current query. Retrieval items are labelled relevant or non relevant, and their information is used to determines the potential utility of items from these documents to improve the query for a subsequent retrieval operation.

Relevance information can be entered directly by users, informed based on user activities or assumed based on document rank. More sophisticated approaches to query reformulation are used in commercial search engines to automatically rewrite queries to improve query document matching.

2.3 Overview of Information Seeking

While retrieval models give an algorithmic method for ranking documents in response to a query, this is merely the external component of a complex cognitive process. Search is "a process, in which humans purposefully engage in order to change their state of knowledge" [29]. Figure 2.2 shows the relationship between the sub processes of human information seeking. This illustrates how IR can support a user in satisfying their information need. Significant research effort has been expended in developing cognitive models of the search process in terms of query

creation and user learning in response to retrieved content. Early contributions to attempt to understand and model the information need, and activities of searchers are modelled in Belkin's proposal of the Anomalous States of Knowledge (ASK) model [10]. This highlights the difficulties of users in specifying the details of their information needs for engagement with IR systems, see Section 2.2.2.

Various extensions to the ASK model taking alternative or complementary approaches to modelling information seeking have been developed since then, e.g. Kuhlthau's Information Search Process (ISP) model [18] which captures the searcher's constructive activity of examining retrieved information in order to extend their state of knowledge. In the ISP model, information seeking is viewed as an active process in which the user attempts to find new information and integrate it with what they already know. Essentially this is a personal activity of sense making as the user understands and interprets the information. The process concludes when the user either satisfies their information need or develops new understanding leading to a reformatted query or possibly abandons the search task.

Vakkari's [121] learning model of search describes how searchers develop, refine and assimilate knowledge gained within their search activities. Observing the behaviour of searchers and creating a conversational agent to work as a collaborative assistant to support them in satisfying their information needs could be based on freestanding rule-based or statistical models. However, a more complete, extensible solution is likely to be achieved if their activities are placed within an information seeking model around which the competence of an information agent can be modelled. In the study discussed in Chapter 4, about user search behaviour [152], we focus on Vakkari's analysis of *Search as Learning* model [121].

In this model search is viewed as a process which incorporate learning, in which engagement with retrieved information leads to revisions in the searcher's personal knowledge structures relating to the topic under investigation. The form of restructuring, modification or elaboration of the user's knowledge as the search progresses depends on their existing knowledge of this or related subjects. Specifically Vakkari

identifies three forms of change in the searcher's knowledge structure associated with a search task:

- **Restructuring**: Occurs when the searcher's knowledge of the subject is currently vague and lacking structure. In this situation the searcher is likely to examine multiple sources of information and potentially have difficulty identifying relevant information without extended investigation of many sources.
- **Tuning**: During which the searcher stabilizes and extends their knowledge of the topic. They may examine less sources than in the case of restructuring and are likely to have sufficient knowledge to enable them to select them more discriminatively and on average to examine them in more detail.
- **Assimilation**: Where the searcher combines new knowledge from examined sources with their existing knowledge structures.

The user's activities of learning about a topic during exploratory search typically involve examination of multiple sources to learn about the topic. Exploratory search entails addressing a need for information that goes beyond merely answering a specific query. In this exploratory setting, the searcher progresses through the search as they learn about aspects of the topic of interest [68]. It is unlikely that there will be a smooth transition between these phases, since the searcher is likely to have different levels of knowledge relating to different aspects of the overall knowledge space needed to resolve their information need. This is a complex cognitive process, but one which needs to be properly considered in the design of a conversational search agents designed to support the user in addressing information need of this type. In order to understand how searchers address information needs with differing levels of knowledge of the topic of their search, and to enable us to explore how engagement with conversational agent can support the search process, we need to study searchers interactions with an existing search engine. The information seeking steps discussed above can potentially contribute to the development of dialogue strategies for conversational search.

2.4 Interactive Information Retrieval

User interaction and experience are important areas to explore for evaluation of the IR system by users. As explained by Belkin [81], the main purpose of an IR system is not to get a high statistical score of precision or recall, but rather to help users to satisfy their information need. Therefore, user search logs and information seeking behaviour become a major area of focus in interactive information retrieval (IIR). The IIR encapsulates multiple aspects of computing including human-computer interaction, traditional IR, information science and psychology. The core focus of IIR is user experience while undertaking search tasks, which also includes changes in user behaviour with respect to the interactions in cognitive dimensions [75]. Designing experiments examining IIR includes consideration of user engagement and user learning. In this section, we overview these topics which will be important the development of conversational search.

2.4.1 User Engagement

The quality of the "user experience" within an interaction relates to user engagement. Positive and negative experiences of the user are highly co-related with user interaction with the system, refered to as "user engagement" [85]. Multiple metrics have been used for measuring and evaluating the experience. These metrics have generally been divided into two categories subjective and objective [67].

1. Subjective: This measure generally focuses on opinion based of user feedback such as answering open-ended questions in a structured or semi-structured interview or questionnaire. This provides an opportunity for the user to self-report their experience with an interactive system.

2. Objective: This measure focuses on user feedback such as eye or mouse (cursor) movement of the user, which provides an indication of the user's actual attention and actions in interacting with a system. This also includes the time taken to complete task, no. of actions in completing task, no. of tasks

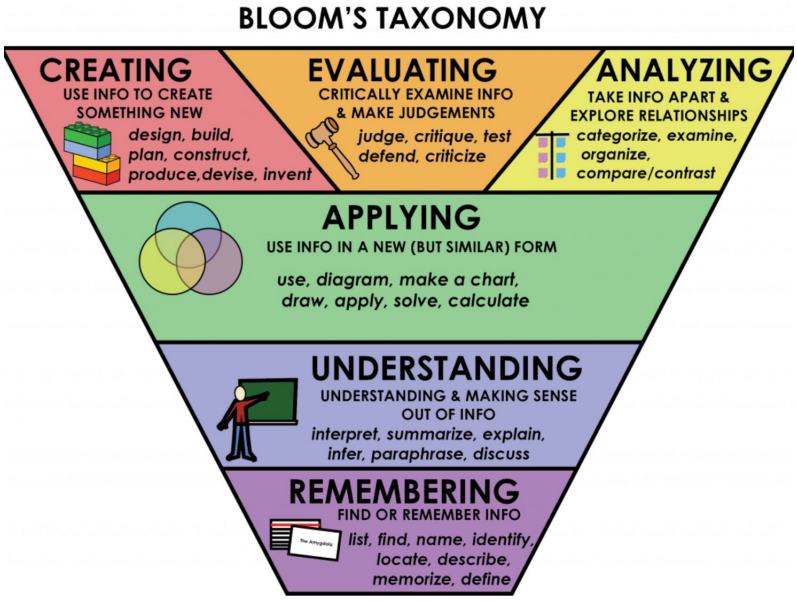


Figure 2.3: Krathwohl's extended version of Bloom's Taxonomy [188, 48].

successfully completed etc.

2.4.2 User Learning and Knowledge Expansion

The process of search and engaging with relevant content produces learning and knowledge expansion by the user. In recent years, measuring the learning and knowledge expansion through search has become a very active area of research. A number of meetings and workshops have laid down the foundation for "searching as learning" [89, 100, 109, 121]. These forums have examined the scope of searching as learning among researchers interested in this topic. In this section, we overview current work on searching as learning and associated knowledge expansion.

2.4.2.1 Overview

Learning is a complex process which combines user experience and their reaction to situations. Learning capabilities and the procedures of various human to human interactions are the primary mode of learning. These interactions include listening, discussing, observing, reading etc.

In this PhD research, we consider the use of Bloom's taxonomy measuring learning Bloom's taxonomy [1] is an hierarchical framework which classifies learning ob-

jectives (cognitive, affective and sensory domains) based on levels of complexity and specificity. Anderson et al. [45] modified and further developed this framework by putting more emphasis on meta cognition, procedural and declarative knowledge. Following Anderson et al. taxonomy, Krathwohl [188, 48] created a framework of six levels with respect to complexity and cognitive dimensions as shown in the Figure 2.3. This revised Bloom's taxonomy and its six dimensions have been used for designing user search experiments which include the consideration of the complexity of the search task and for designing pre-search and post-search questionnaires [74, 104].

In the education sector, schools and colleges have established a standard platform for measuring knowledge and learning through assessments, such as formative, diagnostic, summative [64]. These mechanisms provide the opportunity to regularly measure the progress and development of students [43, 52, 88, 116]. However, measuring learning in an online setting poses challenges for several reasons. These include the short duration of interactions and limited modality options (only speech and only text) for knowledge presentation. The main challenges of measuring learning and knowledge expansion are explained in summary in the next section.

2.4.2.2 Encountered Challenges in Measuring Learning and Knowledge Expansion

The challenges can be classified into four categories:

1. Complexity of diverse learning: There are multiple diverse objectives to learning as identified in Bloom's taxonomy. Knowledge, can be transferred, transformed and created. In addition, there are two broad categories of knowledge: a) declarative knowledge b) procedural knowledge [92].

   (a) Declarative Knowledge: This relates to a collection of facts about someone or something. As per the learning process, declarative learning is the first phase knowledge acquisition, i.e learning of facts.

(b) Procedural Knowledge: This relates to knowledge of the process of doing something. For an example, the skills of a production manager. As per the learning process, procedural learning is the second phase of knowledge acquisition where one starts to acquire skills (procedure) after learning the elementary knowledge (declarative knowledge) about something.

Preparing and planning experiments which measure learning is a complex process. However, certain studies have been developed which enable researchers to focus on measuring acquisition of specific types of knowledge [99, 135].

2. Understanding user expectation: User search behaviour varies based on different dimensions such as: user interest, background knowledge, search task complexity and user experience. It is a challenging and complex task to investigate which dimension relate to user's knowledge gain, and to efficiently and effectively design studies to measure knowledge gain keeping other dimensions constant in experiment. Moreover, the short duration of experimental search sessions in laboratory settings adds further complexity to understanding users expectations [142].

3. User background knowledge anticipation: very often detailed information about the user is not available, thus it is difficult to anticipate their pre-existing knowledge about a search topic. Several studies have been conducted where users are asked to self report their pre-existing knowledge either using a likert scale (0 (low) - 7 (high)) [103] or using methods such as writing a summary describing the topic [107, 51]. Self reporting is a grey area through, depending completely on a user's feedback to describe their current knowledge of a topic.

4. Analyzing and evaluating knowledge gain: This is one of the most sophisticated and complex problems in search settings [51]. Broadly, there are three ways to address this issue:

(a) Comparing self reported Likert score: Studies can be conducted where the user is asked to anticipate their pre-existing knowledge before the

search task and their knowledge level after completing the search task based on a Likert score [103], and then measuring the difference [152]. This method is used in our studies in this PhD research.

(b) Comparison of summaries: A another method is where the user is asked to write a pre-search task and post search task summaries about the search topic. The pre-search task and post-search task summaries are annotated and scored based on three factors: facts, critiques and interpretations. Knowledge gain is measured by evaluating the difference in these scores [96]. This method is also used in this PhD research.

(c) Comparison of the post- and pre- test scores: Studies can also be conducted where the user is asked to attempt a test about the search topic before the search task and after the search task. The resulting test scores are used to evaluate the learning [107]. This method is very popular in fact-based information seeking, but it is very difficult to frame the right set of questions and answers to measure the scores for exploratory information seeking. Furthermore, in an exploratory setting, the answers provided by the searchers are subjective, and measuring the subjective answers increases the complexity of the evaluation. This method is not used in this PhD project due to focus of our research is on exploratory search.

There are always challenges to reporting the precise amount of knowledge gain by users in carrying out a search task since it depends on many external factors, as well factors such mood or attention of the user, the search interface, etc. The methods discussed above can only provide the rough estimates of knowledge gain. The next section outlines some standard methods for measuring learning in search task settings.

2.4.2.3 Methodologies to Measure Learning

Three types of investigation can be conducted to measure expansion of knowledge and learning: a) analysis of search logs, b) crowdsource based investigation, c) lab-based investigation

- Analysis of search logs: It is evident from multiple studies [69, 84] that analysis of the search log can provide insights into user search behaviour. A study was conducted by Eickhoff et al. [99] to investigate user knowledge acquisition based on the log files generated by the Bing search engine[1]. The study found that the behaviour of people varies during a search session based on the knowledge obtained. It also found that users were using new query terms found in the documents or snippets they have just encountered during the search. This is clearly evident in the logs, where the search process itself supports users by enhancing their knowledge, which in turns enables them to formulate better queries.

- Crowd source based investigation: Crowdsourcing is a method used by individuals or organizations to collect feedback about particular service or items from a large number of people based on the multiple factors such as demographic, background knowledge, etc. Multiple studies have been conducted to evaluate search systems and learning using crowdsourcing [86, 98, 115, 135, 189]. Syed and Collins-Thompson [135] conducted a study to investigate performance of a retrieval algorithm by evaluating it through crowdsourcing. Their retrieval algorithm was designed for educational utilities. In this study, they examined users' learning outcomes by measuring variation in pre- and post reading vocabulary quiz to measure learning outcomes associated with Bloom's taxonomy remembering dimensions. The authors found that ranking of the documents based on the density of keywords enhanced the user word and definition learning experience.

[1] https://www.bing.com/

- Lab-based investigation: This is the most popular approach used by the researcher in closed settings. Collins-Thompson et al. [107] carried out a lab-based user study in which they explored the factors and indicators responsible for learning in a web search setting. They observed that observed learning outcomes for searcher was closely matched with their actual learning outcome by analysing a questionnaire, written responses and search logs. Based on the analysis of the post-search survey, the answers searchers wrote were highly correlated with their cognitive learning question scores. The time spent per document while searching by searchers was also positively associated with higher-level cognitive learning question scores. It justifies the lab-based studies are adequate to measure the empirical evidence for the experiment. Based on the above study outcome, this PhD research focused on lab-based investigations.

2.5 Evaluation in Information Retrieval

Evaluation is an integral element of research in IR, specifically measuring the effectiveness of a research system to satisfy a user's information need. At present, there is no standard method for the evaluation of conversational search systems. Evaluating conversational search involves elements of IIR involving the user, and of the effectiveness of the queries created during the conversational search process. We thus need to consider both evaluation of IIR and laboratory-based evaluation of search effectiveness in our consideration of the evaluation of conversational search.

Conversational systems, in general, are typically evaluated using multiple dimensions such as quality of user interactions (depth of communication funnel which means how deep the participants are going into the conversation and the number of dialogue turns), response speed, functionality and performance-based measures (effectiveness and efficiency of an agent to complete the task) [139]. Performance-based evaluation of conversational systems is still at an early stage and investigation of user engagement with conversational systems is an important topic for further

analysis. The absence of existing work on evaluation means that this PhD work will need to explore methods for the evaluation of the use of conversational engagement in search.

We divide evaluation into three areas: evaluation of algorithm IR, IIR and conversational systems. Evaluation of conversational systems is discussed in Chapter 3, we briefly review existing methods for algorithmic IR and IIR evaluation in Appendix A.1.

2.6 Concluding Remarks

This chapter has introduced information seeking models relating to current information retrieval and IIR, challenges associated with learning, user engagement, experimental design and evaluation methods for IR and IIR and the experimental design adopted in our this PhD research.

The next chapter introduces conversational agents and the challenges associated with search and scope of conversational search and information seeking model in a dialogue setting.

Chapter 3

Recent work on Conversational Systems

3.1 Overview

Conversational search is a conceptual advancement on IIR where the user conveys their information need in the form of a dialogue. The use of dialogues in information seeking potentially enables the user to express their information need naturally via multi-modal signals (text or voice). Appropriate use of dialogues may support the construction of better search queries to enable the IR system to better understand the user's information requirement. For example, conversational search may enable the IR system to resolve ambiguity by asking for clarification from the user. This chapter introduces conversational agents, information seeking models in dialogue settings, recent work in conversational search, and challenges associated with conversational search.

3.2 Conversational Interfaces

Conversational interaction (CI) with information systems is a longstanding topic of interest in computing [42]. However, activity has increased greatly in recent years [114]. The key motivation for examining CI is to develop interactive systems which

enable users to achieve their objectives using a more natural and less demanding mode of engagement with computing applications. It is hoped that CI will be less cognitively demanding than interfaces primarily focused on functional requirements for which users are required to develop new mental models to use them reliably. Recent research on CI has focused on multiple topics including: mode of interaction, intelligence of conversational agents, structure of conversation, and dialogue strategy [101, 112, 183]. Progress in CI can be classified into four facet areas: smart interfaces, modeling conversational phenomena, machine learning approaches and toolkits and languages [163, 169, 172].

Smart interfaces are responsible for handling multimodal engagement with users and seek to provide responses in the appropriate modality. The issue of multimodality has featured in the development of many chat interfaces. For example, the University of Rochester built a chat interface which engages with the user to solve complex problems [44, 26]. A number of studies have been conducted to understand the effect of multimodality in CI [183]. With respect to the modeling of conversational phenomena, traditional CI depends on turn by turn engagement where the system and the user wait for each other's response. This began the process of understanding human behaviour, expectation of the next utterance, and generating corresponding responses. Multiple studies have been conducted to understand the incremental processing of knowledge and prediction of next utterances in engagement between users and conversational interfaces [16, 31, 56].

User engagement with CIs is generally mediated by a conversational agent. These agents currently take the form of a 'chatbot'. Current chatbot interfaces have evolved in common with many areas of AI from rule-based systems to the use of data driven approaches using machine learning and deep learning methods [180]. Toolkits have developed to help developers to construct and test chatbot agents for particular applications. The majority of research on conversational agents has focused on question answering and chit chat (unfocused dialogue) systems. Only very limited work has been done on information seeking bots, dating mainly from

the early 1990s [22].

Interest in more general conversational search applications has increased greatly in recent years [129], but much of this work has focused on conceptual issues [105, 136, 137, 147, 159], and their remains little work examining interfaces [174, 186] for user engagement with respect to the user experience, knowledge gain and software usability [156, 159].

3.3 Information Seeking Modelling in Conversational Settings

The modelling of information seeking in conversational search settings is different from information seeking models in conventional search (discussed in Section 2.3). The concept of dialogue-based IR was first introduced by Oddy [8] in a system called THOMAS. This aimed to help the user to select relevant documents without making explicit queries by asking questions to the user and showing them a list of resulting documents.

In 1980, Croft and Thompson [14] introduced Intelligent Intermediary for Information Retrieval (I3R) in a system modelled on an expert intermediary to support users in their search activities. Other work looked at Dialogue Acts (DA) as an approach of dialogue based IR [20, 30]. DAs represent the generic meaning of an utterance at the level of illocutionary force. For the example, the COnversational Roles (COR) model based on DAs developed by Sitter and Stein [20]. The main objective of this model was to guide users through the process of information seeking using a dialogue.

Figure 3.1 illustrates the COR model which features a user A (searcher) and an agent B (information provider) where the square and circle symbols represent the dialogue states. The arrows reflect transition from one state to another state. For example, from step 1-2 the searcher requests information from the agent. The COR model was not very clear with the result representation to the user and did not

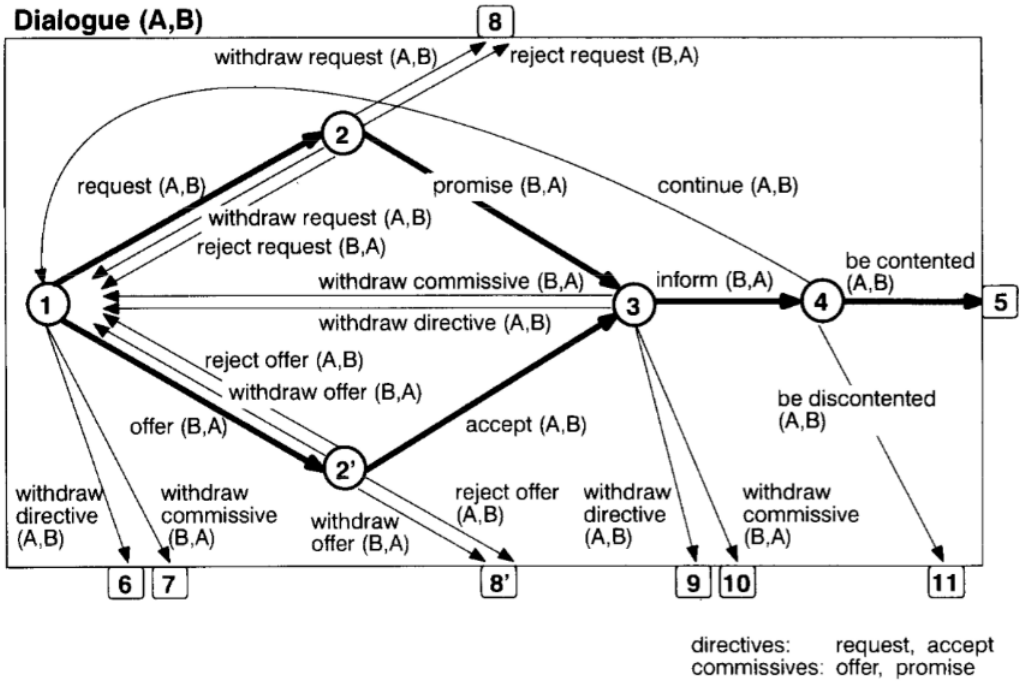


Figure 3.1: Illustration of the COR model [20].

explain the mixed-initiative concept. Mixed-initiative is an important concept in the conversational information-seeking process where either the user or the system has to take the initiative to interact with each other to facilitate the information seeking process. Mixed initiative conversational models allow natural interactions, but are very complex models to implement in a system [112].

Vakulenko *et al.* [170] introduced an advanced information seeking DA model inspired by COR Model. This dialogue-based information seeking model was designed based on the four loop interaction among the components Query, Request, Feedback, Answer (QRFA), as shown in Figure 3.2. Each loop is an interaction cycle between the user and the search agent, and includes the following actions: Question Answering, Query Refinement, Offer Refinement and Answer Refinement. Each user search interaction loop describes a conversational flow in a sequence of information seeking activities. Similar to the COR model, the QRFA model was not very clear with the result representation to the user and did not explain the mixed-initiative concept.

Weisser *et al.* [192] conducted a study with annotated spoken task-oriented

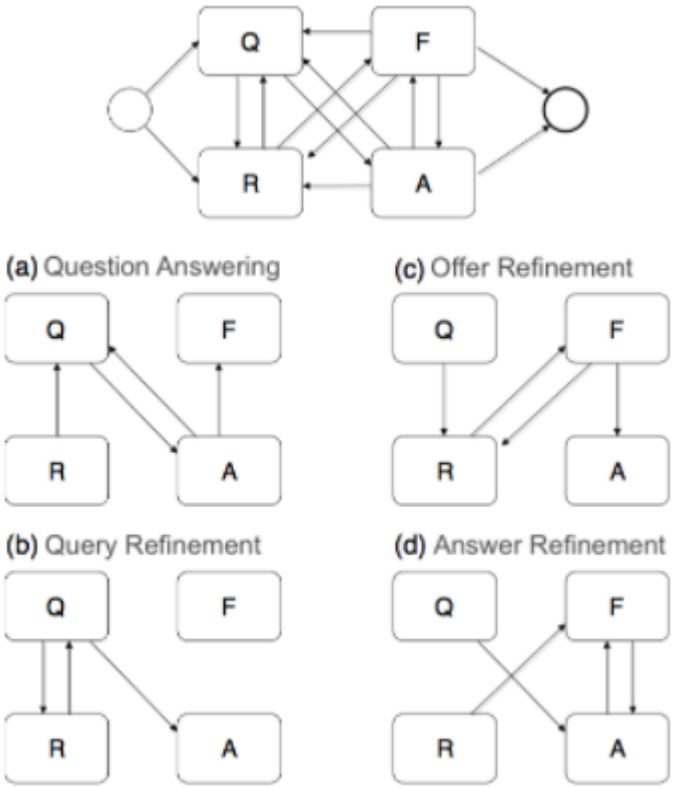


Figure 3.2: Illustration of Query, Request, Feedback, Answer (QRFA) [170].

dialogues using a dialogue taxonomy. The investigation was extended by Loisel et al. [77] by collecting human dialogues during related searches in a control setting. Analysis of dialogues identified common user requirement and proposed a model on the basis of a taxonomy analyzing a medical text corpus. Allen *et.al.* [9] proposed a model based on cooperative behaviour that can be applied in natural language understanding systems. This included a model based on a the method to deliver the generating responses that convey more information than was explicitly requested, where this information should not be too specific or too voluminous for the user. Its implementation in practical scenarios was complex and domain restricted. This was based on the a BDI (Belief, Desire and Intention) model.

Another study conducted by Roeck *et. al.* [37] used natural language techniques to parse the user query into a slot based template which was used to identify keywords to form a query for an IR model. The study conducted by Trippas *et al.* [120] introduced a new model based on spoken interaction for information seeking named as spoken conversational search (SCS). This model was empirically derived

from conversational transcripts collected by performing information seeking activities using a human search agent. This model used a speech only interface and was limited to three conversation turns. This model failed to understand the real life challenges faced by the user while interacting with a computational search agent, as the search agent in the development of this model was a human being. The challenges include understanding the capacities of the agent. Generally, the user has to carry out multiple queries to make the conversational system understand their information need. By contrast, the human intelligence agent can easily understand the user information need. Another challenge is the length of the query phase. The user is restricted to expressing their information need in a limited size of query due to restrictions in conversational system. Searching with the human agent allowed the users to express themselves without any specific word length.

Moreover, simple interaction with a SCS where the agent can clearly understand the user expectations is a very uncommon scenario in real life information seeking. Additionally, it has been found that users learn more when using a combination of visual outputs with audio outputs in contrast to the only audio outputs [61]. Considering the expectations of conversational search [129], the SCS model fails to address the prospect of learning while doing the search.

Another information seeking model are non DA (Dialogues Act) models which are based on the discourse aspect of dialogues. As an example, Belkin *et al.* [13] describe a schema to annotate information seeking conversations between librarians and users to outline the design of an expert system. The schema explained the possibility of extracting a range of information from dialogues including the description, search states and search interactions.

Belkin et al.[27] introduced an advanced model based on scripts, where each script describes the functions of information seeking. Depending on the type of the information need and the associated interactions, the model selected the most appropriate action from the scripts to satisfy the user's information need. Models like scripts [27], COR [20] and QFRA [170] allowed the search system to predict

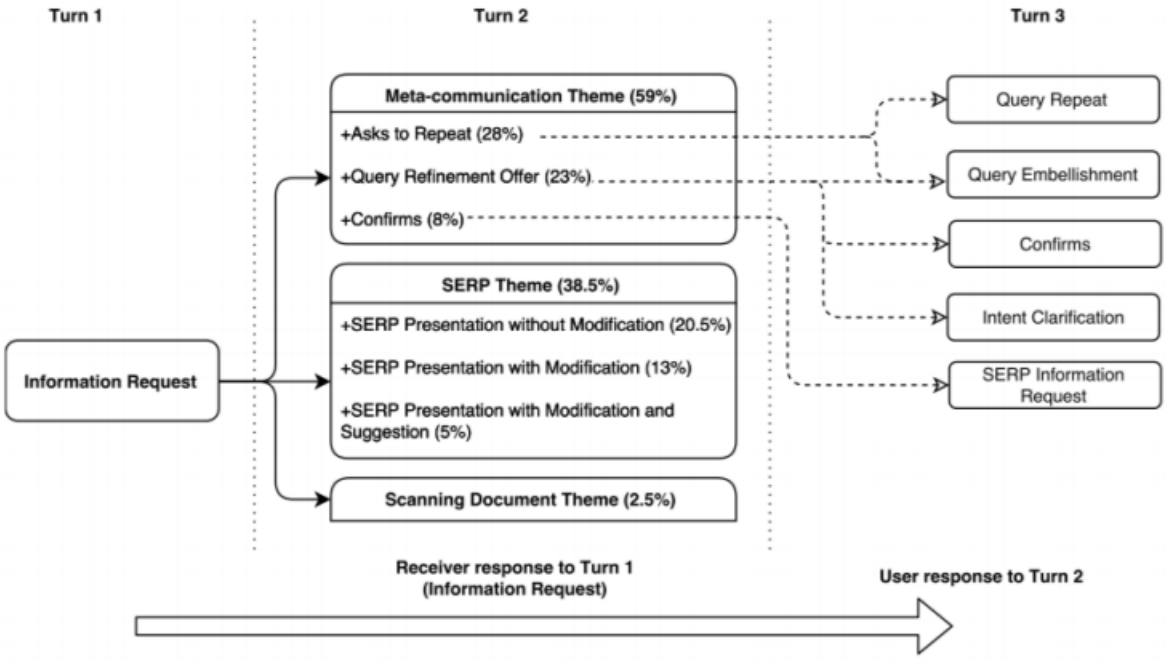


Figure 3.3: Illustration of Spoken Conversational Search (SCS) [120].

the next move based on user previous interactions. These predictions were selected from the planned discourse routine which enabled robust responses to the user.

After referring to multiple information seeking models in conversational settings, we noted gaps in various areas, including representation of results to the user, understanding the user's challenges while working with a conversational system, and implementing these information seeking models in real-life scenarios. In this PhD research, we focus on all these gaps and investigate rule-based and machine learning based multi-view conversational search systems to reduce these gaps and to explore the potential for efficient and effective conversational search interfaces.

Another advanced conceptual information seeking model was introduced by Azzopardi et al. [144] This combine the interaction and action spaces [129, 157]. This is not limited to DA, but may also provide an overview of the potential action to be taken by the user or the search agent. This conceptual framework has not been validated empirically.

Another study [129] investigates conversational approaches to information retrieval and presents a theory and model of CS which explores features of a conversational information retrieval system for users to be able to respond to an IR system in

dialogue settings. In this model, different properties are defined for conversational search, such as user revealment, mixed incentive, etc. The study did not include details of the conversational interface. This conceptual framework has also not been validated empirically.

3.4 Existing Research on Conversational Search

In this section, we review current work on conversational search systems examining three major approaches: a) exploration of the search in conversation system such as smart speaker like Alexa, Google assistant, b) the machine learning approach to explore the potential of extracting information need from the set of dialogues, c) conversational search using human experts and Wizard of Oz methods to explore information seeking behaviour.

3.4.1 Studies on Conversation System

Vtyurina et. *al* [138] conducted a study to explore opportunities for conversational search involving humans, wizards and Google assistant in a complex information seeking task. The study threw light on human-to-human conversations while searching. It also indicated that the humans are not biased towards machines, but that they lack trust when they were dealing with other human beings. The study also indicates various human factors which can effect the process of search, such as trust and social burden. Although the authors discussed two types of user behaviour one where users predefined broad answers, while in the other they like direct answers, explanations regarding reasons for different kinds of behaviour and challenges associated with all three approaches with human, wizards and Google assistants were missing. There thus remains a gap in understanding regarding user behaviour for further investigation in relation to conversational search settings.

Loepez et.*al* [154] conducted a study on conversational agents (Alexa vs. Siri vs. Cortona vs. Google Assistant) to investigate their usability of these different

applications such as access to music services, agenda, news, weather, To-Do lists and maps or directions. Their results showed that even though there are many services available, there is a much to do to improve the usability of these systems. The authors found that Alexa is more accurate in search, for online shopping, while Google assistant is remarkable in its naturalness when answering fact based questions.

Hoy [151] reviewed multiple studies related to the conversational agents and defined a precise definition of conversational agents and discussed the usability and capabilities of conversational agents (Apple's Siri, Amazon's Alexa, Microsoft's Cortana, and Google's Assistant). Hoy summarized the capabilities of agents, e.g. they can ask their assistants questions, control home automation devices, playback media via voice, and manage other basic tasks such as email, to-do lists, and calendars with verbal commands. He also discussed several problems with current voice assistant products. Issues examined included privacy and security controls, which will need to be improved before voice assistants can be used for anything that requires confidentiality [151].

Lopatovska et *al.* [153] conducted a study on Alexa to investigate user interactions, focusing on the types of tasks requested from Alexa, the variables that affect user behaviours with Alexa, and other conversational agents. The results indicated that across all age groups, Alexa was primarily used for checking weather, playing music, and controlling other devices. Users reported being satisfied with Alexa even when it did not produce sought information, suggesting that the interaction experience is more important to the users than the interaction output. More work is required to understand whether users treat Alexa and similar voice-controlled applications as primarily a traditional information retrieval system, a casual leisure system, a control interface for smart home devices, or, simply, a new toy.

The study conducted by Shamekhi et *al.* [118] investigated how a set of vocal characteristics known as 'conversational style' could play a role in engaging users in their interactions with a virtual agent. The authors examined whether a similarity attraction principle influenced how people orient themselves towards agents.

Their results revealed that subjects exhibited similarity attraction toward computer agents, and preferred agents whose conversational style matched their own. The study results contribute to the understanding of how the design of conversational style influences users' engagement and perceptions of intelligent agents, compared to human-to-human interaction.

Budiu *et al.* [191] conducted a study to examine the usability of an intelligent assistant (Alexa, Google Assistant, or Siri) on the variety of the tasks ranging from simple (e.g., weather for the 4th of July weekend) to more complicated (e.g. traffic to Moss Beach during the weekend). Usability testing showed that both voice-only and screen-based intelligent assistants work well only for very simple queries that have fairly simple, short answers. Users found difficulty with complex task such as exploratory information seeking.

This section presented existing work related to conversational agents, intelligent speakers and their use for information seeking and goal-oriented tasks. This allows us to explore the use of conversational agents in different dimensions and identify gaps in existing knowledge. In this study, we explore smart speakers with a screen in complex information seeking activities to study the effect of multi-modality in the search process in a conversational setting.

3.4.2 Machine Learning for Conversational Search Applications

This section outlines the opportunities and challenges for use of machine learning in the process of information seeking in conversational settings.

Conversational search has been a topic of growing research interest in recent years. In 2019 TREC introduced the Conversational Assistance Track (CAsT) [176] which seeks to develop a standard benchmark for the evaluation of conversational search methods. In this track, the main objective was to satisfy a user's information need by understanding a sequence of questions in a conversational format and retrieving relevant documents associated with contextualized query at each turn in

the conversation. Participant submissions focused mainly on three aspects:

- Query Understanding: Appropriate interpretation of the queries across the conversations was particularly challenging. The most popular query understanding technique was deep learning, with 57 % of runs using it. NLP toolkits were used in half of the submitted runs, but there was no improvement in performance observed from from using them.

- Retrieval and Ranking: Using any training data led to improvements in the score. Unsupervised learning approaches were used in 43 % of the runs, but were not found to be effective [176].

- Conversational Context: Almost all runs utilized information of previous turns to interpret the context of the current query and use it to attempt to improve retrieval effectiveness. The title of the conversational topic also played a crucial role in identifying context. Few runs involved all the turns and metadata, since the long description associated with doing this became noisy and was hard to use effectively.

For the ranking process, BERT-based neural models [176] were found to perform the best for this task. However, in some situations, these were outperformed by traditional IR systems. Neural re-ranking approaches were also shown to be effective.

A new data set was introduced for Conversational Question Answering (CoQA) systems [168] at Stanford in 2019. This data set was built for the evaluation of systems that have to understand a series of conversations and to answer questions about them. The key features of this dataset are: answers are free form; and secondly, follow-up questions are more complex, which reflects human conversational behaviours. This data has 80,000 conversations from which 127,000 questions with answers were extracted from seven different domains. After experimentation, it was observed that increasing the size of the history used in answering the questions decreases a model's performance. However, when previous turn data is shown, there is an increase in the model's performance. This was observed to reach higher

performance when two previous turn conversations are shown, implying that most questions in a conversation have limited dependency within a bound of two turns [123] [133].

This section reviews the machine learning approach to understand the challenges with supervised and unsupervised methods to support conversational search. We have studied two conversational search tasks solutions to which use the machine learning approach to extract the information from the dialogues and significant challenges associated with this query extraction methods [160, 167].

3.4.3 Human Experts & Wizard of Oz Experiments

While users of search tools have become accustomed to standard "single shot" interfaces of the form seen in current web search engines, and have learned to use them to good effect, interest in the potential of alternative conversational search based tools has increased greatly in recent years [129]. Traditional search interfaces have significant challenges for users in requiring them to express their information need in fully formed queries, and the idea of agent support conversational based interaction supporting them in the search process is thus very attractive. Multiple studies had been conducted to investigate the potential of conversational search in different dimensions. These studies however have generally involved use of a human in the role of a conversational agent wizard to enable the search process [143, 161, 173].

The concept of computer-mediated dialogue in search is a longstanding goal within IR research, although until recently, not one which had received significant attention [33]. Within this work, one area of study is understanding the processes of a user engaging in search in conjunction with a human intermediary. An early key study reporting a study of this type is described in [12]. The goal was to work towards a model to enable the specification of an *information provision mechanism* which cooperates with the user in an information seeking task. This work was set in the context of academic researchers seeking information in libraries with the support

of professional librarians. Much more recently work by Trippas *et al.* at RMIT [137, 157] has performed similar types of studies focusing on speech only based search, focusing on much more general topics of the type that a user might post to a web search engine .

While understanding the conversational behaviour of searchers in fulfilling their information needs with the assistance of expert searchers, and of the potential for human like intermediaries to assist with web search type activities is of course important, implementation of a conversational agent able to support search in this way would require a highly competent agent. In this study, we approach conversational search from the contrasting objective of understanding current web search type activities as undertaken by users without intermediary assistance, with the aim of identifying opportunities for the inclusion of conversational assistance to support them.

The subjects were given the impression that they were interacting with a machine. While the results of these studies have been interesting and insightful, they have an important limitation in that the agent has full human intelligence. Thus they do not reveal the potential for artificial agents to support search in terms of effectiveness and user acceptance. Studies have also been conducted to investigate the user search behaviour in speech settings where the searcher interacts with the agent (the human "wizard") via speech. These have the limitation of assuming both human intelligence and error free speech recognition, which will generally not be the case in a real system [137, 157].

Some studies have been based on a data driven approach using machine learning methods to extract a query from multiple utterances [176]. This approach's drawback is that the dialogues are not analyzed based on incremental learning over multiple conversations [122, 128]. In contrast, as per the Vakkari model, search is a gradual process, and the user learns with each interaction and adds value to the search process [121]. Other types of studies have developed agents by using an intermediate approach in which a combination of rules to form a dialogue strategy

after generalising user search behaviours [152]. Dialogue strategy guide the user in conversations with the support of a pretrained machine learning model to extract the intent and entities from utterance. Intent helps the agent to understand the user's requirements and to support them in information seeking [166]. From our study of this existing work, we note the gaps that need to be explored in current conversational search.

All three approaches examined in conversational search help us understand the advantages and disadvantages of existing work using each method. For example, usability challenges with conversational systems and limitation with the wizard scenarios. Moreover, recent work also helps us to understand user search behaviour in all three approaches. Additionally, review of the machine learning approach enabled us to understand the challenges of supervised and unsupervised methods to support conversational search. We also analyzed the potential scope for better evaluation methods for conversational search.

3.5 Parameters of Conversational Search

As discussed in the Dagstuhl Seminar report [171], conversational search is a complex process that depends on diverse dimensions such as cognitive load, user experience. This section examines six specific parameters that can influence conversational information seeking and how it affects the user in the search process. We also outline the challenges associated with conversational search and its consequence for the information seeking process.

3.5.1 Essential Factors for Conversational Search

We can identify the following essential factors for the evaluation of conversational search (CS) interfaces.

1. Cognitive Load: Conventional search can impose a significant cognitive load on the searcher [166]. An important factor in the evaluation of conversational

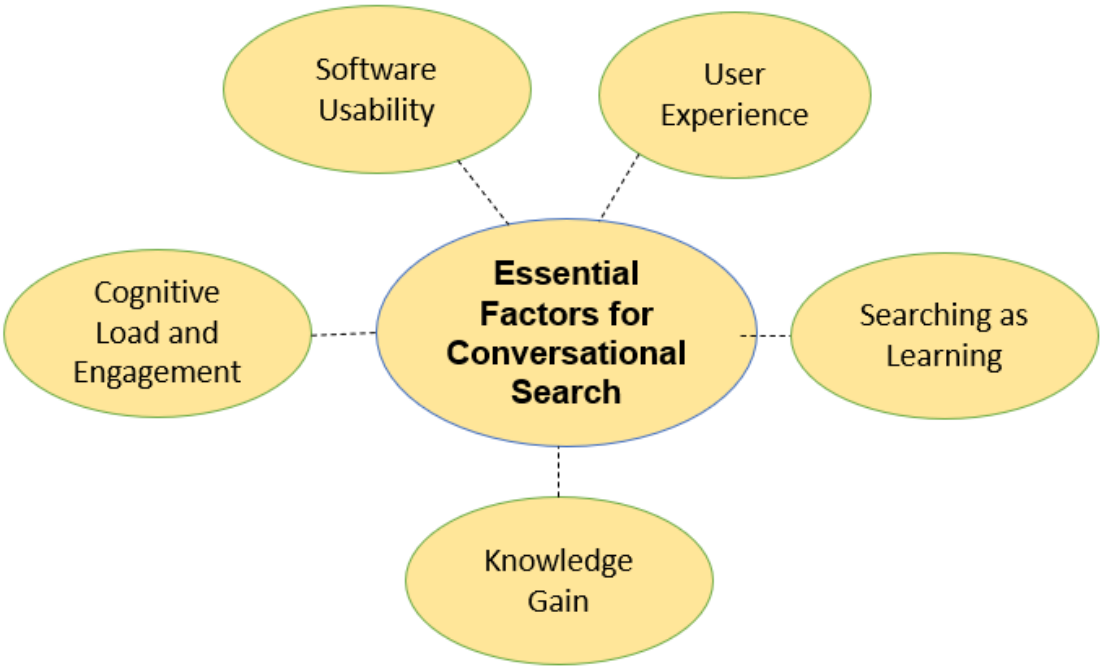


Figure 3.4: Essential factors for conversational search.

systems is measurement of the cognitive load experienced by users while using them.

2. Cognitive Engagement: It has been observed that users get frustrated if they find it difficult to search about their specific topic of interest. Frustration can reduce the user's engagement with a search system and their associated effort to locate relevant information [82, 160, 166].

3. Search as Learning: Learning while searching is an integral part of the information seeking process. high cognitive load and lower cognitive engagement impacts on user learning during the search process [82, 121, 152].

4. Knowledge Gain: Satisfaction of the user's information need is directly related to their knowledge gain about the search topic. Knowledge gain can be measured based on recall of new facts gained after the completion of the search process [96].

5. User Experience (UX): Another important aspect that needs to be considered for evaluation of CS systems is UX. User experience is generally classified

into two aspects: pragmatic and hedonic, where pragmatic quality explores the usage experience of a conversational search system and hedonic quality explores the pleasantness of using the system. Both aspects of UX can be further divided into six components: attractiveness, perspicuity, efficiency, dependability, stimulation and novelty [150]. These factors all provide measures of user ease of use and the dependability of a conversational system [150].

6. Software Usability: CS studies generally do not explore the dimensions of software usability [143, 149, 173, 190]. However, it is important to understand the challenges and opportunities of conversational systems on the basis of software requirements analysis. This allows a system to be evaluated based on real life deployment and to identify areas for improvement. Lower effectiveness and efficiency of a software system can increase cognitive load, reduce engagement and act as a barrier in the process of learning while searching.

3.5.2 Challenges Associated with Conversational Search

Conversational search is a complex information seeking process with numerous associated challenges and effects on users. Despite technological advancements, conversational systems are still at an early stage of development and need much further research in multiple dimensions, such as user experience, usability, task load. Current conversational systems are generally designed for goal-oriented simple tasks such as setting up an alarm, searching for a contract, calling a contract, customer care help desk. However, in relation to conversational search, the current-state-of-the-art systems have to overcome many challenges to support complex information seeking processes in conversational settings. These challenges relates to the technical issues and information space. The major challenges are as follow:

1. Long Conversation: State-of-the-art conversational agents are well suited to fact-based questions and non-exploratory searches, but these conversational agents find it difficult to maintain long conversations without losing the context of all conversation [164].

2. Input and Output Modality: Generally the conversational agents are based on the linear mode of input (speech) where users are restricted to expressing themselves in specific words in a limited time, which restricts conversational systems and their functionality. Therefore, the user is compelled to transmit their information need in a number of limited words in which they can be unable to express their information need [49, 105]. Secondly, they can be spoken agent systems are adequate to present the results in a complex structure such as a SERP, including hyperlinks, videos, images and graphs [91]. Commonly, conversational search presents results in speech or small segment combining text image and speech together in the small space available on smart speaker system screens, which restricts the user to only being able to work with the information that can be held in the user's short-term memory.

3. Extensive Capabilities: Conversational agents are currently becoming very popular for properties discussed later in the section 3.6 but are incapable of providing opportunities for scanning of search results, refining the query or going back to previous results to verify facts or satisfy information needs [166].

4. Other: Very often, users find it difficult to express their information needs in a limited number of words. Current state-of-the-art systems are still at an initial phase of development, and often fail to meet user expectations of the conversational search [6]. Thus, voice-based systems referred to as task-based system inadequate for performing complex information seeking interactions and engagements [111].

3.5.3 Influences on Searchers

The above mentioned challenges cannot be overlooked, as these challenges have huge effects on the search process. The above concerns result in the following consequences which can broadly be divided into four parts:

1. User Frustration: Generally, the user gets frustrated in cases where they are

unable to express their information need to the search system. The user may gets demotivated and not complete the search process or make do with being partially satisfied incomplete information.

2. System Frustration: This case describes the challenge for a system to understand the user's information need and to express its capabilities to the user. In a typical case, the system cannot understand the user accent or speech correctly and provides non-relevant search results. Additionally, the linear input modality fails to satisfy the conversational search principle of "System Revealment", in which the system expresses its capability and potential to the user.

3. Effect on Cognitive Load: It is evident that due to user frustration and system failure, to provide support in complex information seeking leads to an increase the cognitive load of the user. Moreover, systems with small display and without displays at all prevent users from scanning documents or snippets. Thus the user is unable to learn much about the search topic to acquire information to support query modification. Thus, the user can struggle with increases in cognitive load when using conversational system.

4. Effect on User Engagement: Increased engagement with search systems directly correlates with better learning experience and knowledge gain [181]. As discussed in the above section, it is clear that the user finds the current state-of-the-art conversational system challenging to use leading to demotivatation and reduced efforts to perform search operations, leading to poor knowledge gain [152].

3.6 Evaluation of Conversational Search

This section reviews methods which have been adopted to evaluate conversational systems. In particular, we identify the shortcomings of existing work in the evaluation of conversational search.

In this section, we overviews evaluation methods adopted for conversational agents. This includes a detailed review of techniques, tools, and factors considering for evaluation. The main objective of this evaluation is to enable conversational systems to be analyzed to enable understanding of behavior with respect to user needs. Current conversational agents are evaluated using six dimensions:

1. Extensive capabilities : Good extensive capabilities of a conversational system, provide an error-free environment to the user [95, 130, 139]. Extensive capabilities enable the agent to provide better and more robust services to the user, such as speech to text, text to speech, avatar assistant, personalization system, robust functionality to handle any user utterances, etc. This should include spell checking and auto-correction to enable the user to express their thoughts correctly. Furthermore, a conversational system should be able to correctly identify statements and questions. In addition, it should be capable of using an appropriate amount of multimedia content in balance with text [124, 140].

2. User Interaction and Engagement: This is a qualitative measure that deals with user interaction. This measure contains the following parameters: capable of internalizing a conversation, maintaining engagement with the target user [158], respond with meaningful responses, support the user in navigation of the content shown in the conversational system, and designed to answer frequently answered questions based on search history and personal information relating to the user [95, 131]. The above mentioned parameters are good measures of the retention features of the conversational system [130, 139].

3. Response speed: Significant delay in response from a conversational agent gives rise to user frustration [119, 130, 139].

4. Functionality: A qualitative measure that can be classified into multiple variables such as richness of media, navigation tools to support users, multimodality [95, 134].

5. Scalability: Scalability is a key quantitative feature for any system. For example, the number of users supported by a conversational system at the same time, types of server that can accommodate the conversational system, database size [139].

3.7 Conclusion

This chapter has reviewed existing work relates to conventional systems, conversational information seeking and conversational search. This has included approaches used by other researchers to explore the domain conversational search and different information-seeking models. The potential of conversational agents in complex information seeking is explored in later chapters of this thesis. The next chapter begin our investigation by examining user action in conventional search and seeking to identify opportunities for conversational support in the search process.

Chapter 4

Methodology

4.1 Overview

This chapter introduces the experimental design practices used for our interactive information retrieval investigations, a detailed description of our experimental setup and a new framework for the implicit evaluation of CS that includes multiple dimensions to develop a clear understanding of CS processes. The dimensions we consider are: search experience, knowledge gain, software usability, cognitive load and user experience. In this, we take a user prospective approach to understand the evaluation of CS in exploratory settings considering user reporting of information seeking activity.

4.1.1 Experimental Design in Interactive Information Retrieval

Experimental design for the investigation of user search behaviour is a complex process. As highlighted by Kelly [75], user search behaviour is dependent on many factors such as mood, pre-existing knowledge and interest in the search topic. It is very difficult to study the effect of individual factors in the interactive search process. This makes it very challenging to design an experimental setup which allows multiple users to have the same feel or experience while using it. Another complex task is to

understand the relationship between these factors. However, due to the increased interest in the human computer interaction (HCI) and IIR communities, multiple studies have appeared [65, 75, 87, 104, 108] which focus on development of standard practice for design and evaluation of IIR systems. During this PhD research, these studies are used as the source of methods for the design of experiments for our investigations. In the next section we review some of these studies highlighting features important to the design of experiments and investigations. In this section, we discuss the topic of remote studies vs lab-based studies, the effect of limited time duration in experimental setups and the adoption of task sequencing strategies for the allocation of tasks to the user to avoid any biasing effects.

4.1.1.1 Are remote based studies more accurate than lab based studies?

A study conducted by Kelly and Gyllstrom [87] compared lab based vs remote based for IIR experiment. The investigation was conducted on two groups of people from the same population. The first group participated in the experiment remotely and the second group participated in the laboratory. Both the groups were studied on the basis of the user behaviour, search behaviour and evaluation behaviour. For most of the measures, there were no significant differences between the settings. This demonstrated that user behaviour does not change significantly based on experiment search setting. Following these findings, the experiments conducted in this PhD research use either lab based setting or remote setting depending upon prevailing external factors in terms of available user context (restrictions arising from Covid19). The experimental setup and design of each investigation is described in detail in the relevant chapter.

4.1.1.2 What is the effect of time constraint on user search behaviour?

The study conducted by Crescenzi et al. [94] investigated how time restriction and system delays impact on the user during search. In this study, half of the participants conducted the search in a time restricted setting and the other half without a time

constrict. It was observed that the users with limited time for search experienced increased stress and pressure. Moreover, these time restricted participants claimed the task to be more difficult and were less satisfied with the outcome. Following those observations and findings, the experiments conducted in this PhD are not time constricted. Not imposing a time constrict makes the experimental setting closer to an operational search environment. This is important since we do not wish to impose artificial stress on our participants while working with our prototype systems.

4.1.1.3 What is the effect of sequencing?

It is important to ensure that there are no sequence or order effects in an experiment to ensure results are not affected by potential sequencing effects. Sequencing effects can increase the chance that results are due to experimental conditions rather than genuine differences in user behaviour per task resulting from the experimental condition. To avoid any biases on the experimental setup the search tasks are rotated and counter balanced. Studies such as [59, 75] describe the use of the Latin square method to arrange the search tasks to avoid order effects. Based on these methods, the experiments conducted in this PhD research all arrange the search tasks using Latin square sequencing methods. In Chapter 6, Chapter 7 and Chapter 8, each user had to perform search tasks using both the settings, sequenced alternatively so that equal numbers of participants used each setting first.

4.2 Experimental Methodology

In this section we describe the details of the experimental setups for our user studies. The studies aim to enable us to observe and better understand the behaviour of non-specialist searchers whose techniques for use of search engines are generally learned from personal experience. The objective of these studies is to provide input into specification of potential conversational extensions to existing search applications and conversational agents. We seek to gain insights not only into how conversation engagement might directly be incorporated into current user search activities via

standard search engine or conversational agents or a multiview conversational search interface (MCSI), but also where there are opportunities to enhance the user's search experience. The studies aim not just to observe user behaviour in completing their search tasks, but also to gain insights into their objectives at each point in their interaction with the systems used during this PhD. In doing this, we hope to gain insights into the relationship between user actions and information seeking and learning, in order to be able to make use of this in the design of future conversational search agents.

4.2.1 Information Needs for Study

```
It is late, but you can't get to sleep because a sore throat has
taken hold and it is hard to swallow.  You have run out of cough
drops, and wonder if there are any folk remedies that might help you
out until morning.
```

Figure 4.1: Example backstory selected from UQV100 test collection.

For our search tasks we wish to give searchers realistic information needs which could be satisfied using a standard web search engine. The objective is to provide standardised search tasks throughout this PhD study. so users' search approaches on different search platforms with a consistent cognitive complexity load focuses on tasks. This enables us to to measure the tradeoff between each search platform in controlled settings. This also helps to avoid any types of bias effect due to variations in the cognitive complexity of the search tasks. In order to control the form and detail of these, we decided to use a set of information needs specified within *backstories*. The backstories that we selected where taken from the UQV100 test collection [106]. This consists of 100 backstories created from the TREC 2013 and 2014 Web tracks. In previous work, the creators of the UQV100 test collection developed an earlier collection of backstories in which they classified search tasks according to their level of cognitive complexity based on the Taxonomy of Learning [48], from which they adopted three of the five cognitive dimensions: *Remember*,

Understand and *Analyze* defined as follows:

- Remember: retrieving, recalling and recognizing the information from memory.
- Understand: developing the meaning of verbal and non-verbal communication through exemplifying, explaining, interpreting, classifying, comparing and inferring.
- Analyze: dividing the problem into small parts, examining how each part relates to each other, and to an overall structure through breaking, organizing, and attributing.

The search tasks in this earlier study were based on TREC tasks from 2002, 2003 and 2004, and were labeled with their level of cognitive complexity [102]. While all three types may benefit from search using conversational interaction, we decided to focus on the most cognitively engaging *Analyze* type in the expectation that this would require a greater level of activity to satisfy the information need. However, we chose to use the UQV100 backstories since they relate to more recently created queries for web search tasks.

Since the UQV100 topics were not provided with these labels, we selected a suitable subset as follows. The UQV100 topics were provided with labeled estimates of the number of queries would need to be entered to address the described information need and the number of documents that would need to be accessed in order to satisfy their information need. This labelling provided by the developers of the UQV100 topics are referred as backstories in this thesis [106]. We used the product of these figures (number of queries and documents) as an estimate of the expected cognitive complexity, and then manually selected 12 of the highest scoring backstories that we rated as the most suitable for use by general web searchers, e.g. not requiring specific geographic knowledge or of specific events. This set of backstories (as shown in Appendix B.7) were used in the studies of Chapter 5, 7 and 8. Figure 4.1 shows an example backstory selected for use in our study.

As a source of information needs for our study in Chapter 6 (Conversational

```
Martin Luther King Jr.  was an American Baptist minister and activist
who became the most visible spokesperson and leader in the civil
rights movement.  You have to find Information about the personality
using Alexa skills (as per the search setting) and based on your
information gain, you have to write a short summary (in the
questionnaire) about the person mentioned above and fill the
questionnaire accordingly.
```

Figure 4.2: Example backstory for our investigation of Alexa as an exploratory search tool.

agents), we provide participants with backstories requiring information about an individual which we anticipate users to able to address using a single Wikipedia autobiography page. While the user should be able to address general information need from the single Wikipedia document, this is a long and complex document required considerable engagement from the user.

An example of a backstory expressing an information need of this sort is shown in Figure 4.2. We developed twelve backstories for which full review of the corresponding Wikipedia autobiography page is a cognitively complex task, such a task would be classified as class "Analyze" within the Taxonomy of Learning [48].

4.2.2 Experimental Procedure

Participants in our study were required to complete a search session consisting of multiple search tasks based on our selected backstories. As part of their search session they had to complete a questionnaire before, during and after undertaking each task. In this section, we first give details of the standard structure, and then describe our use of questionnaires and our experimental setup, and follow this with the procedures used for our studies (ethical permission, recruitment, pilot studies and thematic studies).

4.2.3 Standard Structure

As shown in the Figure 4.3, the standard procedure for the studies conducted during this PhD started from ethical permission approval, followed by recruitment of the

subjects for the experiments. A pilot study was conducted with a small number of subjects to identify and address any issues with the design of the study. After this, the main study carried with the main subject group. Every subject has to give consent for their participation in the study. In each study, individual subjects had to do two search tasks or one search task per system as per the study design discussed in each chapter. The subjects need to fill the relevant questionnaires and to participate in semi-structured interviews based on the study design of each study discussed in this PhD. The details about each step mentioned above discussed in the individual next sections.

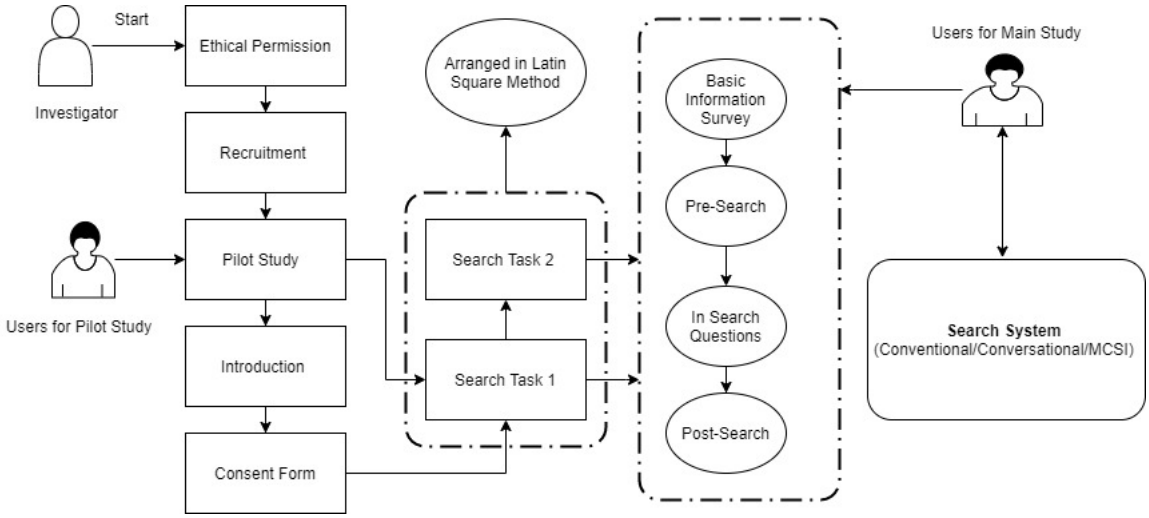


Figure 4.3: Schematic representation of experimental process including user engagement.

4.2.3.1 Questionnaire

While conducting the search the user had to complete an online questionnaire in a Google form while undertaking their search activities. The detailed questionnaire can be found in appendix B.1. For each search task the participant completed a questionnaire divided into four sections (Questions varied per section according to the study design):

- *Basic Information Survey*: Participants entered an assigned user ID, age, occupation, task ID to be undertaken. Used in the studies of Chapter 5, 6, 7, 8.

Questions
What search query did you enter?
Why did you choose these words?
Which results look useful to you?
(Please enter the document names and links)
How did you decide which results are useful?
If you opened any documents, why and what did you learn from them?
Did this query enable you to satisfy the search task?
If not, what will be your strategy to refine the query for the next iteration?

Table 4.1: In search questions from the questionnaire.

- *Pre-Search*: Details of their pre-existing knowledge with respect to the topic of the search task to be undertaken. Used in the studies of Chapter 5, 6, 7, 8.

- *In search Questions*: Details of their search activities, including details of the queries entered, documents examined, what was learned. This section was filled for each query entered while seeking to satisfy the current information need. Used in the studies of Chapter 5.

- *Post-Search*: Post-search feedback from the user. Used in the studies of Chapter 5, 6, 7, 8.

4.2.3.2 Experimental Setup

Participants used a setup of two computers arranged with two monitors side by side on a desk in our laboratory. One monitor was used for the search session, and the other to complete the online questionnaire. Participants carried out their search tasks using the standard Google web search engine using a Google chrome browser, conversational agent or MCSI (based on study design of each chapter). The search question section of the questionnaire is shown in Table 4.1. A web-tracking plugin was installed in the browser to capture details of the participant's activities with the browser. In addition, all search activities were recorded using a standard screen recorder tool to enable post-collection review of the user activities.

4.2.3.3 Pilot Studies

A pilot study with two undergraduate students in Computer Science was conducted using two additional backstory search tasks to see how long it took them to complete the sections of the study, gain insights into the likely behaviour of participants, and to generally debug the experimental setup. We conducted a separate pilot study for each investigation during this PhD.

Participants were given printed details of the instructions for their search sessions and each backstory in printed form at the beginning of each task. Results from the pilot study are not included in the analysis. Each of the pilot search tasks took around 30 minutes to complete. Feedback from the pilot study was used to refine the specification of the questionnaire to make it easy for users to understand each question and the corresponding instructions.

Based on the results of the pilot study, each participant in the main study was assigned two or one (based on time and efforts) of the selected 12 search task backstories with the expectation that their overall session would last around one hour. Pairs of backstories for each session were selected using a Latin square procedure to ensure an even distribution of search tasks among participants and to avoid task sequencing effects. The procedure (alternate sequence for each setting) to arrange backstories to avoid any sequence and bias effect is implemented for each investigation during this PhD.

4.2.3.4 Thematic Analysis

The summaries, videos and interviews were thematically labelled by two independent analysts (one male and one female) and Kappa coefficients were calculated (approx threshold mean 0.85 for all studies) [7]. Disparities in labels were resolved by mutual agreement by analysts. The independent coders were PhD research students working in Irish Universities and having sufficient experience in user-centered design experiments. The same coders were assigned throughout the PhD experiments to keep the whole process consistent. The category of themes are ‘user friendly and

robustness', 'interactive and engaging', etc. The example of labels is shown in Appendix B.3.4. The questionnaire in the interview dealt with user search experience, software usability and cognitive dimensions and was qualitatively analyzed. Some other example of categories are helpful, relevant, easiness, etc. The sample coding example is provided in Appendix B in Table B.3.3. This was done using an annotation schema relevant to our research aims designed after investigating the data; the complete response dataset was then coded using these data-derived codes [93]. The interview questions is shown in Appendix B.3.

4.2.3.5 Ethical Permission

Approval was obtained from the DCU Research Ethics Committee prior to beginning the data collection in all the user studies conducted during this PhD. As shown in Table 4.2, the experiment name and its corresponding ethical permission number.

4.2.3.6 Recruitment

All recruitment of participants was conducted after ethical permission and approval from the university. The process of recruitment was conducted through different modes, including advertising to different university groups through email, people to people (P2P), including groups and individual advertisements. The population for each experiment varied depending upon external factors such as COVID 19 restrictions. The population distribution of each studies discussed in the individual chapters. Although the focus of this PhD project is independent of the age or gender categories, we tried always to seek to achieve a balance across age groups and gender ratio to avoid any potential bias effects. During this study, we noticed the balance can be maintained between the genders by keeping the headcount low. An increased headcount, based on the demography of the specific location, tends to shift the bias towards the majority community. In each study, 1-2 participant records were excluded from analysis, since they did not follow the experiment instructions carefully. The data provided in the each chapter is for those participants whose

data was used for analysis. The whole experiment activity voluntary (which allow them leave the experiment in the middle, if they do not want to continue) and participants were informed about the whole process. The next section include the evaluation framework techniques adapted for the user studies.

Experiment Name	Ethical Permission No
Chapter 5 (Study with Conventional Search)	DCUREC2018_093 (main application)
Chapter 6 (Study with Conversational Search)	DCUREC2018_093 (amendment)
Chapter 7 (Study with Rule based MCSI)	DCUREC2018_093 (amendment)
Chapter 8 (Study with Machine learning based MCSI)	DCUREC2018_093 (amendment)

Table 4.2: Details about ethical permission.

4.3 Framework for the Implicit Evaluation of Conversational Search Interfaces

Most CS studies have focused on user search experience of the task or the usability of CS systems [149, 190]. This has provided feedback focused on user search experience. As part of this PhD research, we introduce a new framework for implicit evaluation of CS based on six factors related to the needs of CS, as discussed in Chapter 3.

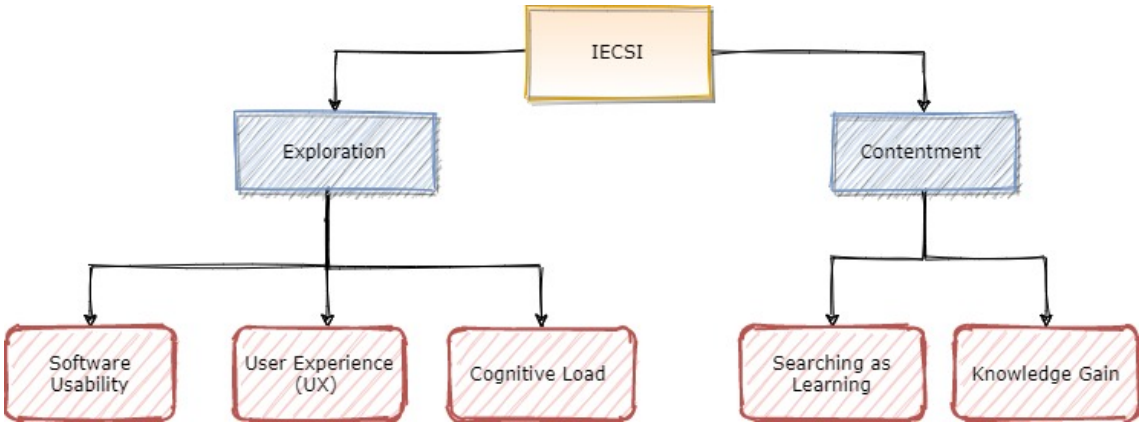


Figure 4.4: Components of the Implicit Evaluation of Conversational Search Interfaces (IECSI).

	Topics (0 (very low) - 7 (very high))
Search Formulation (Per-Search)	Background Knowledge Interest in Topic Anticipated Difficulty
Content Selection	Actual Difficulty Text Presentation Quality Average number of Docs Viewed per Search The Usefulness of Search Results Text Relevance
Interaction with Content	Cognitively Engaged Suggestions Skills System Understanding Input Average Level of Satisfaction
Post Search	Search Success Presentation of the Search Results Expansion of Knowledge after the Search Understanding about the Topic

Table 4.3: Characteristics of the search process [121] illustrating changes in knowledge structures.

4.3.1 Designing our Conceptual Framework

Our CS evaluation framework is divided into two segments: Exploration and Contentment, as shown in Figure 4.4. This section describes the use of standard questionnaires focused on multiple dimensions to form our Implicit Evaluation for Conversational Search Interface (IECSI).

4.3.1.1 Exploration Segment:

This segment focuses on exploring CS interfaces, and is classified into three components: Software Usability, User Experience, and Cognitive Load.

1. Software Usability: Usability is an important consideration for the evaluation of interactive software. IBM Computer Usability Satisfaction Questionnaires enable psychometric evaluation from the perspective of the user, and are known as the Post-Study System Usability Questionnaire (PSSUQ) Administration and Scoring [28]. The PSSUQ includes four dimensions: overall satisfaction score (OVERALL), system usefulness (SYSUSE), information quality (INFOQUAL) and interface quality (INTERQUAL), which includes sixteen param-

eters that would be asked through survey after the user used the system.

2. User Experience (UX): UX is measured using a questionnaire for interactive product known as User Experience Questionnaire (UEQ-S) [66, 132, 150]. This questionnaire also enables us to analyse and interpret outcomes by comparing with a benchmark dataset of outcomes for other interactive products. This questionnaire also provides us with the opportunity to compare interactive products with each other. UEQ-S contains two meta-quality dimensions: pragmatic and hedonic. Each dimension contains four different parameters as shown in Table 4.4. Pragmatic quality explores the user experience of a CS system. Hedonic quality explores the pleasantness of using the system.

3. Cognitive Load: An important consideration in the evaluation of CS interfaces is their impact on the user's cognitive load during the search process. To measure the user's workload, the NASA Ames Research Centre proposed the NASA Task Load Index [15]. This is a multi-dimensional rating procedure which provides a measurement of the overall workload during a process or event. This workload is classified into six subscales: mental, physical, temporal, own performance, effort and frustration. Out of these six dimensions, three are related to the demand imposed on the subject due to the task (mental, physical and temporal) and the remaining three to the interaction of the subject with the system (effort, frustration and performance). This implicit evaluation enables us to examine the cognitive load and cognitive engagement of the user while using a system.

4.3.1.2 Contentment Segment:

This segment focuses on information need satisfaction during the search process. It includes a questionnaire based on interaction while searching, learning during searching and knowledge gain arising from the search activity:

1. Searching as Learning: As discussed in Section 3.5.1, it is important to observe

	Negative	Scale	Positive
Pragmatic quality	obstructive	1 2 3 4 5 6 7	supportive
	complicated	1 2 3 4 5 6 7	easy
	inefficient	1 2 3 4 5 6 7	efficient
	confusing	1 2 3 4 5 6 7	clear
Hedonic quality	boring	1 2 3 4 5 6 7	exciting
	not interesting	1 2 3 4 5 6 7	interesting
	conventional	1 2 3 4 5 6 7	inventive
	usual	1 2 3 4 5 6 7	leading edge

Table 4.4: Scales Pragmatic quality and Hedonic quality.

Parameter	Definition
Dqual	Comparison of the quality of facts in the summary in range 0-3, where 0 represents irrelevant facts and 3 specific details with relevant facts.
Dintrp	Measures the association of facts in a summary in the range 0-2, where 0 represents no association of the facts and 2 that all facts in a summary are associated with each other in a meaning.
Dcrit	Examines the quality of critiques of topic written by the author in range the 0-1, where 0 represents facts that are listed without analysis and 1 where both advantages and disadvantages of the facts are given.

Table 4.5: Summary comparison metric based on the Wilson's work to verify the knowledge gain [96].

whether a CS system supports the user effectively in their engagement with the search system and enables the user's knowledge gain arising from the search process. To better understand this process, we decided to separately measure the factors of both user interaction and modification of their mental knowledge structures. We developed a questionnaire [152], inspired by Vakkeri's model of search as learning [53, 121] based on Table 4.3, to capture details of user interaction behaviour.

2. Knowledge Gain: To measure knowledge gain, the user is required to write a pre-search summary and a post-search summary about the search topic. This summary is manually evaluated by independent assessors (atleast two assessors) per summary in three sub dimensions: Quality of Facts (Dqual), Interpretation (Dintrp) and Critique (Dcrit), as shown in Table 4.5 [96].

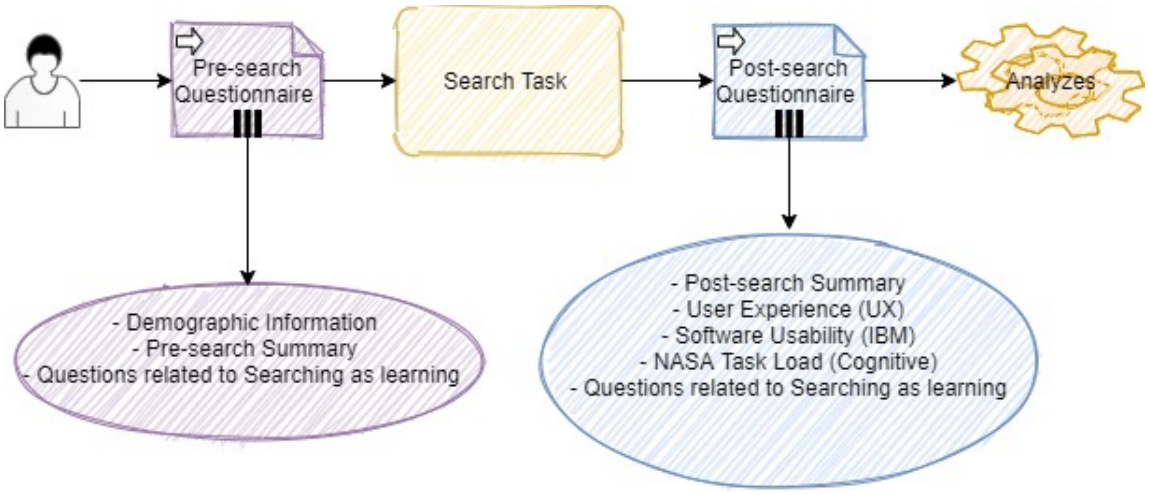


Figure 4.5: Evaluation process including pre-search and post-search questionnaires.

4.3.2 Developing the Evaluation Process

The overall evaluation process is shown in the Figure 4.5. Users complete a pre-search questionnaire and then a post-search questionnaire to analyse the knowledge expansion and user behaviour during the search. Both questionnaires are based on the metric described in Section 4.3.1, and as shown in Figure 4.6. To maintain uniformity, subjects rate each parameter on a 7-point Likert scale [103], where the scale ranges from 1 (very low) to 7 (very high) on each questionnaire. The evaluation is conducted from two perspectives: a) comparison of a conversational interface with a conventional search system, b) evaluating only a conversational interface based on a provided benchmark:

- **Comparison of conversational interface with conventional search system:** This evaluation method enables comparison of conventional and conversational search interfaces based on 5 dimension metrics, as described in Section 4.3.1. The user completes two search tasks one each using each search setting (conventional and conversational). For each task, the user completes a pre-search questionnaire and a post search questionnaire. This analysis is intended to provide better insights into the operation of a CS system and contrasting user opinions of each type of interface.

- **Evaluating CS interface based on benchmarks:** Most of the metrics

introduced in the framework have a standard benchmark. Pre-search and post-search questionnaire scores for each dimension of evaluation can be compared using their standard benchmarks. This not only provides an estimation of evaluation, but also provides an opportunity to explore the conversational interface with the standard system benchmark. Furthermore, this allows us to understand user expectation in general in all dimensions. This provides empirical measurability of a CS interface in the light of the benchmark. Moreover, this can help us to understand how far or close the current CS interface is from the user's expectations.

As per the above perspective, it is very important to analyse the data critically including use of statistical significance tests. If the results are significant, this can be used to develop a separate benchmark for the CS interface to assist other researchers in comparing their studies on CS interfaces.

4.3.3 Implementation and Analysis of the Framework

As noted earlier, the user is required to complete pre-search and post-search questionnaires. We developed these questionnaires (shown in Appendix B.1) by combining the dimensions introduced earlier in Section 4.3.1. The details of the pre-search and post-search questionnaires are described below:

1. **Pre-search Questionnaire:** This focuses only on contentment, and contains questions on demographic details of searcher, background knowledge of the searcher about the search topic, interest in the search topic, searcher experience of using conversational system.

2. **Post-search Questionnaire:** This focuses on contentment and exploration, and contains questions on knowledge gain after search, based on interactions (e.g., How many documents reviewed by user?), software usability, UX, cognitive load, etc.

	Metric	Pre-Search	Post-Search	Evaluation
Contentment	Searching as learning	User will be asked to fill the Questionnaire after reading the search task. This is to understand the state of user in the phase of search formulations in search process as shown in the Table 2.	User will be asked to fill the Questionnaire post search task. This is to understand the state of user in the phase of Content Selection, Interaction as shown in the Table 2. This questionnaire will also allow user to provide direct and indirect feedback about the search success, expansion of knowledge and interaction experience.	• To keep the uniformity, subjects can rate each parameter on a 7-point Likert scale where 1 (very low) and 7 (very high) on each questionnaire. • Some parameters are subjective.
	Knowledge Expansion	User will write a summary based on Pre-existing knowledge on search topic.	User will write a summary on search topic based on knowledge gain after search.	Summary will be coded by at least two independent assessors (Kappa Coefficient >.85) based on three categories such as Quality, Interpretation and Critiques as shown in Table 2.
Exploration	Software Usability		User will fill the Psychometric Evaluation for software from the perspective of the user By IBM discussed in tool.	• Evaluated using Sixteen as mentioned in Table (Likert scale 0-7). • To verify your system is as per the standard, the overall average is greater than 3.

Figure 4.6: Implicit evaluation for conversational search interface metrics including contentment and exploration phase.

Each question is evaluated based on a Likart score (0,7), except for the knowledge gain metric which is evaluated based on analyzing the questionnaire of contentment phase. As described, the framework is classified into two sections: exploration and contentment, as shown in Figure 4.6. Details are as follows:

1. **Exploration:** This questionnaire investigates user experience by exploring the system's satisfaction of the user's information need. As such, conversational interfaces are evaluated based on the post-search questionnaire. The mean score of each question is calculated based on the number of users. Analysis is conducted using both Quantitative Analysis and Qualitative Analysis.

   (a) Quantitative Analysis: This is based on the mean score, statistical testing (T test) is carried out based on the population and nature of the experiment. When comparing a conventional system and a conversational system, we are able to perform dependent significant testing (T test), since the population undertaking the experiment in both settings is the

same. If we are comparing the mean score of the conversational interface with a standard benchmark, we can conduct independent significant testing (T test). This statistical testing enables us to understand how systems differ. Additionally, each dimension discussed above in Section 4.3.1 has a standard tool or metric for analysis and comparing it with the benchmarks.

(b) Qualitative Analysis: The different dimensions are annotated based on the comparison of the means. A mean value between 2 and 4 represents a neutral evaluation of the corresponding scale (yellow dimension), a mean > 4 represents a positive evaluation (green dimension) and mean < 2 represents a negative evaluation (blue dimension). After comparing the mean, each question is annotated based on these dimensions. The dimensions are annotated by two independent analysts with the Kappa coefficient (Approx 0.85), then the dimension is counted for each section such as software usability, user experience [117, 155], cognitive load. As per the dimension, the aspects of the interface that need to be improved can be identified. For example, if software usability gets more red dimensions, then the interface needs to be improved with respect to software usability.

2. Contentment: The questionnaire to investigate contentment with respect to information need can be analyzed by comparing the user's pre-search knowledge and post-search knowledge. As discussed earlier, contentment evaluation is designed to investigate user learning while searching, and their knowledge expansion arising from the search process. The analysis can again be conducted using both Quantitative Analysis and Qualitative Analysis.

(a) Quantitative Analysis: Based on the mean score of search as learning and knowledge gain (the difference between post-search and pre-search summaries of each setting (conventional system and conversational system))

parameters, statistical testing (T test) can be applied.

(b) Qualitative Analysis: Search as learning questions, as shown in Table 4.3, are annotated, evaluated and analyzed based on different dimensions as discussed in the exploration Qualitative Analysis Section 1. Pre-search and post-search summaries can be compared based on the parameters discussed in Table 4.5. The summary is scored against all these factors by two independent analysts with the Kappa coefficient (Approx 0.85) [7]. For each parameter, the difference between pre-search and post-search summaries is calculated to capture the gain in knowledge about search topic after completing the search task.

4.4 Concluding Remarks

The concept of CS remains an ongoing topic of research. A crucial part of this work is the experimental methodology and its evaluation. Studies of CS to date have mainly been based on user experience. This overlooks interaction with the system and changes in the user's knowledge. In this chapter, we introduce a standard experimental methodology and evaluation framework that covers multiple dimensions to study the effectiveness of a CS interface.

In the next chapters, following the methodology discussed above, we design and develop information seeking processes through a standard search system, a smart speaker conversational agent and a novel MCSI, which is later evaluated based on the components of the IECSI framework.

Chapter 5

User Search Behaviour in Conventional Search

5.1 Overview

Conversational search presents opportunities to support users in their search activities to improve the effectiveness and efficiency of search while reducing their cognitive load. Any limitations in the competency of these conversational agents restrict the situations for which they can replace human intermediaries, such as librarians. In order to move towards the development of a system to enable conversational search of this type, we need to understand capabilities required of the agent in information seeking. To progress our understanding of these, we report a study examining the behaviour of users when using a standard web search engine, designed to enable us to identify opportunities to support their search activities using a conversational agent.

5.2 Motivation

One of the motivating objectives of conversational search is to support more natural interaction between users and content information archives via dialogue-based engagement. The ideal conversational interface might perhaps be seen as an auto-

mated surrogate for a human assistant with the competence to support search with the intelligence of such an assistant. An aspiration of this sort is, at least in the short to medium term, unrealistic. Search tasks requiring this sort of support will continue to need to engage human intermediaries to satisfy the searcher's information need. At this time, research in conversational search should seek to support IR activities where users work, or could work, unassisted with search applications. For example, to enable more effective or efficient engagement with web search engines, reducing the cognitive load on the searcher necessary to achieve their objective, while improving search effectiveness. The entity with which the user engages would be a surrogate for a less skilled human intermediary. The question then arises, in order to provide useful support to a user's search activities, what does the associated conversational agent need to be able to do? In agent terms, what *competence* does it need? How does the user learn to *trust* the activities of the agent?

Support for web search requires the conversational agent be able to handle queries relating an information need in any topic area of interest to the searcher. Information needs in these areas can be of different forms, ranging for example from straightforward factoid questions to more exploratory needs requiring consultation of multiple retrieved items. A further factor is the varying knowledge of the subject area of the searcher, and their consequential interest in different items from those available in the information archive.

In order to begin to specify the necessary functionality of a conversational search agent in this setting, we need to understand what users seek to do in order to satisfy their information needs, and to study their current actions as they try to achieve this. Models of information seeking, such as those of Belkin [11], Kuhlthau [18], and Vakkari [121], provide mechanisms for understanding the cognitive processes associated with satisfying a user's information need. Regardless of the interaction mechanisms they support, search applications seek to provide mechanisms to resolve information needs. The manner in which the user learns about the subject, or progresses in the resolution of their information need, will depend on the manner

of their interaction with the search application. Current applications rely on the user to create their search query, and to work with retrieved surrogates such as snippets and potentially with query reformulation and further searches, to address their information needs. The argument for conversational search is that it has the potential to better direct completion of search tasks to improve efficiency, reduce cognitive load or user effort, and potentially to enable more successful completion of search tasks, by providing proactive support, than users are able to do with current entirely user driven tools [152, 129, 171, 177].

In this chapter, we report a study examining users' search behaviours as they seek to satisfy predefined information needs specified in backstories using standard engagement with a commercial search engine. This study contrasts with other studies which have approached understanding the potential for conversational search by studying search by using two individuals with one taking he role of the searcher and the other the search agent intermediary (wizard) [143, 190]. The results of our study are designed to provide input into the design of a conversational search agent to support common web search tasks.

In this chapter, we have investigate RQ1 introduced in section 1.5.1 of Chapter 1.The answer to this question obtained is summarized in Section 9.2.1 of Chapter 9 in this PhD study.

The remainder of this chapter is structured as follows: Section 5.3 describes the methodology of our study, Sections 5.4 and 5.5 present results and analysis from our study, Section 5.6 summarizes the findings and finally Section 5.7 concludes and outlines directions for the next stage of our work.

5.3 Experimental Methodology

In this section we describe the details of our user study. The study aims to enable us to observe and better understand the behaviour of non-specialist searchers whose techniques for use of search engines are generally learned from personal experience. The objective of this study is to provide input into specification of potential conver-

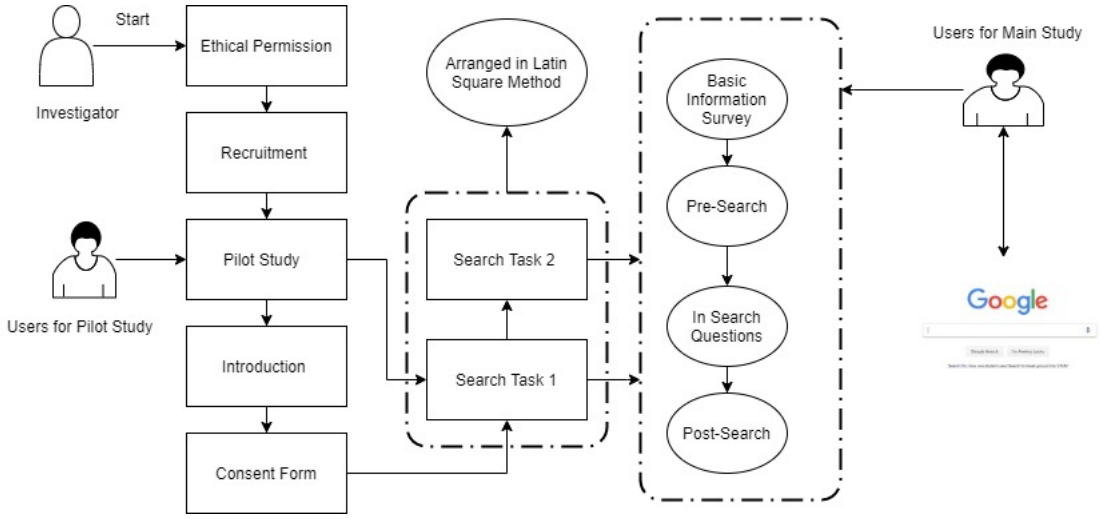


Figure 5.1: Schematic representation of experimental process including user engagement.

sational extensions to existing search applications. We seek to gain insights not only into how conversation engagement might directly be incorporated into current user search activities, but also where there are opportunities to enhance the user's search experience. The study aims not just to observe user behaviour in completing these tasks, but also to gain insights into their objectives at each point in their interaction with the search engine. In doing this, we hope to gain insights into the relationship between user actions and information seeking and learning, in order to be able to make use of this in the design of future conversational search agents.

5.3.1 Information Needs for Study

For our search tasks we wish to give searchers realistic information needs which could be satisfied using a standard web search engine. In order to control the form and detail of these, we decided to use a set of information needs specified within *backstories*. The details of these backstories were presented in Chapter 4.

5.3.2 Experimental Procedure

Participants in our study were required to complete a search session consisting of multiple search tasks based on our selected backstories. As part of their search

session they had to complete a questionnaire before, during and after undertaking each task discussed in methodology Chapter 4.

The procedures followed in this investigation as shown in Figure 5.1.

5.3.2.1 Experimental Setup

Participants used a setup of two computers arranged with two monitors side by side on a desk in our laboratory, as described in Chapter 4. Participants carried out their search tasks using the standard Google web search engine using a Google chrome browser. In addition, all search activities were recorded using a standard screen recorder tool to enable post-collection review of the user activities. The data provided in the Table 5.1 is for those participants whose data was used for analysis. Approval was obtained from the DCU Research Ethics Committee prior to beginning the data collection, details were given in Chapter 4.

5.3.2.2 User Behaviour Categories

In our study we were interested to analyze user behaviour with a standard user-driven search tool in terms of the interactions they make and to seek to understand the reasons for their behaviour. In doing this, we divided search behaviour into four categories:

- **User type A**: The user enters one query and selects one document from the retrieved list to fulfill the information need.
- **User type B**: The user enter one query and opens multiple documents as they seek to fulfill their information need.
- **User type C**: The user performs more than one iteration of querying and inspection of results in order to fulfill their information need.
- **User type D**: The user performs more than one search interaction, but only opens a single document. This behaviour may arise if the user has difficulty expressing their information need in a single query, leading to reformulations

and multiple runs, or if there is little relevant information available, but this suggests that they open more than one document, potentially one for each query.

- Additionally there is the possibility of the case where the user issues one or more queries, but does not select any of the retrieved items, this may indicate that either the user retrieves no relevant items, cannot identify retrieved relevant items or is able to satisfy their information need from one or more of the snippets displayed in the retrieved ranked list(s).

Experiment	Age	No. Male (M)	No. Female (F)	Ratio (F/M)
Chapter 5 (Conventional Search with backstory)	18-25	1	2	2:1
	26-35	6	6	1:1
	36-45	1	0	0:1
	Total	8	8	1:1
Chapter 5 (Conventional Search without backstory)	18-25	3	2	2:3
	26-35	8	5	5:8
	36-45	1	0	1:0
	Total	12	7	7:12

Table 5.1: Details of age distribution throughout this investigation.

5.3.2.3 Pilot Studies

A pilot study with two undergraduate students in Computer Science was conducted using two additional backstory search tasks to see how long it took them to complete the sections of the study, gain insights into the likely behaviour of participants, and to generally debug the experimental setup as discussed in Chapter 4.

User Behaviour	No. of Interaction	Average time Per Search (in minutes)	No. of Documents viewed	No. of Search task
User Type A	1	22	1	2
User Type B	1	21.7	>1	8
User Type C	> 1	34.6	>1	22

Table 5.2: Types of user behaviour.

5.4 Main Study: Part 1

Participants for the main study were recruited on a voluntary basis with the majority being postgraduate students studying either on MSc or PhD programmes. All participants were affiliated with our University or another public University in Ireland. A total of 17 subjects completed a search session. One participant was found not to have followed the instructions appropriately, and their contribution was excluded from analysis. Thus there were a total of 32 search tasks completed.

5.4.1 Information Seeking Behaviour

Table 5.2 summarizes the following features for the three classes of interactive search behaviour observed in our study:

1. **No. of interactions**: The number of queries used by the participant to complete the search task.

2. **Average time per search**: The average time take by a search with each class of interaction behaviour.

3. **No. of documents per interaction**: No of documents opened for a query.

4. **No. of Search tasks**: The total number of search tasks which follow the particular type of user behaviour.

Table 5.3 shows a summary of the information collected during the study. Based on our analysis of the data gathered, we make the following observations:

Search Stage	Variables	Type A	Type B	Type C
Search Formulation	Interesting Topic (1-5)	4	4	3.7
	Background Knowledge (1-5)	2.5	2.8	2.2
	Difficulty level (1-5)	2.5	2.1	2.9
Content Selection	Average of number docs viewed per search	1	2.5	4.4
	Why chooses these results?	-Top three -Relevance	-Top three -Authenticated source -Relevance	-Search recommend -Learn about topic -Relevance to search topic
Interaction with Content	Average Level of Satisfaction (1-10)	9.00	8.66	8.16
	Total Time Taken for search	22 minutes	21.7 minutes	34.6 minutes
Post-search	Understanding of the topics	6.5	8	7.6
	Expansion of knowledge a search	4	4	4

Table 5.3: Characteristics of search process with back story view.

1. **User type A**: Users showing this behaviour spent on average of 22 minutes studying a single document. This document was able to entirely satisfy their information need, and there was no opportunity to study learning behaviour from the examination of multiple documents. This behaviour was only observed for 2 of the 32 completed search tasks, in each case for a different backstory. Thus, the search behaviour was not determined by the requirements of the backstory. The existing knowledge of the topic for these searches averaged 2.5 out of 5. In this case the user exhibited sufficient existing knowledge of the subject to be able to use the provided document snippet and other metadata to identify a document which is able to satisfy the information need

beyond their existing knowledge of the subject [121].

2. **User type B**: The number of documents opened was on average 2.5. The average background knowledge of the topic of the information need was 2.8 out of 5. The average time spent viewing each document was approximately 9 minutes, which is less than half of the time taken by type A users examining their single document. Users thus spent an average of 21.7 minutes examining documents. Although more common than Type A behaviour, this behaviour was still not common, and we observed this behavior for only 8 search tasks.

   Type B users claim a slightly greater initial knowledge in comparison to the Type A users[1] which may encourage them to make greater exploration of the topic by examining multiple documents.

3. **User type C**: The average time spent on each search task was 34.6 minutes, with an average of 2.6 queries and 4.4 documents opened per search task. The average prior knowledge of the search topic was 2.2 out of 5. The average time spent on viewing each document was approximately 7.8 minutes, which is nearly one third of the time taken by type A users and less than the time taken on an average by type B users.

   Type C users on an average have less prior knowledge of the search topic than Type B and Type A searchers. It appears that this motivates them to carry out multiple search interactions with reformulated queries and to open multiple documents to satisfy their information need. Type C users follow all the paradigm of change in knowledge structure (restructuring, tuning and assimilation) following the search process [121]. This behaviour was by far the mostly commonly observed in our study being seen for 22 search tasks.

[1]It should be noted that is calculated in both cases based on very small numbers of search tasks.

5.4.2 Analysis of Search Activities

In this section we consider our results in terms of Vakkari's model of Search as Learning as discussed in Section 2.4.2. From the summary of the data provided by the participants in Tables 5.3 and review of the videos of the search sessions we can make the following observations.

1. **Type A searchers**: For type A searches, the user has a good knowledge of the topic of the information need, and despite the complex nature of the information need is able to form a query able to retrieve a single document which completely addresses the information, or at least does so to the extent that they regard the task as completed on the basis of reading this document. In other cases a similarly knowledgeable searcher could form an equally good query, but not be able to find a single document addressing the information need. Suitable documents were observed to come from sources such as Wikipedia or a specialist website relating to the topic of the information need.

   In terms of the searcher's knowledge structures relating to this information need, their prior knowledge of the topic means that they will already have a structured understanding of the topic, and that their cognitive activity will consist of tuning and assimilation of the information. The fact that the required information is contained in a single document makes the user's task easier, although the time taken to engage with the document to identify the necessary information means that the user is still observed to undertake a large amount of work. With respect to conversational search, the question arises of whether an agent could improve the efficiency with which the information is accessed.

2. **Type B searchers**: The time spent with each document is less than the single document in type A searches, which suggests that users are able to identify the relevant material relatively quickly. The overall time for a task is taken up by the process of accessing detailed information from across the opened

1st Interaction
User Query 1: `folk remedies for sore throats`
No of Documents viewed: 1
Doc Title: `10 Natural Home Remedies for Sore Throat - Global Healing Center`
2nd Interaction
User Query 2: `folk remedies to help soothe a sore throat`
No of Documents viewed: 1
Doc Title: `22 Natural Sore Throat Remedies to Help Soothe the Pain`

Figure 5.2: Queries issued in one search session for the example backstory shown in Chapter 4.

documents, and to interpret it for tuning and assimilation in their learning of the knowledge relating to the task. Their prior knowledge of the topic again means that the cognitive activities will focus on tuning and assimilation, but in this case it is likely to be more demanding since it is spread across consultation with multiple documents. In this case, this is a more demanding activity since this takes place across multiple documents. Again the follow on question to these observations, is to consider whether a conversational agent might be given the competence to improve the efficiency of the user's engagement with this information to satisfy their information need. There would appear to be scope here for a conversational agent to assist the searcher in their engagement with snippets and documents as they seek to address their information need.

3. **Type C searchers**: Participants on average show less knowledge of the topic of the search task. Figure 5.2 shows the queries issued and summary responses in one of the search sessions for the backstory shown in Figure 4.1. To satisfy the information, these searchers needed to issue more than one query and to engage with documents retrieved across more than one query. These findings are consistent in that the searcher's initial query may not be sufficiently well informed to enable it to retrieve all the required information in a single run. It may also be that the multi-faceted nature of some of these search tasks means that no single query was able to obtain all the required information.

User Behaviour	No. of Interaction	Average time per Search (in minutes)	No. of Documents Viewed	No. of Search Task
User Type A	1	23.2	1	5
User Type B	1	22.7	>1	11
User Type C	> 1	33.3	>1	22

Table 5.4: Types of user behaviour without backstory view.

We analyzed this point in more detail in our study. The learning activities in this case are likely to be more of knowledge involving a combination of restructuring, tuning and assimilation as the searcher progresses through the completion of the search task. In completing a search task, these participants are essentially engaging in a limited dialogue with the search engine. In addition to the exploration of conversational opportunities for the search considered for Type A and Type B searches, for Type C, we can also consider how an agent might support the search through the multiple query phases.

Examination of the contents of both the initial and reformulated queries showed that most of the content or topic related words contained in them are found in the written backstories. A notable point related to the contents of the reformulated queries is that they contain very few occurrences of words found in documents opened in the previous search. On consideration allowing participants to keep the written backstory visible while carrying out the search does not represent the operational situation for many web searches where the user relies on recall of related words to form queries, and may be more reliant on words from retrieved documents in the reformulation.

Thus, we repeated our study providing participants with the backstory, and then removing it from their view while carrying out their search. We report results and analysis of this further study in the next section.

Search Stage	Variables	Type A	Type B	Type C
Search Formulation	Interesting Topic (1-5)	3.8	3.5	3.5
	Background Knowledge (1-5)	2	1.9	2.3
	Difficulty level (1-5)	2.6	2.7	2.6
Content Selection	Average of number docs viewed per search	1	3.7	4
	Why chooses these results?	-Top three -Relevance	-Top three -Authenticated source -Relevance	-Search recommend -Learn about topic -Relevance to search topic
Interaction with Content	Average Level of Satisfaction (1-10)	9.00	8	8.18
	Total Time Taken for search	23.2 minutes	22.72 minutes	33.4 minutes
Post-search	Understanding about the topic	7	7	7.6
	Expand of Knowledge After search	3.8	3.7	3.9

Table 5.5: Characteristics of search process without backstories view.

5.5 Main Study: Part 2

A total of 19 participants completed the second study, and completed 38 search tasks. 11 subjects participated in both studies. These participants were assigned different search tasks in the second study. Using some of the same participants in both studies allowed us to compare their behaviour across both experimental conditions.

Search condition	Same search behaviour in each session	Different search behaviour in each session
With backstory	8	8
Without backstory	12	7

Table 5.6: Types of search behaviour with search session.

5.5.1 Information Seeking Behaviour

Results of our second study with the revised backstory condition are shown in Tables 5.4 and 5.5. Analyzing the results in terms of Type A, Type B and Type C behaviour, we observe the following:

1. **User type A**: Participants spent on average of 23.2 minutes studying a single selected document. They were observed to spend more time reading the document in the new condition without the backstory. This behaviour was only observed for 5 of the 38 completed search tasks, in each case for a different backstory. The existing knowledge of the topic for these searches averaged 2 out of 5.

2. **User type B**: The number of documents opened was on average 3.7 which is higher than in the maintained backstory condition. The average time spent viewing each document was approximately 6.1 minutes, which is less than the time taken by Type A users examining their single document. Participants thus spent an average of 22.7 minutes examining documents which is also longer than in the backstory condition. This behaviour was observed for 11 search tasks. The background knowledge was 1.9 out of 5 which is below average, and again slightly less initial than Type A users, which may encourage them to read more documents.

3. **User type C**: The average time spent on each search task was 33.4 minutes, with an average of 3.1 queries and 4 documents opened per search task. The average prior knowledge of the search topic was 2.3 out of 5. The average time spent on viewing each document was approximately 8.3 minutes, which is nearly one third of the time taken by Type A users in both conditions and

more than the time taken on average by user Type C in condition 1 and type B on condition 2. The time taken by C user in condition 2 is less than time taken by User B in condition 1. Type C users have on average more prior knowledge of the search topic than Type B and Type A searchers in condition 2.

This behaviour was the most commonly observed in our study being used for 22 search tasks. In the case of Type C users, It was observed that Type C behavior was different in second condition with the backstory removed.

We summarize the overall findings from this study in Tables 5.3 and Table 5.5. We observe different search behaviour in the search sessions shown in Table 5.6.

From these results and review of the videos of the search sessions we can make the following observations.

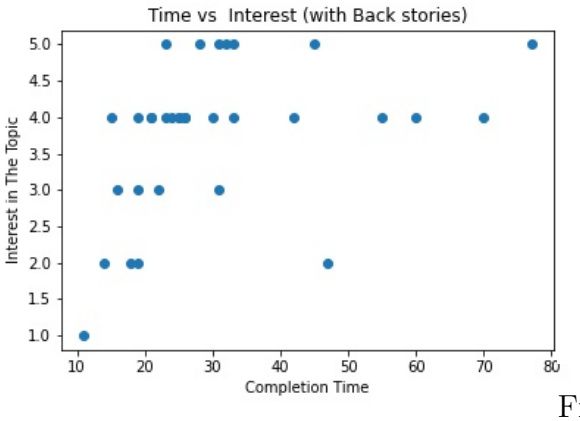


Figure 5.3: Correlation graph of completion time vs interest (Correlation coefficient 0.37) (with Back story)

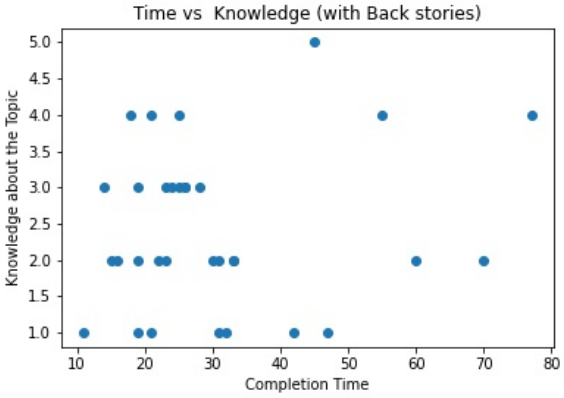


Figure 5.4: Correlation graph of completion time vs background knowledge (Correlation coefficient 0.07) (with Back story)

1. **Type A searchers**: We observe differences in the behaviour of Type A searchers in both conditions. Type A searchers in the second condition show less interest, less knowledge and greater difficulty in completing the task than Type A searchers in the maintained backstory condition.

2. **Type B searchers**: The time spent with each document is less than the single document in Type A searches, which suggests that searchers are able to

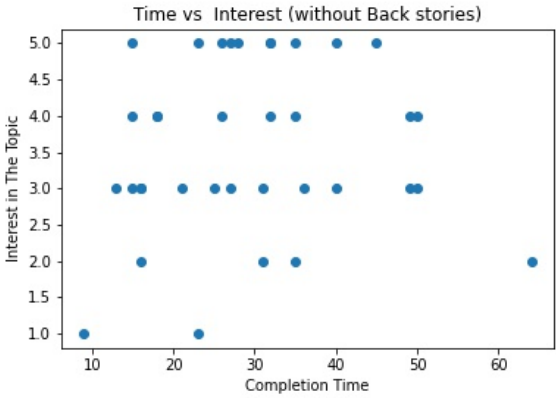


Figure 5.5: Correlation graph of completion time vs interest (without Back story) (Correlation coefficient 0.09)

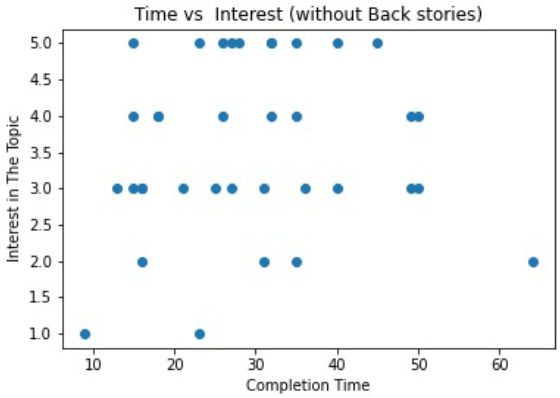


Figure 5.6: Correlation graph of completion time vs background knowledge (without Back story)

identify the relevant material relatively more quickly in Type B condition 1 and Type A users in both the conditions. The Type B user in condition 2 refer to more documents with the less average reading time of each document. This shows type B users were examining more documents and spending less time reading each one. This also reflects that the user might be able to structure and tune their knowledge by shallow reading of multiple documents. This behaviour direct us towards a conversational approach of IR.

3. **Type C searchers**: Type C users in the removed backstory condition spent less time reading each document in comparison to other user behaviours in both conditions. Type C users claimed to have very good background knowledge with above average interest in the topic. On average type C searchers viewed four documents per session.

We have also calculated the correlation between background knowledge vs task completion time and task completion time vs interest of the user in the backstory as shown in Figures 5.3, 5.4, 5.5 and 5.3 in both the settings. We did not observe any significant correlations in the variables.

5.6 Summary Analysis

In summary, the reformulated queries for the second and subsequent queries in the Type C tasks make much more use of words appearing in retrieved documents rather than those in the written backstories. This is not really surprising since the retrieved documents provide the searcher with a ready source of topically relevant words which they can choose from to improve or revise their query statement.

With respect to conversational search, this finding is interesting because we could consider incorporating a conversational agent in the analysis of the retrieved documents and selection of words or the automated reformulation of the query. This could go beyond recommendation of words for query expansion in current IR systems with the agent taking a more proactive role.

Ideally the agent should consider the features of the content actually engaged with by the searcher, their original query and recommended expansion words which they select, to determine the searcher's knowledge of the topic under investigation and their interests within this topic, and use this to support the search.

By taking on tasks currently driven by the user, such as examination of documents to determine how to reformulate queries, and actively personalising the behaviour of the search tool, taking account of the searchers subject knowledge and interests, a conversational agent should be able to reduce the overall load on the user in completing their search task. Information seeking models can provide a basis for a structure for the behaviour of such agents to support search, taking account of the behaviour of the searcher as they learn about the subject under consideration while they work to address their information need.

Multiple classes of user task completion behaviour action sets were observed and their potential scope for incorporating with conversational search can be stated as follows:

- Type A: Issue single query, open singles document, and spends much time engaging with it. Conversational system could support user exploration of the document to reduce time and complexity of satisfying information need.

- Type B: User issues single query, and then engages with multiple retrieved documents to satisfy information need. A conversational agent might assist in identifying useful content found in these documents in an integrated way to improve efficiency of satisfying the user's information need.
- Type C: User issues initial query and one or more reformulated queries and consults multiple retrieved documents. Conversational agent could assist with:
  - reformulating queries.
  - displaying potentially relevant retrieved content.

5.7 Concluding Remarks

In this chapter, we studied user search with a standard commercial search engine for cognitively demanding search tasks, to inform the development of conversational search systems for use in these tasks. We experimented with two conditions, with a back story shown during search and with the back story removed. We hypothesised interactive information-seeking behaviour into four types. We observed the first three types in our study. Analysis and observation of the user's behaviour for these three types can broadly be explained in terms of Vakkari's learning model of search in which searchers develop, refine and assimilate knowledge gained within the search.

The next chapter presents the study to explore the intelligent conversational speaker's potential in information seeking progress. Thus, we directed our investigation into understanding user search behaviour in a conversational setting, which could offer more insight into understanding the need for conversations. The next chapter also explores the potential of embedding dialogue strategies in the exploratory information seeking process via smart speaker.

Chapter 6

Examining the Potential for Conversational Exploratory Search using a Smart Speaker Digital Assistant

6.1 Overview

Online digital assistants, such as Amazon Alexa[1], Google Assistant[2], Apple Siri[3] have become very popular in recent years. These applications provide a range of services to their users, a key function among these services is their ability to satisfy user information needs from the sources available to them. Many users regard these applications as providing search services similar to online search engines. However, while it is clear that they are often able to answer factoid questions effectively, it is much less obvious how well they support exploratory type search tasks. Since these applications represent an important state-of-the-art conversational search application, in this chapter we examine the search behaviour of a popular online digital

[1] https://www.amazon.co.uk/b?ie=UTF8&node=14100223031
[2] https://assistant.google.com/platforms/speakers/
[3] https://www.apple.com/siri/

assistant.

We describe an investigation examining the behaviour of the standard Amazon Alexa for exploratory search tasks. Our user study shows that Alexa often provides poor support for these types of search task. We then propose extensions to Alexa designed to address these shortcomings. Our Custom Alexa application extends Alexa's conversational functionality for exploratory search. A further user study shows that our extended Alexa application enables users to more successfully complete exploratory search tasks, and is also well accepted by our test users.

In this chapter, we have investigate RQ2 and partially RQ3, as introduced in section 1.5.2 and section 1.5.3 of Chapter 1. The answers to these questions obtained are summarized in section 9.2.2 and section 9.2.3 of Chapter 9 in this PhD study.

6.2 Motivation

Conversational digital assistants embedded in smart speaker applications offer users a range of services including simple command and control of networked smart home appliances, accessed through conversational engagement. Digital assistants are often demonstrated using requests such as fetching recipes or latest weather forecasts. While it is clear that they are often able to address such requests, which are generally satisfied by single items or factoids, it is much less clear how well current search applications support more exploratory information needs, and what additional functionality might be required to address any identified shortcomings.

While conventional IR systems, such as web search engines, rely on the searcher's ability to browse retrieved content in an efficient manner, smart speaker systems are largely driven by spoken interaction, sometimes involving multi-modal output. User access to returned information in spoken form has a much lower bandwidth than visual review of textual. This suggests that digital assistants must select information to be returned in spoken form with higher precision than is the case for conventional IR systems. While this is straightforward when providing the user with answers to factoid questions, it is much more complex when considering more exploratory type

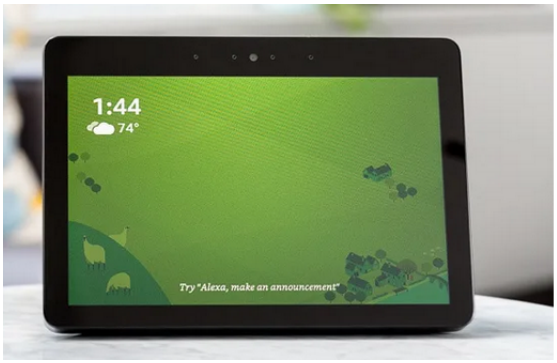


Figure 6.1: Amazon Alexa Echo interface.

search tasks where the searcher may need to interact with multiple longer items.

One way to limit delivery of extraneous information is to partition the search process into smaller incremental tasks where the searcher engages with the digital assistant using a conversational search process [129].

While in some contexts spoken only engagement is possible, the operational platforms of many digital assistants enable some form of multi-modal interaction. For example, smartphones, tablets and dedicated platforms such as the *Amazon Echo Show*. While the incorporation of visual presentation of information in these contexts reduces the problems associated with limited bandwidth due to speech only information transmission, their often small size and user expectation of search efficiency means that we should probably still seek some form of iterative incremental search engagement. For our study we adopt the Amazon Alexa assistant operating on an *Amazon Echo Show* platform[4] as shown in Figure 6.1.

The Echo Show enables conversational interaction with the Alexa assistant, and incorporates a tablet sized screen to enable multi-modal engagement. In the remainder of this chapter we first introduce the Amazon Alexa application, and then describe our study of complex search using Amazon Alexa based on the result of this study, we then introduce, develop and evaluate our Custom Alexa application.

[4]https://www.amazon.co.uk/amazon-echo-show-5-compact-smart-display-with-alexa/dp/B07KD7TJD6

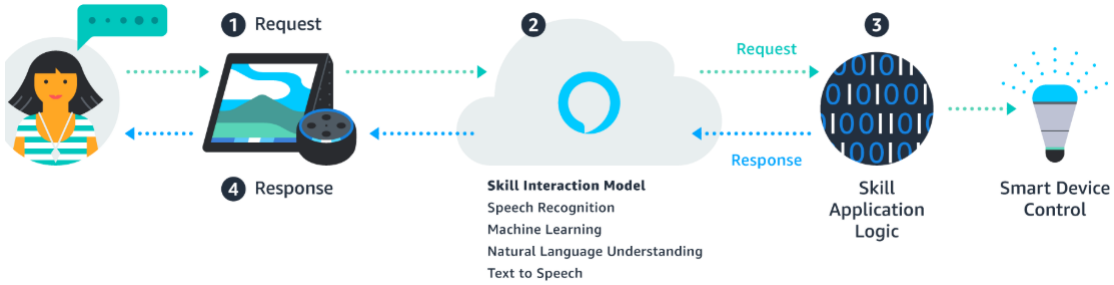


Figure 6.2: Alexa workflow [125, 185].

6.3 Search Using Amazon Alexa

In this section we begin by outlining the features of the Amazon Alexa application, and then introduce the search tasks used for our investigation of its use as a tool for exploratory search.

6.3.1 Amazon Alexa

The Amazon Alexa digital assistant provides a wide range of information provision services and control of applications to users. Alexa can operate on a range of dedicated hardware platforms including Amazon Echo, Amazon Echo Show, Amazon Dot, and related hardware, as well as an application running on more general platforms. Alexa performs voice-operated functions while communicating through a local WiFi Internet connection or other wireless connection with Amazon's AWS cloud servers, or other networked devices, to carry out these functions [125, 185].

The workflow of a standard user engagement with Alexa is divided into four steps: receiving a spoken instruction or request, interaction mode (responsible for speech recognition and *intent* identification), *skill* application logic (action after triggering the intent), and response, as shown in Figure 6.2. Where an *intent* is defined by Amazon as actions that fulfill spoken requests from the user, and a *skill* is an application which enables Alexa to perform an operation. A key feature of Alexa as a research tool is that new skills can be created to enable Alexa to perform new or extended operations [162]; it is for this reason that we chose to use Alexa for our investigations. Specifically we base our study on the use of an Amazon

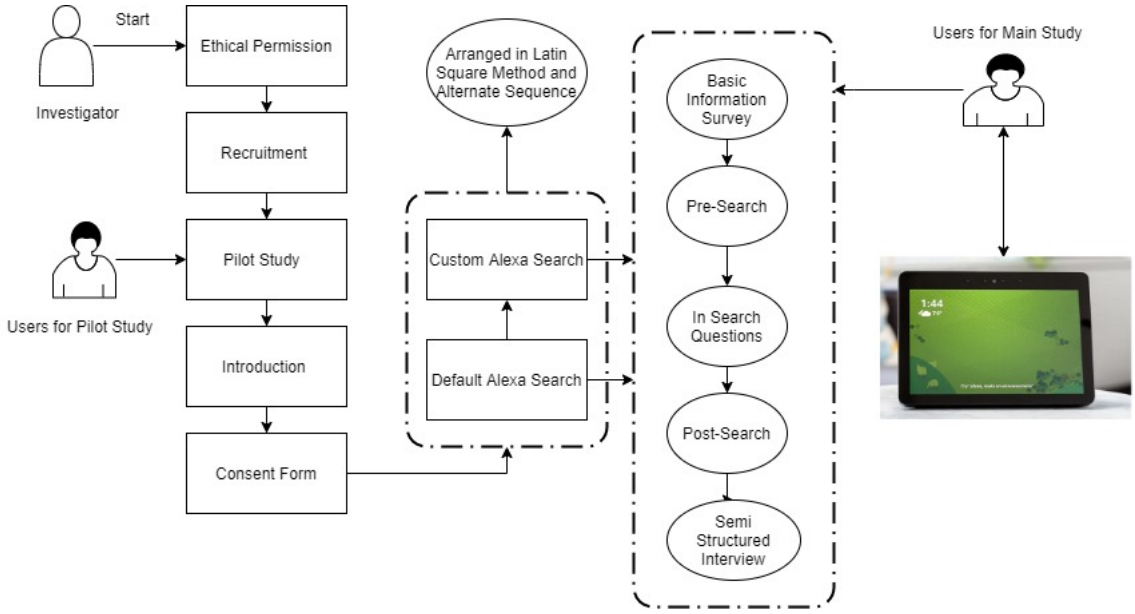


Figure 6.3: Procedure of the Alexa information seeking study.

Echo Show device which combines spoken interaction with the availability of a high quality 7-inch touchscreen display which can be used within applications to enable multi-modal interaction. Technical details of the Echo Show can be found in [141].

6.3.2 Exploratory Search Tasks with Alexa

In this study we first examine the ability of the standard Alexa assistant to support exploratory search tasks using Alexa's default conversational interaction features. As a source of information needs for our study, we provide participants with backstories requiring information. Details of the backstories expressing information needs of this sort were presented in Chapter 4.

6.3.3 Experimental Procedure

Participants in our study had to complete a structured search session as shown in Figure 6.3. They were given printed instructions for their search session. At the start of the session, they were provided with an opportunity to familiarize themselves with using Alexa for 5-10 minutes before starting the main study. Each participant had to complete one search task using an assigned backstory. They were given the printed backstory to study before they began the search. Following on from

our experience with use of backstories in the investigation described in Chapter 5, this was then withdrawn to prevent them from simply copying the details from the backstory as the basis of their query. The search session included completing questionnaires before and after carrying out the search task. The questionnaire included asking about the participant's expectations and experience of the search process, and writing a short summary relating to the topic of the backstory before and after carrying out the search task. After completing the search they were also required to attend a semi-structured interview to gather details of their experience of the search activity.

While participants completed the search tasks using the Echo Show, the questionnaires were completed online using a standard desktop computer. All search activities were video recorded for post-collection review of the users activities. The data provided in the Table 6.1 is for those participants whose data was used for analysis. Approval was obtained from the DCU Research Ethics Committee prior to beginning the data collection, details were given discussed in Chapter 4.

Search tasks were evaluated by analyzing the self reported questionnaires, interviews and the recorded videos. All the details from the interviews and video recordings were assigned to response categories by independent analysts, as discussed in Chapter 4. Some example of categories are helpful, relevant, easiness, etc. The sample coding example is provided into appendix in Table B.3.3.

Experiment	Age	No. Male (M)	No. Female (F)	Ratio (F/M)
Chapter 6 (Conversational Agent both settings)	18-25	14	3	3:14
	26-35	8	6	3:4
	36-45	0	0	NA
	Total	22	9	9:22

Table 6.1: Details of age distribution throughout this investigation.

6.3.3.1 Pilot Studies

Prior to the main study, a pilot study was carried out by two undergraduate students in Computer Science using two additional backstory search tasks, as noted in Chapter 4.

6.4 Behaviour of Alexa for Non-factoid and Exploratory Search

Each participant in the main study was assigned one of the selected backstories from the 12 tasks. Backstories were arranged by latin square method to avoid the learning effect. In total 33 subjects participated in the experiments of which results for 2 subjects were not included for analysis due discrepancies in their data. This study examined the user expectations of the Alexa assistant to support exploratory search and their experiences when they sought to use Alexa to address these information needs.

The following research questions were investigated during the study.

1. What are the challenges and opportunities to support exploratory conversational search using Alexa?
2. What characteristics of Alexa prevent it from functioning as an effective tool for complex information seeking?
3. What are the main expectations of users for conversational search systems?

6.4.1 RQ1: What are the challenges and opportunities to support exploratory conversational search using Alexa?

6.4.1.1 Challenges

Attempting to use the default Alexa assistant to address the exploratory information needs expressed in our backstories led to considerable user frustration with poor

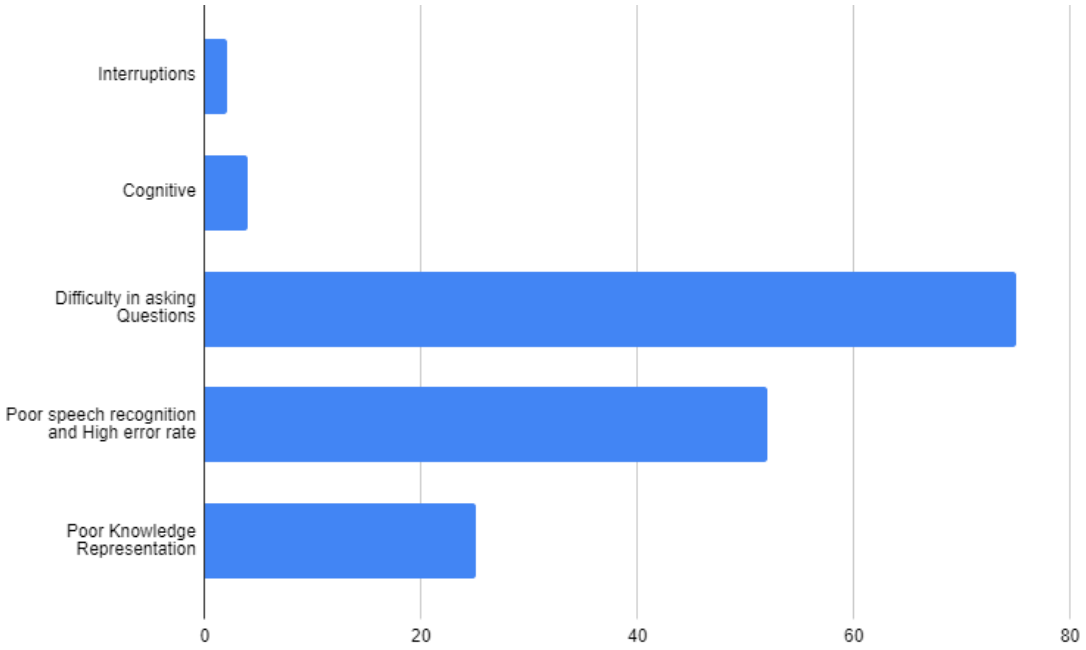


Figure 6.4: Limitations with Default Alexa.

success in addressing the information needs. From analysis of user feedback we identified the following challenges.

1. Task Success: For 62% of search tasks, either Alexa did not provide a response or gave irrelevant answers for the user query.

2. User frustrations and feedback: It was found that on average participants took approximately 5.5 minutes with an average 14.1 interactions (turns) to complete their search task. Feedback from the user surveys clearly indicates high levels of user frustration due to which the user ended their search early.

3. Major limitations of Alexa in exploratory search:

   A number of limitations were identified from observation of user interactions and their feedback. We divide these into five broad categories as shown in Figure 6.4.

   (a) Limited search support: From the responses to queries provided by Alexa, it became clear that Alexa only supports either fact based answers or simply starts reading from the beginning of a long Wikipedia document.

(b) Poor speech recognition and high error rate: While not directly related to its search capabilities, around 52% of participants noted that they experienced frustration arising from poor speech recognition and high word error rates while interacting with Alexa. For example, one participant noted that *"Alexa was not able to understand my voice and its frustrating and tiring to ask same thing again and again"*. Such errors can result in problems of participants being able to frame their desired query as they wish to, and lead to Alexa making mistakes in interpreting the query correctly, leading to incorrect responses.

(c) Difficulty in asking questions: This was the most important reason identified by participants, 75% of whom indicated that they had difficulty in creating queries. They were unable to search effectively since they had no background knowledge about the subject. Alexa offered no formal support to them in forming queries, and Alexa's answers were too precise to enable the searcher to build their knowledge of the subject [121] as they might with a standard web search engine.

(d) Others: A number of other factors were identified, The key ones are "Interruptions" (2%) and "Cognitive Load" (4%). Participants observed that they were unable to complete entry of queries due to interruptions by Alexa. In these cases, Alexa took a partially completed query as finished, and interrupted in the middle of process of entering the query providing results which may confuse the searcher or force them to repeat or reformulate the query.

6.4.1.2 Opportunities

Our investigations using the standard Alexa application highlight some critical areas which provide opportunities to improve exploratory search using Alexa.

1. Background knowledge support and effective knowledge representations: The Alexa application provides fact based answers, but does not support the user

to learn and refine their search. After conducting this study, we propose that the user should be provided with relevant information related to their search query based on facts which could help them to explore and analyse, to improve their ability to form effective queries to continue exploration of the topic.

2. Priming, dialogue-driven approach and interactive search process: For the standard Alexa, we observe that the search was not able to actively engage the user in the search process. Moreover, the efforts made by the searcher to identify and explore documents to satisfy their information need were very frustrating. To reduce the frustration and enhance search effectiveness, we propose to introduce a dialogue driven approach to the search process.

6.4.2 RQ2: What characteristics of Alexa prevent it from functioning as an effective tool for complex information seeking?

6.4.2.1 Fact-based Search

Current conversational assistants such as Alexa are designed to answer fact-based questions, like a traditional question and answering system, in contrast to traditional IR systems, which enable users to perform exploratory search operations.

6.4.2.2 Poor Information Seeking Strategy

Our investigation found that Alexa did not support interactive information seeking. This was reported by more than 45% of participants, as shown in Figure 6.5. The conversational agent clearly has not been designed to follow typical information seeking strategies to help the user. The information seeking strategies of this sort include two important factors: exploration and learning, which can be further subdivided into acquiring knowledge, interaction with information sources, engagement with information sources, comparing, reasoning, analysing evaluation, discovery, planning and forecasting [50].

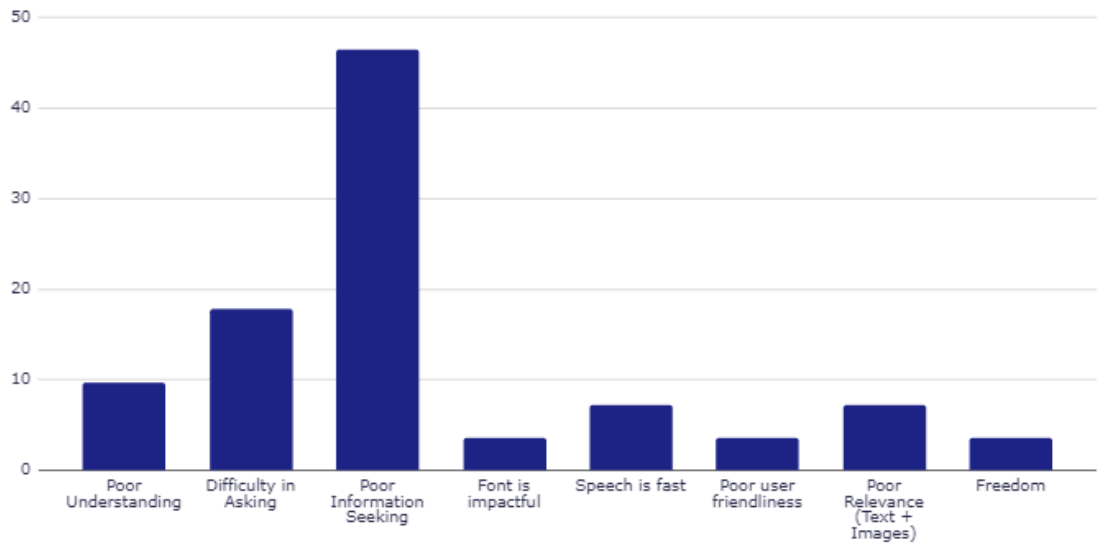


Figure 6.5: User feedback from search experience with Default Alexa for exploratory information seeking tasks.

As shown in Figure 6.6, the three major reasons for poor information seeking in standard Alexa are reported to be: lack of background knowledge (due to which the user was unable to create the right query), Alexa not being able to correctly recognise the user's query, and poor representation of knowledge by Alexa.

6.4.3 RQ3: What are the main expectations of conversational search systems?

From our study we found five major expectations of our participants for search in conversational systems.

6.4.3.1 Exploratory

A conversational search system should provide a broad information space to the user to provide the opportunity for them to explore a space of relevant information and to narrow the exploration to focus on addressing their information need. For example, one user said *"It is difficult to come up with question (query), I didn't realize what I wanted to hear, It would be better, I can explore the information or select the options of my choice"*

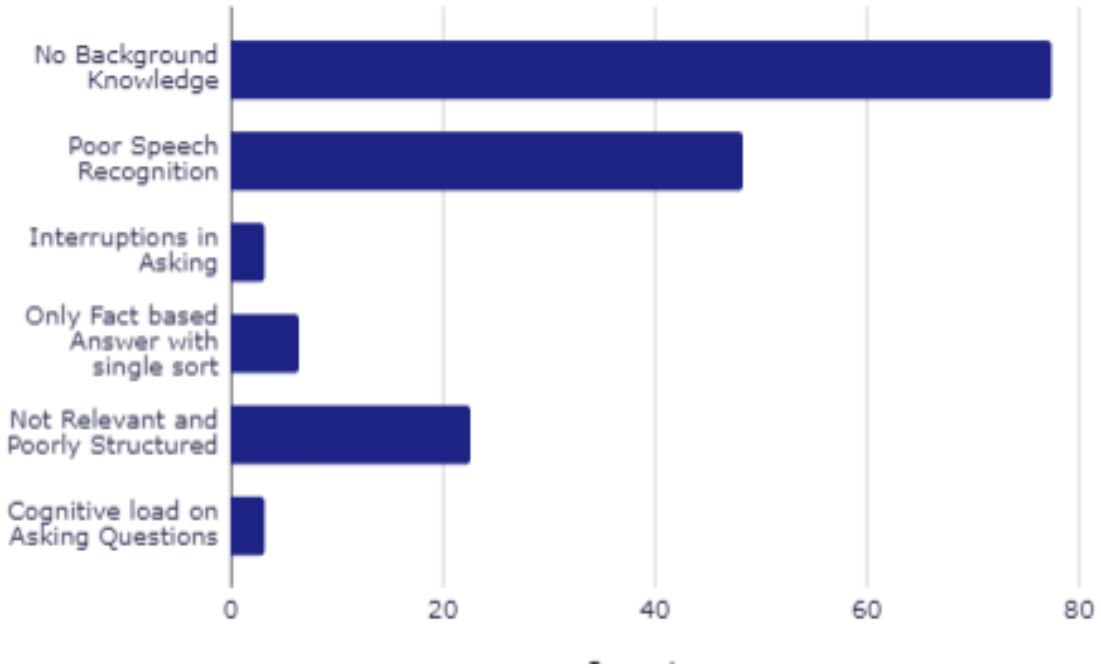


Figure 6.6: Cause for poor information seeking with Default Alexa.

6.4.3.2 Content Selection

In general, the standard Alexa system presents very limited information to the searcher in response to a query. The information returned by Alexa was often not suitable, and in many cases the searcher had to issue more queries in order to even begin to address an exploratory information need.

Our investigation showed that the important variables with respect to the user experience while attempting to satisfy their information need were as follows: the average number of interactions, the number of successful interactions, the number of unsuccessful interactions, the average time to complete a search process, and the quality of the presented text. Based on our results, we can conclude that the average interaction failure rate is around 62%, which is very high. We can see from these figures that this engagement is very inefficient, as a result of this the user is struggling to learn since most of the interactions result in failure, leading to demotivation and frustration of the user.

6.4.3.3 Content Interactions

Searcher interactions include use of multiple Alexa skills including: navigation skills, presentation skills and the speech skills of the conversational agents.

1. Navigation skills: A conversational search agent should support the user in navigation through the information space of the documents.

2. Presentation skills: The user expects presentation in different modes. More than 85% of users considered a combination all three dimensions (Text, Speech and Images) to be the best means of presenting the information, with only around 6.5 % of the users believing that satisfying information needs with speech and text only to be sufficient.

3. Speech skills: The speech skills can be classified based on multiple parameters, including speech speed, speech recognition, interruptions, speech content and its length. As shown in Figures 6.5 and 6.6, the searchers expected a speech system to support standard speaking rate, query length, less interruption by Alexa in the middle of conversation and good speech recognition accuracy.

6.4.3.4 Information Representations

In our study, we found that 18% of the searchers reported that the information represented during the search process (in default setting) was poor. As shown in Figures 6.5 and 6.6, the searchers felt that the information shown was poorly structured and not relevant to the task. The searchers were not satisfied with the combination of text and images. They also found difficulty in maintaining the contextual information flow during the task. We observed that the information representations can be refined based on various parameters: length of the text (optimal), relevance of the information to the query, structure of presentation, flow of conversation and information.

6.4.3.5 Conversational Properties

The study suggests properties that the participants indicated every conversational search system should have.

1. On boarding: This is the initial interaction in which a new user is introduced to the system in which it explains its competencies.

2. System as teacher: The user expects a system should ease their interactions by revealing its capabilities and essentially teach the user how to use it. In our study we observed that users who have previous experience with the conversational application interacted with it for longer (7.2 minutes) than users who using it for the first or second time (5.8 minutes).

3. Incite: We observed that most of the interactions were one way with the system unable to engage in useful dialogue with the user. However, ideally a conversational application should engage in back and forth dialogue with the user to assist them in reaching their goal.

4. Diverge flow and course corrections: A conversational application should be robust. As such, it should be able to handle any unexpected entries from the user, and make use of this input to guide the user towards their goal.

6.4.4 Dialogue Strategies to Support Content Engagement

As a result of our initial pilot investigations with the standard Alexa application, we sought to develop a revised a dialogue strategy for the Alexa assistant with the goal of improving its ability to support exploratory search. We refer to this revised Alexa application as it Custom Alexa search, which incorporates Duck Duck Go and Wikipedia API. This is designed to enable a user to carry out exploratory interactive search with Alexa as shown in Figure 6.7 where the user is interested in finding about "Mahatma Gandhi" by entering a query ("Who is Mahatma Gandhi?") and Alexa understands the intent of the user ("Search"). It takes the query and passes it to

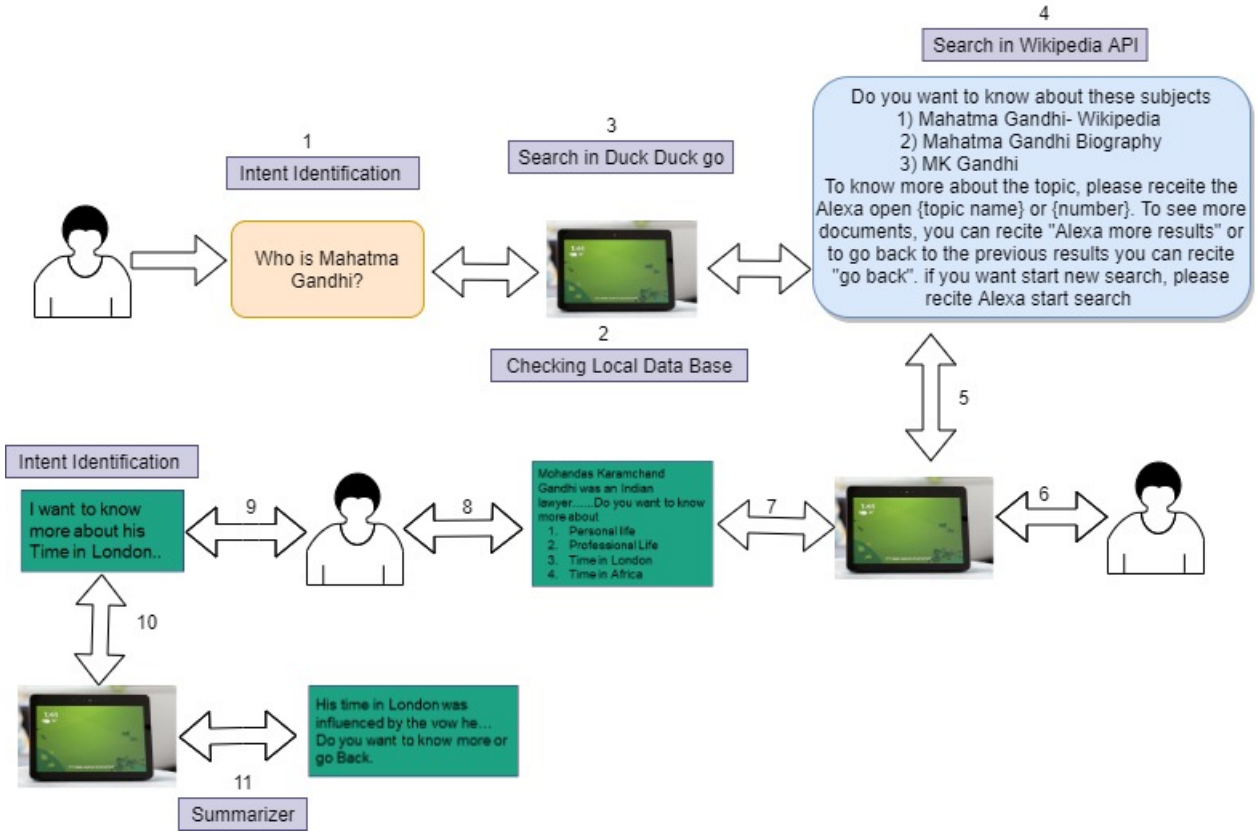


Figure 6.7: Flow of the Custom Alexa search in exploratory information seeking.

the Duck Duck Go API and asks the user to select any title of interest from the retrieved results or to ask for more retrieved results. Once the user chooses the title "Mahatma Gandhi" then the same title searches in Wikipedia, and shows the introductory summary from the Wikipedia page and presents the subsection titles of the documents, from which the user can select and explore further.

We implemented this as an Alexa skill, and then deployed it as a prototype using the Amazon Echo show. As skill is a custom function defined for Alexa to the specific propose. The Custom Alexa skills were developed iteratively using a series of pilot studies with informal feedback from participants prior to the formal evaluation. We investigate its effectiveness using a study following the same experimental setup as used in our exploration of the standard Alexa application for exploratory search.

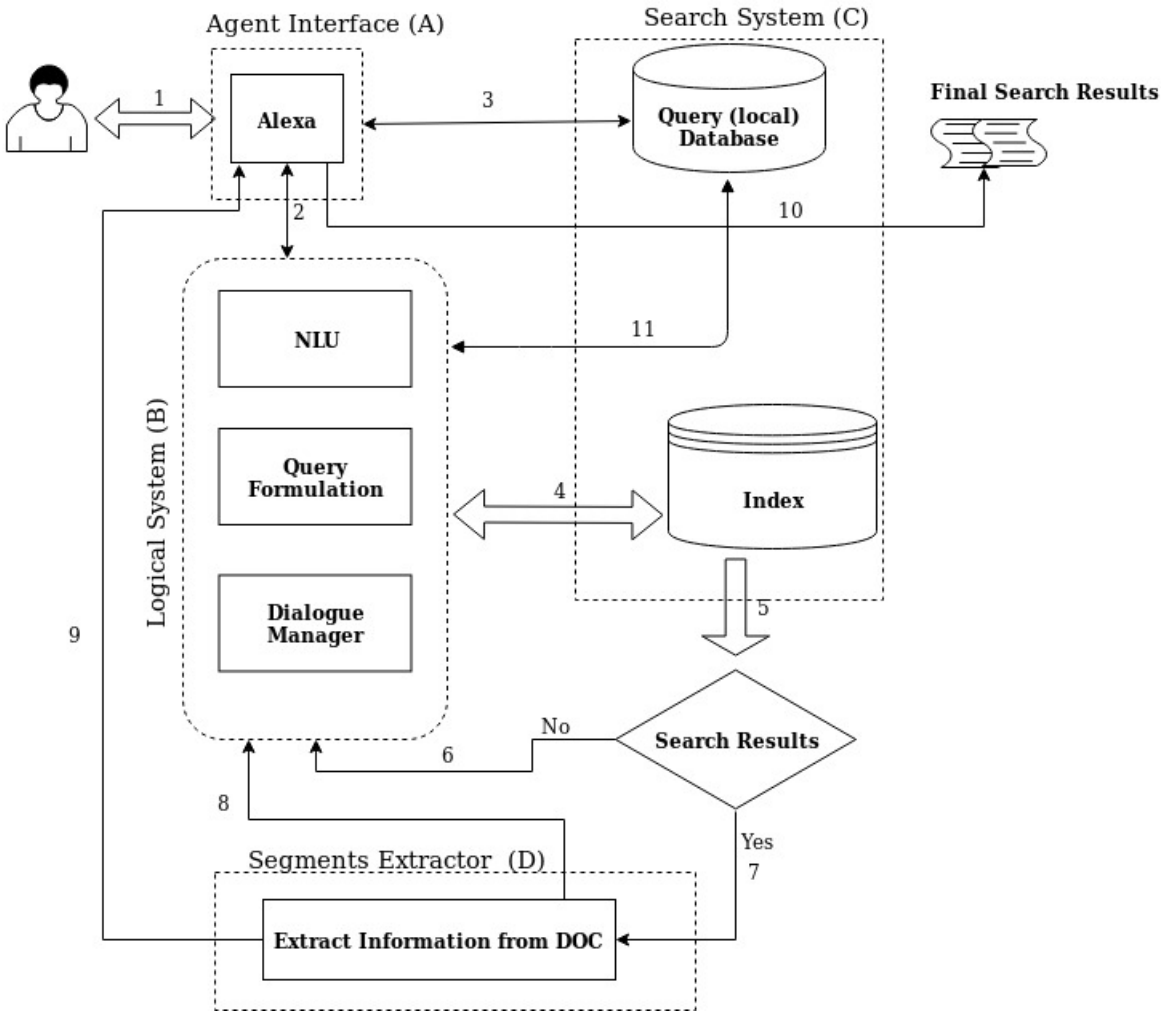


Figure 6.8: Conceptual diagram of our agent and chat process from step 1 (greet) to step 10 (displaying the results) of Custom Alexa.

6.4.4.1 Study Details

A total 31 search sessions were conducted using same participants as the standard Alexa study, but with a different backstory assigned to each user following a Latin square backstory assignment process to avoid biasing effects between participants and assigned backstories. The participants were again given the opportunity to familiarise themselves with the application for 5-10 minutes prior to beginning the search task, had to complete pre-search and post-search questionnaires during the search session, and also to participate in a semi-structured interview at the conclusion of the search task. Responses were again assigned to categories by independent analysts.

In the following subsection, we describe the dialogue strategy for our custom

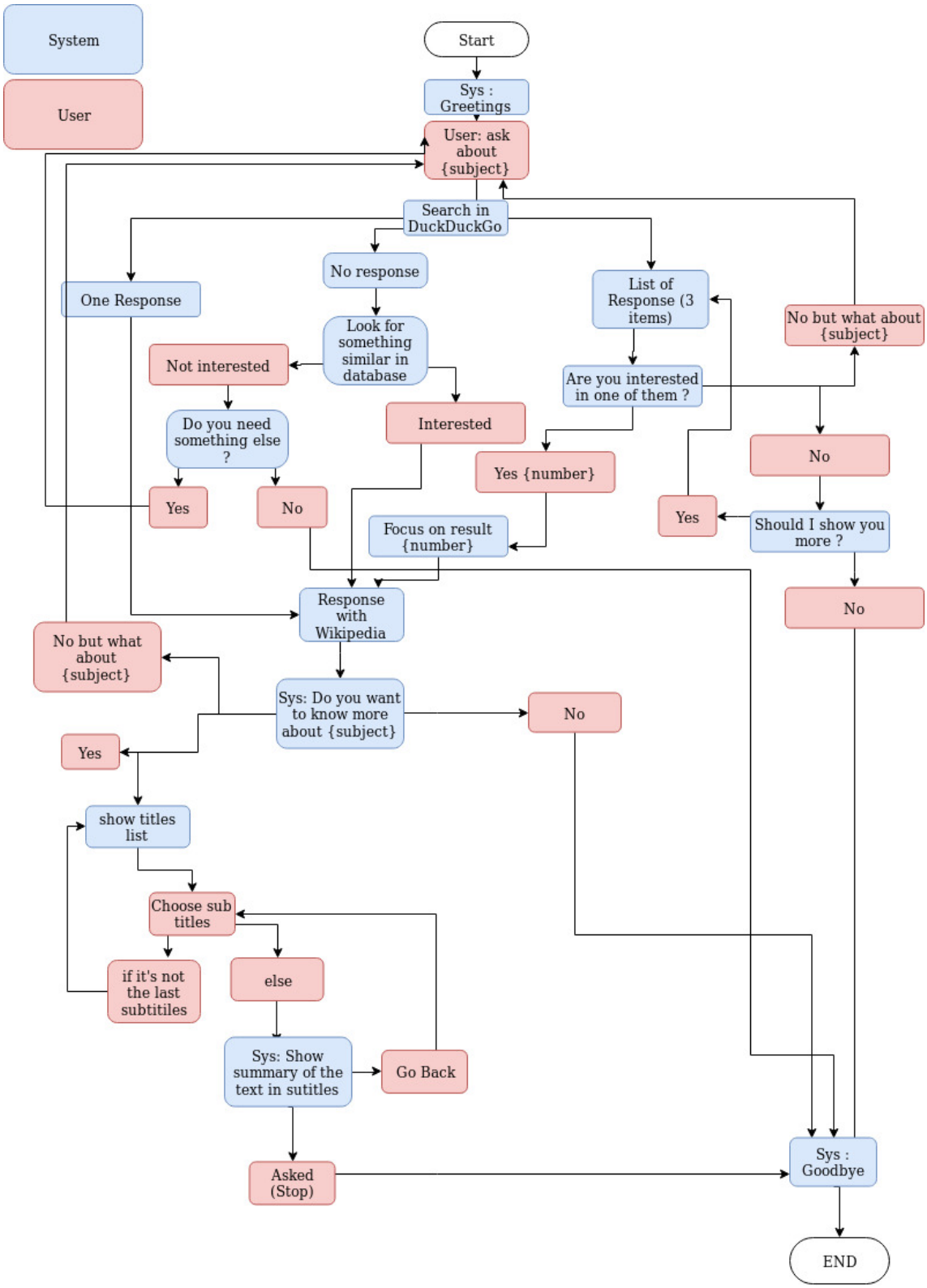


Figure 6.9: Flowchart for Custom Alexa.

search application, the search process used and additional associated functionality. The flowchart of the custom Alexa application is shown in Figure 6.9.

6.4.4.2 Dialogue Strategy

The dialogue strategy was designed to enable users to search for and explore the content of long retrieved documents, and to facilitate interaction between Alexa and the searcher via a dialogue. The dialogue strategy has two major components: developing the skill (e.g. search and greetings) using an Amazon Developer account, and the second part to embed the search process into the dialogues.

1. Alexa Skills: We designed two intents for the Custom Alexa skills. These actions were developed to fulfill spoken requests entered by the searcher. Each intent has at least one trigger utterance, a predefined word or phrase which the user might say to invoke the intent. The intents are *Greeting* (trained to answer greetings, unexpected questions and non-relevant questions with respect to a search) and *Search* (trained to identify a search query, pass it to the search system, and to present the response from the search system to the users). Each intent was trained using likely user utterances with a corresponding response which would be expected by the users. We selected 80+ alternative utterances for both intents. The utterances were collected using a small survey among a group of undergraduate students. The training model is an inbuilt function of the Alexa cloud service, which is a black box. Training can be started by supplying a sample of intents and entities. Our entities were general topics which we considered to be noun objects.

2. Embedding the Search Process with Dialogue: Custom Alexa supports the identification and classification of different input intents by the user. Once Alexa identifies the input as the search intent, it passes the user input to the search process which extracts the search query from the user input. The query is extracted by entity identifier of Alexa. The extracted query is searched based on similarity matching in the query archive (local data base used to store each

successful query made by the user as shown in Figure 6.8) and which asks for confirmation from the user regarding the search query extracted. The objective of the local database is to suggest a query based on an archive of past successful queries and to help the user with query formulation, which was a major challenge in the default Alexa setting. This helped in correct query identification with the goal of reducing the error rate because of incorrect query and improve reliability of relevant search results. This acknowledgment from the user triggers the search process. All the responses from the search process were embedded within dialogue based on our dialogue strategy as shown in Figures 6.7 And 6.9. Priming technique has been adopted to embedded dialogue with information. Priming is the concept where the agent drove the user by providing options.

6.4.4.3 Search Process

The search process is triggered once the user has confirmed the search query extracted from the user input to Alexa. Figure 6.8 shows the search process workflow. This contains the following sub-modules:

1. Calling Duck Duck Go API: The query is passed to the Duck Duck Go search API. The titles of the top 3 documents returned by Duck Duck Go are displayed by Alexa on the Echo Show screen, as shown in Figure 6.7. The searcher can then select one of these from these results by saying 'Open 1' or 'Open <document name>' or the searcher may request more results by rejecting the displayed items by saying 'No, show me more results'. Alternatively, the searcher may change their query and restart the search process by saying 'Alexa start search', which restarts the search process. Alexa initiates the search again by saying 'I am Adapt search agent, I can help you in the search process, please ask your query; For an example "Who is harry potter?"'. Once the user has selected selects an item from the displayed results, the dialogue strategy triggers the Wikipedia API.

2. Calling Wikipedia API: The title of the user selected document is passed to the Wikipedia search API. The section and subsection headings of the highest ranking retrieved item are then shown to the user. The dialogue strategy then provides options which enable the searcher to select sections and subsections of the returned documents. These selected parts of the document are then summarised using the summarisation component outlined below. The document navigation options enable the searcher to explore the individual summarised parts of the document.

3. Calling the Summarizer: The summarizer from study [167] was used to display the important content of the section of a document selected by the user on Echo Show. The summarizer algorithm is shown in algorithm 1. The algorithm used the default parameters based on the above discussed study. The Echo Show displays the summary along with further sub-options to explore more, as shown in Figure 6.7 and algorithm 1. The searcher can explore further subsections or go back to the previous view. As soon as the user chooses any section or sub-section to explore, the summarizer extracts the summary of that specific subsection

Algorithm 1: Pseudo code Custom Summarization Method

Result: Custom Summarization Algorithm

1. Split the whole paragraph (conversation) into sentences
2. Each sentence is considered as individual document
3. Calculation of term frequency of each word of document
4. Calculation of inverse document frequency (IDF) of each word from document
5. Calculation of TF-IDF score of each word
6. Calculate the normalized TF-IDF score of each sentence by summing the TF-IDF score of each sentence and dividing the sum by word length of the sentence
7. Top 50% top scoring sentences extracted
8. DBSCAN clustering algorithm is applied on the extracted sentences with parameters eps = 0.1 and minimum sample = 2 where eps is responsible to consider two points as neighbour if the distance between them is less than eps value and minimum sample size is number of points needed to form cluster
9. DBSCAN divides the sentences into n clusters
10. Cosine similarity score is calculated between the n number of clusters and section name
11. Based on the cosine similarity score, top 70% of n clusters have been extracted for keyphrase extraction, which are arranged in order as it lie in actual paragraph.

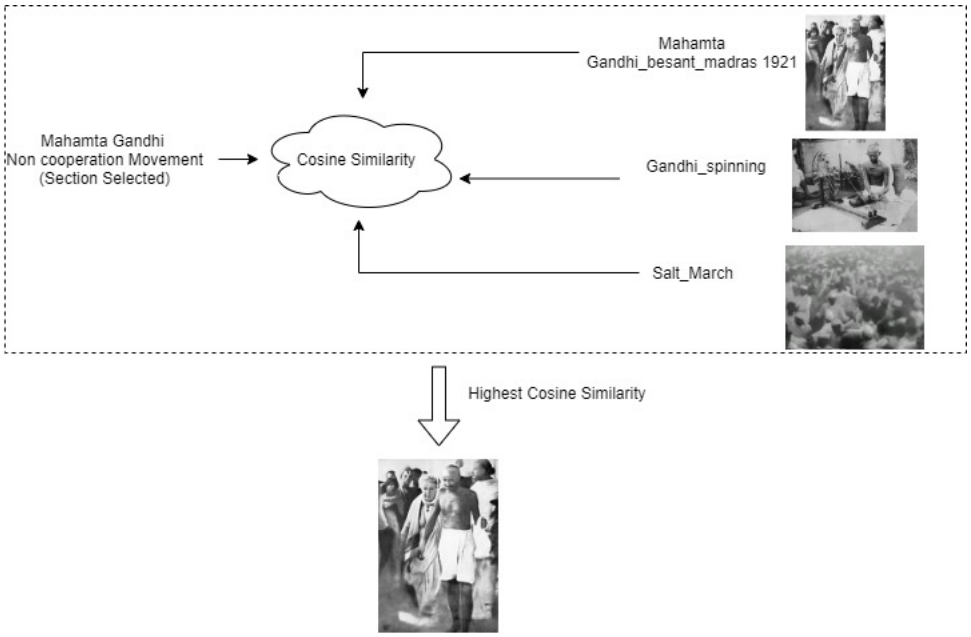


Figure 6.10: Use of cosine similarity between subsection title and image labels to select of image for display.

6.4.4.4 Additional Functionality

Our Custom Alexa skill also selected images associated with the displayed subsections. For images are provided by the Wikipedia API. We use cosine similarity between the images labels and the title of the selected Wikipedia subsection are used to determine which images to display with the text, as shown in Figure 6.10. The image with the highest similarity is shown with the contents of the selected subsection.

6.5 Investigating the Effectiveness of Customised Alexa for a Exploratory Search

We investigated the effectiveness of our Custom Alexa application for exploratory search using the study design described above. This examined two additional research questions.

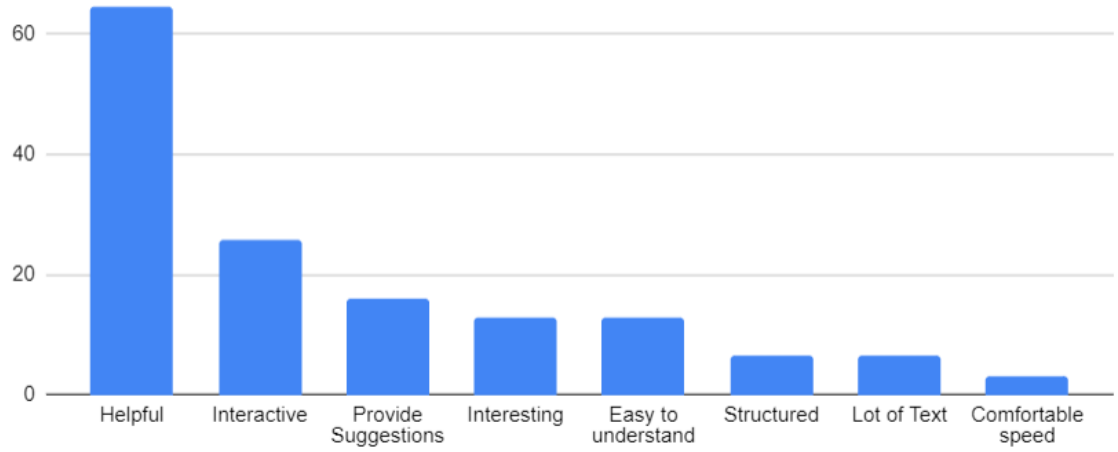


Figure 6.11: Participant feedback on Custom Alexa setting.

6.5.1 RQ4: How well does the Custom Alexa dialogue system support exploratory search?

The custom Alexa application was designed to address the shortcomings identified in our initial investigation of Alexa for exploratory search. In particular, it was designed to enable the user to explore longer returned documents to locate relevant information. This research question is divided into three sub-questions.

6.5.1.1 RQ4(1): How effectively does custom Alexa communicate information to the user?

1. Dialogue Strategy: In the semi-structured interview, carried out in this second study, searchers reacted positively to our Custom Alexa application. Around 45% of users found it helpful. Additionally, 12% of searchers claimed in this Custom Alexa setting, Custom Alexa provided them with suggestions: query suggestions (query from archive), documents list. Users also found Custom Alexa interesting, easy to understand with comfortable speed and structured, unlike what they experienced with standard Alexa Figure 6.5.

2. Structure of representation of information: Representation of the information was one of the key criteria of the search process. We break down the information structure into two major components: the content of document and

representation of the information.

(a) Content of document: Around 87% of participants were satisfied with the information displayed by the custom Alexa setting. The custom setting was able to satisfy around 85% of the user's information needs as shown in Figure 6.4.4.

(b) Structure of text: Test structure refers to the way Alexa display the information on the screen. Around 77% of searchers were satisfied with text structured provided by the Custom Alexa setting, while 9.7% users were more content with standard Alexa text structure. The custom setting was able to satisfy 75% of the searchers. The reasons behind preferring Custom Alexa were ease of use (7.5%), ease of information seeking (37.5%), its interactive nature (30%) and that it is more relevant and informative (25%).

6.5.1.2 RQ4(2): How to verify the user understanding, satisfaction and search success in the dialogue-based exploratory search process?

The questionnaire, interviews and interaction videos enable us to interpret the searchers' understanding, satisfaction and success in using the Custom Alexa application. Our analysis is divided into two major components: search experience and knowledge expansion via summary comparison.

1. Search experience: The searchers rated (out of five) a range of variables comparing the Default setting and Custom setting of Alexa. The searchers rated Custom Alexa skill higher for most of the variables (as shown in Table 6.2) in comparison to the Default search setting.

2. Knowledge expansion via summary comparison: To verify the expansion in knowledge, we conducted a comparison of pre-task summary and post-task summary for the Default and Custom Alexa settings using a standard comparison methodology. The summary comparison is based on three standard

Variables	Default	Custom	P Value
Text Quality*	3.1	3.9	0.00012
Navigation Skills*	2.7	4.1	>0.00001
Speech Skills*	3	4.2	0.00013
Presentation Skills*	3.3	3.9	0.0007
Better Understanding	3.3	3.7	0.8502
Knowledge Expansion*	3.2	3.9	.00146
Cognitive Engagement	3.4	3.9	0.1649
Search Session Success*	3.1	3.7	0.01354
Suggesting Skills*	2.5	3.7	0.0014
Alexa Stop	1.8	1.3	0.28242
Ease of MultiModal	3.7	4	0.16067

Table 6.2: Comparison between Default setting and Custom setting with statistical testing: two tailed T paired test with 95% confidence where * indicate, statistically significant in P<0.05.

Parameters	Default Alexa	Custom Alexa	P <0.05
D-Qual*	14	35	0.00904
D-Intrp*	17	37	0.01577
D-Crit*	15	21	>0.00001

Table 6.3: Summary comparison of Default and Custom Alexa on quality, interpretation and critique (Two tailed T independent test) [96].

parameters named: D-Qual, D-Intrp and D-Crit as explained in Table 4.5 [97].

The difference between all factors in pre-search task with post-search task is greater in custom (C) setting than the default (D) setting as shown in Table 6.3.

The summaries are annotated by two independent accessors based on the parameters shown in Table 4.5 [97]. The parameters' specifics were presented in the methodology described in Chapter 4. This indicates that the searcher wrote a better summary with more facts and analysis after using the Custom Alexa search setting. An example of a summary shown in Table 6.4.

	Pre-search summary	**Post-search summary**
Default Alexa	Nelson Mandela was African anti-apartheid revolutionary who served as president of South Africa. He served as a president from 1994 to 1999.	Nelson Rolihala Mandela born on 18/07/1918. He married to Gracie Marchel. He was member of national party which was NP-Led government at that time. He died on 05/12/2013. at Johannesburg.
Custom Alexa	Annie Besant was a British socialist and supporter of both Irish and Indian self-rule	Annie Besant had written over 3000 books. For the crown against nation (1886) and why do I not believe in god (1887) was famous book among 3000 books. She freed in September 1917 and become Indian national congress party leader for a year. At her early life she had contact with Brayant and may's match fact in london where womens used to get paid poorly. She was truly fought for freedom and education right. She was indeed women right activist

Table 6.4: Example summaries related Pre-search and Post-search using Default Alexa and Custom Alexa.

6.5.1.3 RQ4(3): Can priming help in information seeking and reducing errors in conversations?

During our interview sessions, users were asked about their experiences using the Alexa Custom setting. They answered questions relating to two dimensions: i) reasons to prefer Alexa Custom setting (Figure 6.12), and ii) what are the challenges

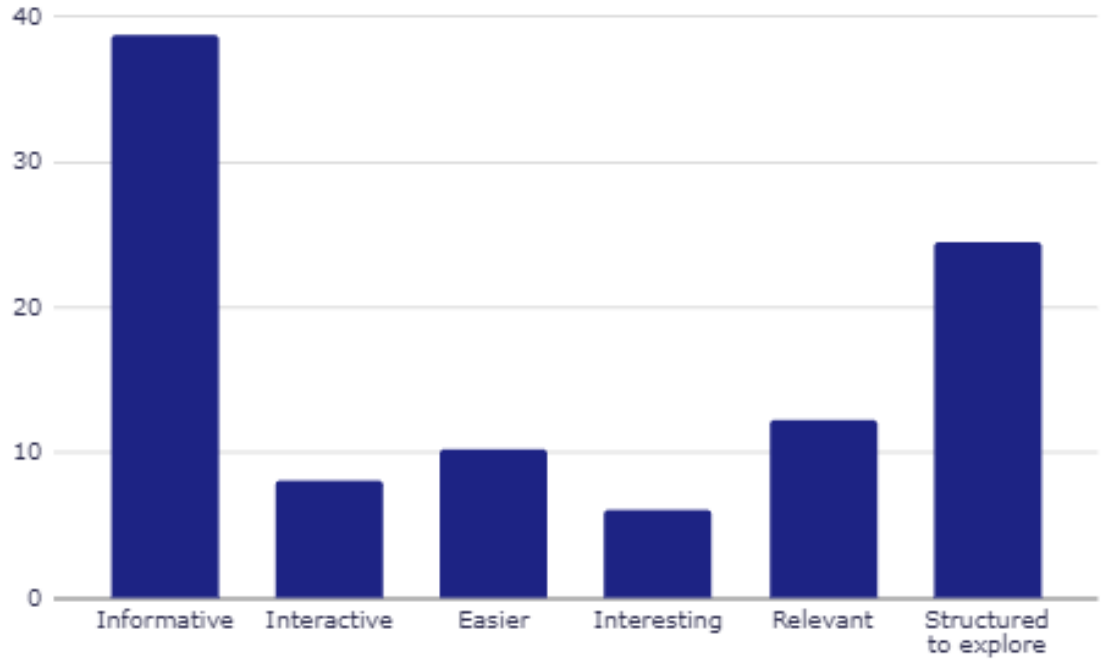


Figure 6.12: Reasons for preferring the Custom Alexa in information seeking.

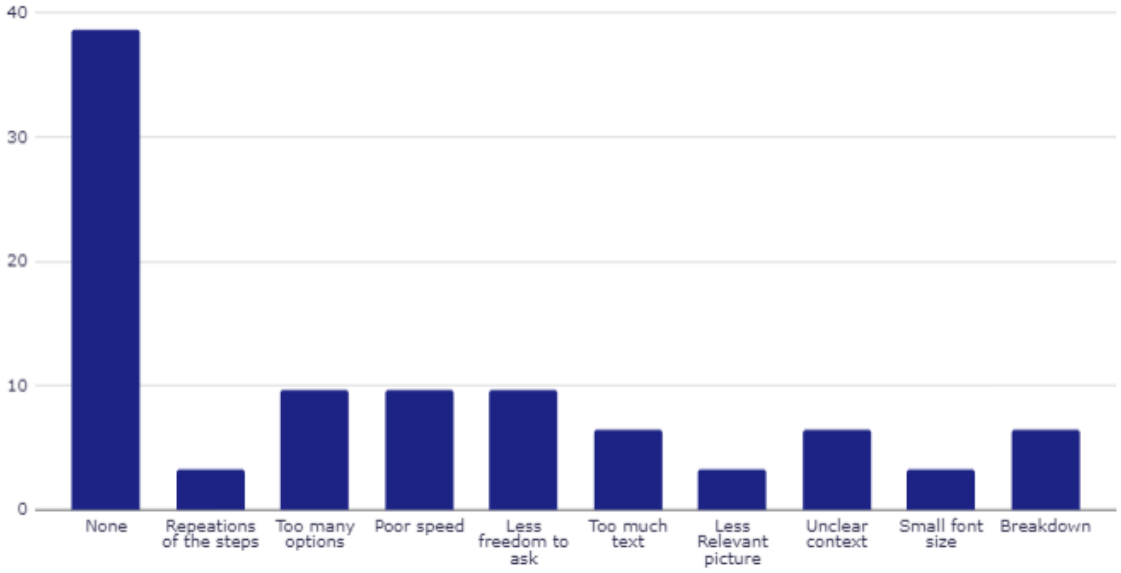


Figure 6.13: Challenges of using Custom Alexa experienced by users in information seeking.

of using the Alexa Custom setting (Figure 6.13)?

1. Reasons to prefer Alexa Custom setting: The top three reasons to choose Alexa custom setting were: Navigation and Directed Search (13%), Relevant and More Informative (21%), and Options and Suggestions (19%). Overall, the users found it informative, well directed search, and that it provided options which gave them the opportunity for exploration throughout the search process, as shown in Figure 6.12.

2. Challenges of using Alexa Custom: Around 40% of searchers were happy and did not find any challenges in using Alexa Custom setting shown in Figure 6.13, in contrast to the Default setting where around 95% people found it challenging, as shown in Figures 6.5 and 6.6. The three top challenges were: too many options (10%), slow speed (10%), and less freedom to enter long queries (10%). We consider these results to be positive outcomes for Alexa Custom setting, since searchers found the response of Default Alexa setting too quick (i.e. spoken responses were delivered too fast to be fully comprehended), and that it was unable to provide suggestions and options during the search process.

6.5.2 RQ5: What is the user search behaviour and experience with Default Alexa and Custom Alexa in an exploratory search setting?

Our final research question focuses on comparing user behaviour patterns during an exploratory search using the Default Alexa and Custom Alexa applications based on user interaction and self-reported answers in the questionnaires.

1. Custom Alexa: Based on our analysis of user interactions (Table 6.2), we observe that participants found the Alexa Custom application more cognitively engaging than the Default application. This observation implies two conclusions: that the custom application can hold the participant's interest in the

search process, and also that participants were able to learn and understand more using the Custom Alexa application.

Some users reported a lower level of knowledge of the topic before commencing the search task. However, they were interested in the topic, which led them to engage with relevant content with a very high of interaction during which they explored in great depth. Other users who also began with less knowledge had notably less interest in the topic. They showed strong engagement with a limited amount of content, but did not explore the retrieved content so widely. Other users with little initial knowledge of the topic had a very high interest in it, which led them to engage more with diverse sources of content, but with less interaction and less detailed examination of specific areas of content. The majority of users were comfortable with the multi-modality of the engagement and were satisfied with the exploratory custom search interface.

2. Default Alexa: In this setting, some users with less background knowledge of the topic engaged repeatedly with the limited content by repeating the same queries to enhance their understanding and search experience with topic. Other users restricted themselves to only a few queries since the poor speech recognition that they experienced led to frustration. Some users expected more options (sections and sub sections) and suggestions to be given by the system as per convention of their previous experiences with IR systems.

   Generally, cognitive engagement with the Default system was less in comparison to the Alexa Custom application. Most of the users were not very comfortable with the multi-modality available in the Default setting. In general we found that the individual pieces of information provided in Default responses were not sufficient to develop a broad knowledge of the topic, resulting in poor post-search summaries.

6.6 Concluding Remarks

We have described a study examining use of the standard Alexa assistant application for exploratory search tasks. This demonstrated that while it is generally found that Alexa is able to answer factoid type questions quite successfully, it is not able to support the requirements of more exploratory search tasks. Our study highlighted these shortcomings in terms of examining multiple retrieved items and specifically engaging with larger items in order to locate relevant information.

In response to these issues, we proposed and implemented a customised Alexa application specifically designed to address the identified challenges. A second study examining our Custom Alexa application showed that it was able to successfully address the identified problems and was well received in terms of usability by the participants in our experimental study.

While our study shows how existing commercial conversational digital assistant applications such as Alexa can be successfully extended to support exploratory search in a manner which is easily used, this is only an initial prototype. In the next chapter, we build on our analysis of opportunities to provide conversational support in a standard search engine reported in the previous chapter, and our study of conversational search using a digital assistant, to develop a prototype multi-view conversational search interface to argument a standard search engine.

Chapter 7

A Conversational Search Interface Supported by a Rule-Based Agent

7.1 Overview

In this chapter we describe a laboratory-based study comparing user behaviour for a conventional search interface (CSI) with that of an agent-mediated multiview conversational search interface (MCSI) which extends the CSI. User reaction and search outcomes of the two interfaces are compared using the implicit evaluation framework using five analysis methods: workload-related factors (NASA Load Task), psychometric evaluation for the software, knowledge expansion, search satisfaction and user interactive experience introduced in Chapter 4. Our investigation uses scenario-based search tasks, and shows the MSCI to be more interactive and engaging than the CSI corresponding, with subjects claiming to have a better search experience.

In this chapter, we investigate partially RQ3, RQ4 and partially RQ5 as presented in Section 1.5.3, Section 1.5.4 and Section 1.5.5 of Chapter 1. The answers to these questions obtained in this Phd study are summarized in details in Section 9.2.3, Section 9.2.4 and Section 9.2.5 of Chapter 9.

7.2 Motivation

Recent years have seen significant research interest in conversational search (CS). Such systems assume the presence of an agent which engages in a dialogue with a searcher to support them in satisfying their information needs [129]. While there has been much discussion of the potential of CS methods, there is little work reporting investigation of operational conversational prototypes, and in particular how these compare with conventional search systems used to perform the same search task. Those studies of CS which have appeared generally adopt a human "wizard" in the role of the search agent [137, 143]. These studies make the implicit assumption that an agent can interpret the searcher's actions with human like intelligence. In this study, we take an alternative position using an automatic rule-based agent to support the searcher in the CS interface and compare this with the effectiveness of a similar conventional search interface to perform the same search tasks. The goals of our study thus include both better understanding of how users respond to CS interfaces and automated agents, and how these compare with the user experience of a CSI for the same task.

A CS system should enable the searcher to develop their query and address their information need via a series of interactions. In this process the conversational agent allows the user to learn about their topic of interest by incrementally assisting them in developing their search query within a dialogue enabling them to move towards satisfying their information need [129]. Such a means of engagement with a search system can potentially reduce cognitive load by supporting the user in creating a query which describes their information need in detail over multiple conversational steps.

Work to date on CS has focused on information communication via a spoken dialogue channel [137, 157]. While this directly represents human-human dialogue, and can be effective for simple question type searches, the rate of information communication and the capacity for human comprehension of spoken information means that this does not offer an efficient means of information communication for more

complex or exploratory search tasks.

Previous studies of CS interfaces have focused on chatbot type interfaces which limit the information space of the search [143, 173], and are very different from conventional graphical search interfaces. Search via engagement with a chat type agent [149] can result in the development of quite different information-seeking mental models [157] to those developed in the use of standard search systems [10], meaning that it is not possible to directly consider the potential of CS in more conventional search settings. We are interested to consider how user mental models of the search process from conventional search engines will respond in a CS conversational setting to enhance the user search experience.

The ubiquity of CSIs mean that users have well established mental models of the search process from using these tools. With respect to this, it is important to consider that it has been found in multiple studies that subjects find it difficult to adapt to new technologies, especially when dealing with interfaces [23]. Thus, when presented with a new type of interface for an equivalent search task, it is interesting to consider how users will adapt and respond to it.

In this chapter, we introduce a desktop based prototype MCSI to a search engine API, shown in operational example videos at link1[1]. Our interface combines a CS assistant with an extended standard graphical search interface. The interface agent takes the form of a personal assistant which works beside the user, rather than sitting between the user and the search engine [24]. The user is able to engage directly with the search system, while receiving guidance from the search assistant both to help them to form their query and also potentially to guide their interaction with retrieved content. A useful way of thinking about this is to consider the search agent as taking the role of a librarian assisting a searcher in a library looking for relevant information or items.

Our CSI works similarly to a conventional search system where the user issues a query and receives a SERP response (links and snippet) where links connect to the

[1] `https://drive.google.com/open?id=1AoS5Nrnj7nGrPIsRAiA96ttwvzzwkpCK`

original full documents.

For our study of conversational engagement with a search engine and contrasting it with more conventional user-driven interaction, we adopt a range of implicit evaluation methods. Specifically we use cognitive workload-related factors (NASA Load Task) [15], psychometric evaluation for software [28], knowledge expansion [96], search satisfaction [152] and user interactive experience [150]. Our findings show that users exhibit significant differences in the above dimensions of evaluation when using our MCSI and a corresponding CSI.

This chapter is structured as follows: Section 7.3 provides a technical overview of our MCSI, the methodology for our investigation and a brief explanation of our experimental procedure, Section 7.4 provides our results, and includes analysis, findings and hypothesis testing and Section 7.5 concludes.

Figure 7.1: Conversational Agent incorporating: chat display, chat box, information box, query box with action buttons for Enter and Clear, and retrieved snippets and documents. Green outline indicates the MCSI setting and red block indicates the Conventional Search Interface (CSI) setting.

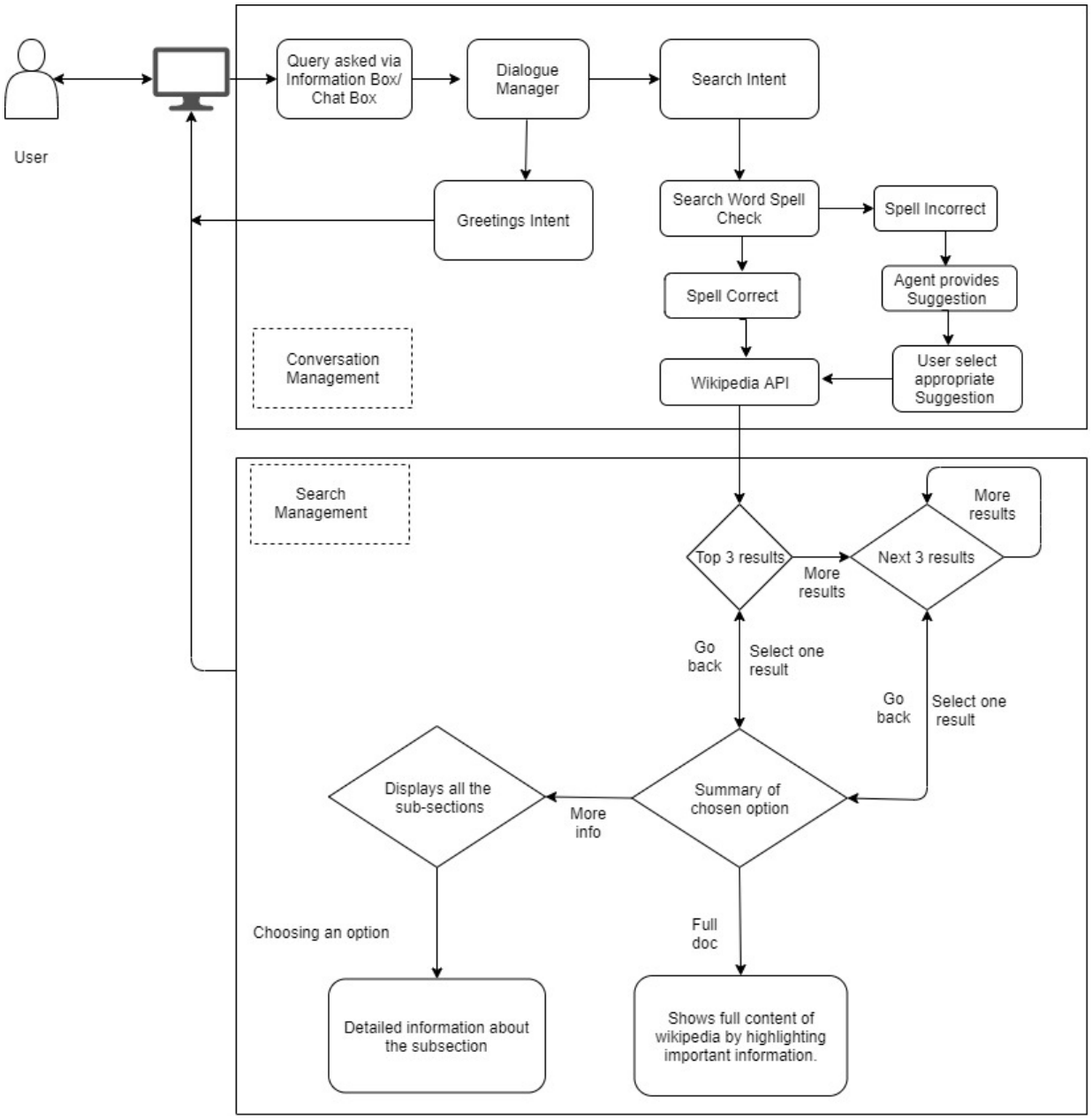


Figure 7.2: Workflow of the conversational model.

7.3 Methodology

In this section we describe the details of our prototype MCSI and of our user study. The study aims to enable us to observe and better understand and contrast the behaviour of searchers using a CSI and our prototype MCSI.

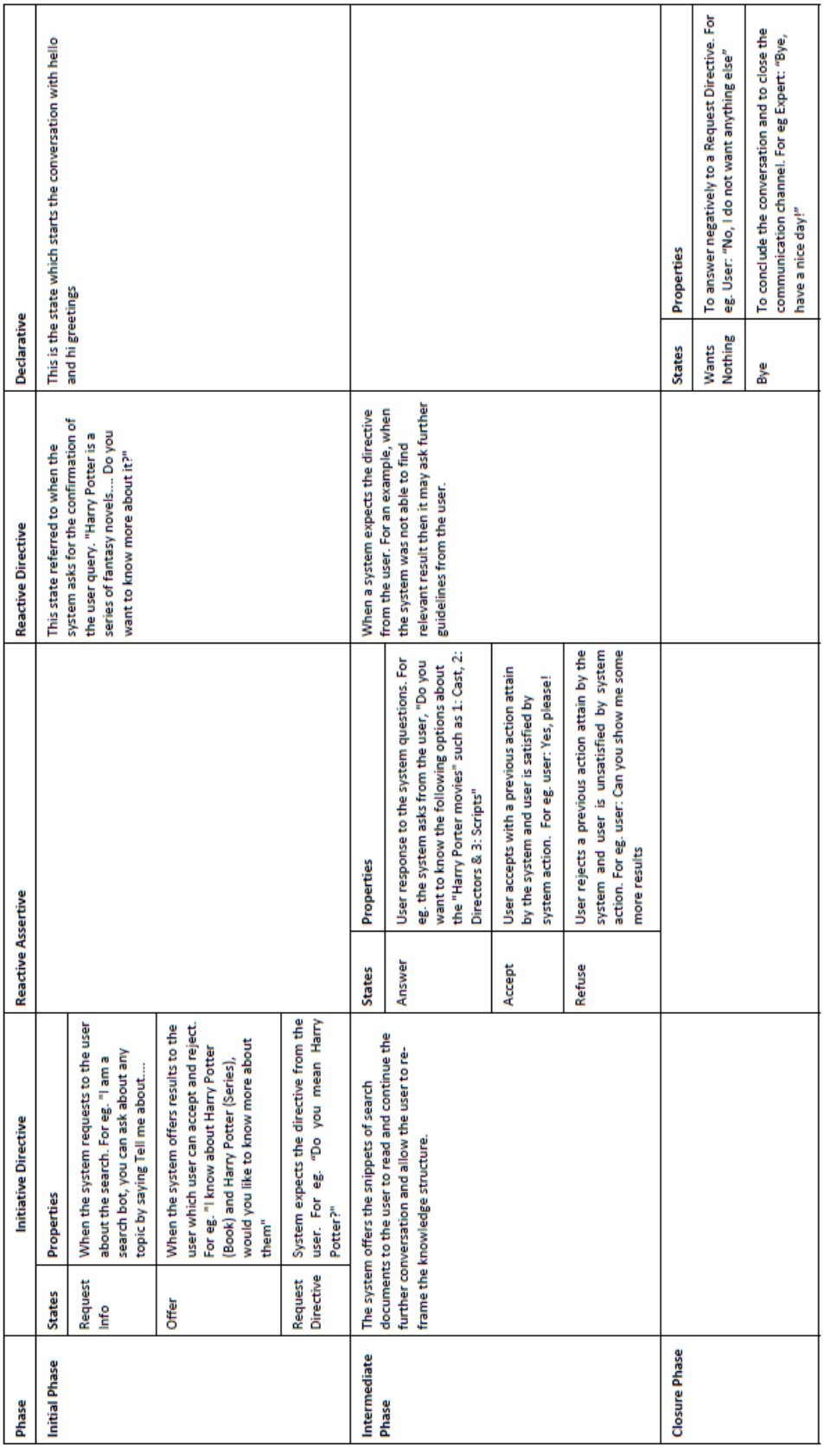


Figure 7.3: Overview of conversational dialogue strategy.

7.3.1 Prototype Conversational Search System

In order to investigate user response to search using a MCSI and to contrast this with a comparable CSI with the same search backend, we developed a fully functioning prototype system, shown in Figure 8.1. The interface is divided into two distinct sections. The righthand side corresponds to a CSI with which the user can interact, and the lefthand side is a text-based chat agent which interacts with both the search engine and the user. Essentially the agent works alongside the user as an assistant, rather than being positioned between the user and the search engine [24]. The Web interface components are implemented using the web python framework flask and HTML, CSS, and JS toolkits. The agent is controlled by a logical system and is implemented using Artificial Intelligence Markup Language (AIML) scripts [2]. These scripts are used to identify the intent of the user and to access a spell checking API [3], and are responsible for search and giving responses to the users. Since the focus of this study is on the functionality of the search interface, the search is carried out by making calls to the Wikipedia API [4]. The interface includes use of an algorithm (based on Custom Alexa algorithm introduced in Section 6.4.4.3) of Chapter 6 which highlights important segments within long documents associated with queries to enable the searcher to skim through them. The interface includes the following components:

Chat Component:

1. Chat Display: Shows the conversational dialogue between the search agent and the user.

2. Chat Box: Enables the user to insert chat.

Search Component:

1. Query Box: Enables the user to use traditional search methods by entering their own query or one suggested by the chat agent.

[2]https://www.botlibre.com/script?category=AIML
[3]https://pypi.org/project/pyspellchecker/
[4]https://pypi.org/project/wikipedia/

2. Retrieval Results: Display of returned document snippets and document contents.

Action Button: Enables the following operations:

1. Enter: Enter the text or query in the Chat Box or Search Box.

2. Clear: Clear the text from the screen.

The search agent communicates with the user via the Chat Box. The agent performs various potential actions which include seeking clarification of ambiguous queries, suggesting words to improve the query, and possibly to answer questions based on user selection of the response or operations.

The search assistant can accept the following user commands:

1. Full Doc: Opens the full document with highlighting of important sections.

2. More Info: Displays all subsections of the selected document.

3. Go Back: Displays the previous results.

4. More Results: Displays more search results.

5. More About: Shows the overall summary from Wikipedia of the selected document.

6. Option: Provides detailed text about selected specific sub-sections.

7.3.1.1 Dialogue Strategy and Taxonomy

After investigating user search behaviour [152] and dialogue systems [20, 22, 77, 192], we developed a dialogue strategy and taxonomy to support CS. The dialogue process is divided into three phases and four states as shown in Figure 7.3. The three phases include processes:

- To identify the information need of the user,

- to present results in the chat system,

- and to continue the dialogue until the user is satisfied or the user aborts the search.

The agent can seek confirmation from the user, if the query is not clear, it can also correct the query and reconfirm the query from the user to make the process precise enough to provide better results. The agent can also highlight specific information in long documents to help the user to direct their attention to potential important content. The user always has the option to interrupt the ongoing communication process by entering a new query directly into the Query Box. The communication finishes with the user ending the search with success (satisfaction of their information need) or with search failure.

The dialogue strategy is structured as follows (Figure 7.3).

1. Initial Phase: This consists of three dialogue states: Initiative Directive (includes: request information, offer and request directive), Reactive Directive, and Declarative.

2. Intermediate Phase: This starts after the searcher executes a search query. This consists of the following states: Initiative Directive, Reactive Assertive (answer, accept and refuse), and Reactive Directive.

3. Final phase: This is the closure phase where the user ends the engagement by being declarative (Want nothing, Bye).

7.3.1.2 System Workflow

The system workflow is divided into two sections: Conversation Management and Search Management, as shown in Figure 7.2.

1. Conversation Management: This includes a Dialogue Manager, a Spell Checker and connection to the Wikipedia API. The Dialogue Manager validates the user input and either sends it to the AIML scripts or self handles it. We use AIML scripts to implement the response to the user as per the dialogue

strategy. If the user input is directed to the AIML scripts then the system response is determined by the AIML script which could further classify the user's intent, if unable to identify the intent, it may ask user to clarify the query again. The two major categories of intent are: *greeting* and *search*:

- Greetings: The greeting intent is responsible for initializing, ending the conversation and system revealment.
- Search: The search intent is responsible for directing the user input to the Spell Checker or the Wikipedia API and transferring control to search management.

The Spell Checking module is responsible for checking the spelling of the query and asking for suggestions from the user (for an example: if the user searches for "vsiual" then the system would ask: Do you mean "visual"?). Once the user confirms "yes" or "no", then the query is forwarded to the Wikipedia API.

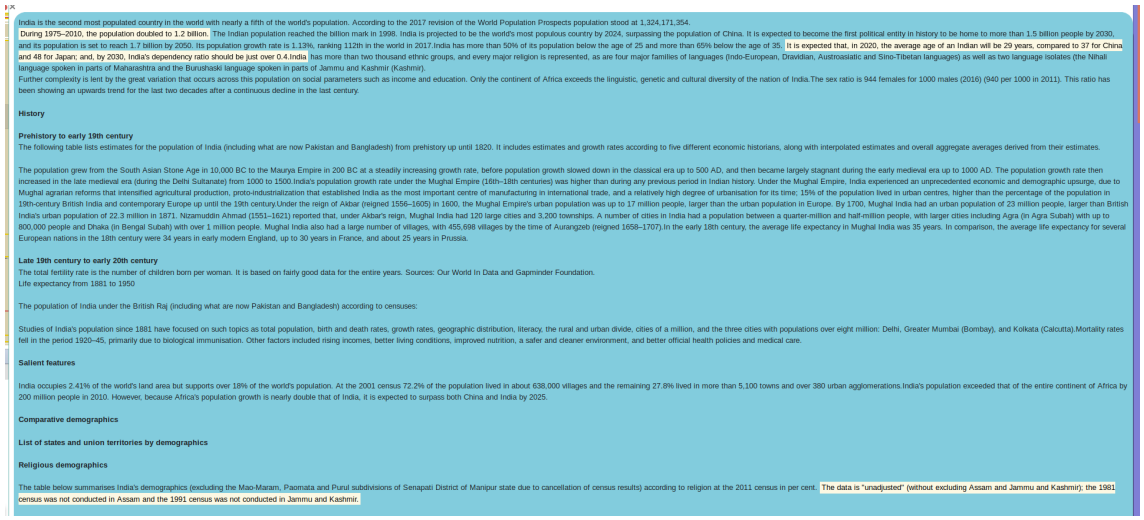
2. Search Management: This is responsible for search and displaying the top 3 search results to the user in the Retrieval Results box. The user may also look for sub-sections from a selected documents. The search agent can give detailed information about the option selected or display more results. Search management also has an option to look at the full document based on an algorithm discussed in Section 6.4.4.3 of Chapter 6. This opens a long document with important sections with respect to the query highlighted.

7.3.1.3 User Engagement

The prototype user interface provides the user with the flexibility to interact with both the search assistant and directly with the search engine. The user can explore a chosen document by selecting multiple subtopics presented from the document at the request of the user, which enable user to select and read topics of interest.

If the user commences a search from the Retrieval Results box, the assistant

initiates a dialogue to assist them in the search process by showing them snippets of the topic, by presenting sections and sub-sections of the selected document, and directing them to further operations that they can do to satisfy their information need. The system also provides support to the user to support them in reading full documents. As described above, important sections in long documents are highlighted to ease reading and reduce cognitive effort.


India is the second most populated country in the world with nearly a fifth of the world's population. According to the 2017 revision of the World Population Prospects population stood at 1,324,171,354.
During 1975–2010, the population doubled to 1.2 billion. The Indian population reached the billion mark in 1998. India is projected to be the world's most populous country by 2024, surpassing the population of China. It is expected to become the first political entity in history to be home to more than 1.5 billion people by 2030, and its population is set to reach 1.7 billion by 2050. Its population growth rate is 1.13%, ranking 112th in the world in 2017.India has more than 50% of its population below the age of 25 and more than 65% below the age of 35. It is expected that, in 2020, the average age of an Indian will be 29 years, compared to 37 for China and 48 for Japan; and, by 2030, India's dependency ratio should be just over 0.4.India has more than two thousand ethnic groups, and every major religion is represented, as are four major families of languages (Indo-European, Dravidian, Austroasiatic and Sino-Tibetan languages) as well as two language isolates (the Nihali language spoken in parts of Maharashtra and the Burushaski language spoken in parts of Jammu and Kashmir (Kashmir).
Further complexity is lent by the great variation that occurs across this population on social parameters such as income and education. Only the continent of Africa exceeds the linguistic, genetic and cultural diversity of the nation of India.The sex ratio is 944 females for 1000 males (2016) (940 per 1000 in 2011). This ratio has been showing an upwards trend for the last two decades after a continuous decline in the last century.

History

Prehistory to early 19th century
The following table lists estimates for the population of India (including what are now Pakistan and Bangladesh) from prehistory up until 1820. It includes estimates and growth rates according to five different economic historians, along with interpolated estimates and overall aggregate averages derived from their estimates.

The population grew from the South Asian Stone Age in 10,000 BC to the Maurya Empire in 200 BC at a steadily increasing growth rate, before population growth slowed down in the classical era up to 500 AD, and then became largely stagnant during the early medieval era up to 1000 AD. The population growth rate then increased in the late medieval era (during the Delhi Sultanate) from 1000 to 1500.India's population growth rate under the Mughal Empire (16th–18th centuries) was higher than during any previous period in Indian history. Under the Mughal Empire, India experienced an unprecedented economic and demographic upsurge, due to Mughal agrarian reforms that intensified agricultural production, proto-industrialization that established India as the most important centre of manufacturing in international trade, and a relatively high degree of urbanisation for its time; 15% of the population lived in urban centres, higher than the percentage of the population in 19th-century British India and contemporary Europe up until the 19th century.Under the reign of *Akbar* (reigned 1556–1605) in 1600, the Mughal Empire's urban population was up to 17 million people, larger than the urban population in Europe. By 1700, Mughal India had an urban population of 23 million people, larger than British India's urban population of 22.3 million in 1871. Nizamuddin Ahmad (1551–1621) reported that, under Akbar's reign, Mughal India had 120 large cities and 3,200 townships. A number of cities in India had a population between a quarter-million and half-million people, with larger cities including Agra (in Agra Subah) with up to 800,000 people and Dhaka (in Bengal Subah) with over 1 million people. Mughal India also had a large number of villages, with 455,698 villages by the time of Aurangzeb (reigned 1658–1707).In the early 18th century, the average life expectancy in Mughal India was 35 years. In comparison, the average life expectancy for several European nations in the 18th century were 34 years in early modern England, up to 30 years in France, and about 25 years in Prussia.

Late 19th century to early 20th century
The total fertility rate is the number of children born per woman. It is based on fairly good data for the entire years. Sources: Our World In Data and Gapminder Foundation.
Life expectancy from 1881 to 1950

The population of India under the British Raj (including what are now Pakistan and Bangladesh) according to censuses:

Studies of India's population since 1881 have focused on such topics as total population, birth and death rates, growth rates, geographic distribution, literacy, the rural and urban divide, cities of a million, and the three cities with populations over eight million: Delhi, Greater Mumbai (Bombay), and Kolkata (Calcutta).Mortality rates fell in the period 1920–45, primarily due to biological immunisation. Other factors included rising incomes, better living conditions, improved nutrition, a safer and cleaner environment, and better official health policies and medical care.

Salient features

India occupies 2.41% of the world's land area but supports over 18% of the world's population. At the 2001 census 72.2% of the population lived in about 638,000 villages and the remaining 27.8% lived in more than 5,100 towns and over 380 urban agglomerations.India's population exceeded that of the entire continent of Africa by 200 million people in 2010. However, because Africa's population growth is nearly double that of India, it is expected to surpass both China and India by 2025.

Comparative demographics

List of states and union territories by demographics

Religious demographics

The table below summarises India's demographics (excluding the Mao-Maram, Paomata and Purul subdivisions of Senapati District of Manipur state due to cancellation of census results) according to religion at the 2011 census in per cent.; The data is "unadjusted" (without excluding Assam and Jammu and Kashmir); the 1981 census was not conducted in Assam and the 1991 census was not conducted in Jammu and Kashmir.

Figure 7.4: Sample of the long document in MCSI.

7.3.1.4 Review of Long Documents

In the study conducted reported in [152], we noted that users could spend considerable time reviewing long documents, as shown in Figure 7.4. Our MCSI supports these users by highlighting important segments (using the custom algorithm described in Chapter 6) in Section 6.4.4.3 with respect to the user's query. This facility also provides the user with the opportunity to select subsections within a document instead of needing to read the full document. This setting helps the user to put less effort into engaging with these documents.

7.3.1.5 Conventional Interface

To enable direct comparison with our MCSI, the CSI for our study was formed by using the conversational interface with the agent panel removed and the document

highlighting facilities disabled. The searcher enters their query in the Query Box, and document summaries; are returned by the Wikipedia API. Full documents can be selected to be viewed to satisfy the user's information need.

7.3.2 Information Needs for Study

For our investigation, we decided to use a set of information needs specified within *backstories*, discussed in chapter 4.

For search backstories, we use the same selection procedure as discussed in Section 5.3 of Chapter 5.

7.3.3 Experimental Procedure

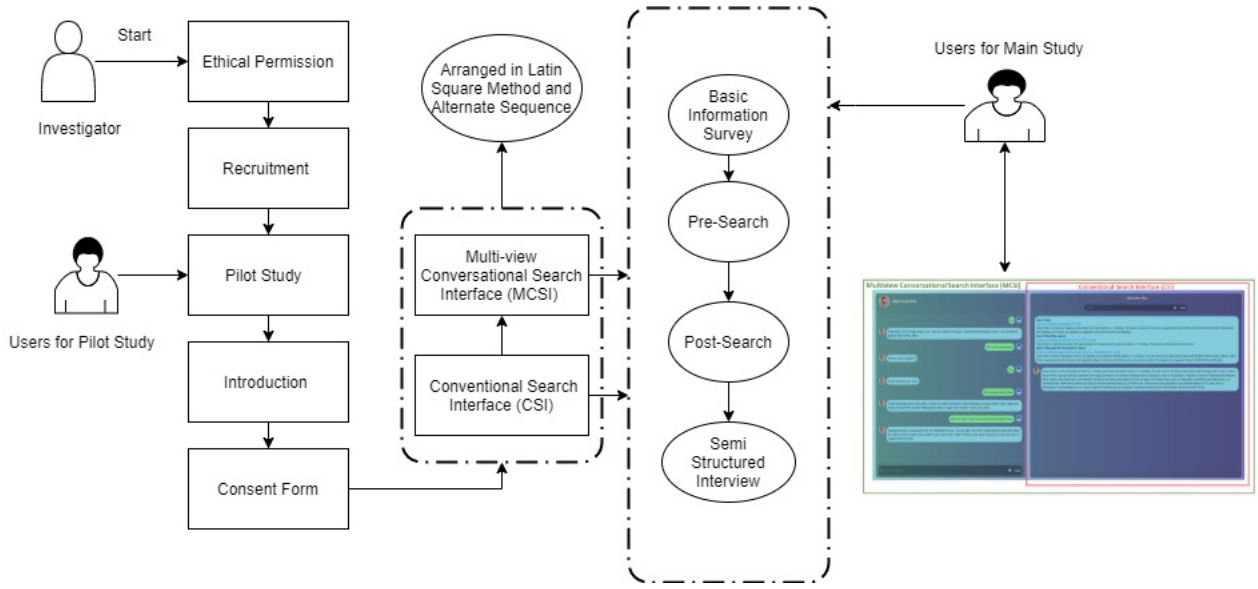


Figure 7.5: Experiment procedure for information seeking tasks to investigate CSI and MCSI interface.

As shown in Figure 7.5, participants in our study had to complete search tasks based on the backstories using the MCSI and the CSI. The sessions were designed to assign search tasks to avoid potential sequence related biasing effects, discussed in Methodology Chapter. Each user search session consisted of multiple backstory search tasks. While undertaking a search session, participants were required to complete a pre- and post task search questionnaires. In this section, we first give

our details of our practical experimental setup, then outline the questionnaires, and finally describe a pilot study undertaken to finalise the design of the study.

7.3.3.1 Experimental Setup

The experimental setup was same as, that used for the study in Chapter 5 except for, the search interface used by the participants in this study.

Participants were given printed details of the instructions for their search sessions and each backstory in printed form at the beginning of each task. Based on our findings in Chapter 5, where we discovered that subjects tend to search more naturally without having access to the backstory while searching for tasks, subjects in this study were not allowed to access the backstory once they started the task.

7.3.3.2 Questionnaires

Participants completed two questionnaires for each search task. The questionnaire was divided into three sections: basic information survey, pre-search and post-search, as described in Chapter 4.

Participants completed the questionnaire online in a Google form as shown in the Appendix B.1.3.

7.3.3.3 Pilot Study

A pilot study was conducted with two undergraduate students in Computer Science using two additional backstory search tasks. This enabled us to see how long it took them to complete the sections of the study using the CSI and the MCSI, to gain insights into the likely behaviour of participants, and to generally debug the experimental setup.

7.3.3.4 Study Design

Based on the results of the pilot study, each participant in the main study was assigned two of the selected 12 search task backstories with the expectation that

their overall session would last around one hour. Subjects completed one task per system. Pairs of backstories for each session were selected using a Latin square procedure and after every six tasks the sequence of allocation of the interface was rotated to avoid of sequence effects in use of the interfaces [3].

Each task was repeated 4 times with the expectation that this would give sufficient results to be able to observe significant differences where these are present. Since there were 12 tasks, this required 24 subjects to participate in the study. In total, 27 subjects participated in our study (excluding the pilot study), we examined the data of 25 subject, since 2 subjects were found not to have followed the instructions correctly. The study was conducted in two phases. Each user had to perform search tasks using the CSI and MCSI with the sequencing of their use of the interfaces varied to avoid learning or biasing effects as described above. The data provided in the Table 7.1 is for those participants whose data was used for analysis.

Experiment	Age	No. Male (M)	No. Female (F)	Ratio (F/M)
Chapter 7 (MCSI Rule based Agent both settings)	18-25 26-35 36-45	9 8 0	4 5 0	4:9 8:8 NA
	Total	17	9	9:17

Table 7.1: Details of age distribution throughout this investigation.

As well as completing the questionnaires, the subjects also attended a semi-structured interview (questions in Appendix B.3) after completion of their session of two tasks using both interface conditions. The videos and interviews were thematically labelled by two independent analysts. The category of themes are 'user friendly and robustness', 'interactive and engaging', etc. The example of labels is shown in Appendix B.3.4. The questionnaire in the interview dealt with user search experience, software usability and cognitive dimensions and was qualitatively an-

Task Load Index	CSI Mean	MCSI Mean	Percentage Change	P-Value
Mentally Demanding	4.16	3.68	11.54	.273795
Physically Demanding	3.12	2.76	11.54	.441676
Hurried or Rushed	3.34	2.76	14.81	.213878
Successful Accomplishing	4.28	5.32	-24.3	.016199
How hard did you have to work to accomplish?	4.44	3.96	10.81	.270243
How insecure, discouraged, irritated, stressed, and annoyed were you?	3.32	2.40	27.71	.071443

Table 7.2: Task load index to compare the load on user while using both the systems (MCSI and CSI) with independent T two tailed test .

alyzed. Based on the analysis of interviews, out of 25 participants 92% of them were happy and satisfied with the MCSI. MCSI liked by the majority of the subjects without any effect or preference of using order in search process. This shows that there is no sequence effect arising from the order of the interfaces in the search sessions. In all conditions, subjects prefer the MCSI. Each hypothesis of the study was tested using a T Test. Each hypothesis was evaluated on a number of factors which contribute to the examination of each dimension as described below.

7.4 Study Results

The MCSI was compared with the CSI using the implicit evaluation framework examining multiple dimensions: cognitive load, knowledge gain, usability, search satisfactions and user experience introduced in Chapter 4. The description as follows:

7.4.1 Cognitive Dimensions

In terms of cognitive load, the user was asked to evaluate the CSI and MCSI in 6 dimensions [15] as shown in Table 7.2.

1. **HO: Users experience a similar task load during the search with the different interfaces**: The user evaluated the system based on six parameters

Topic	CSI Mean	MCSI Mean	Percentage Change	P value
Easy to use*	4.04	5.96	47.52	.000059
Simple to use	4.48	5.92	32.14	.003526
Effectively complete my work*	3.92	5.64	43.88	.000226
Quickly complete my work*	3.72	5.76	54.84	.00003
Efficiently complete my work*	3.88	5.76	48.45	.000045
Comfortable using this system*	4.16	5.88	41.35	.000471
Whenever I make a mistake using the system, I recover easily and quickly*	4.04	5.44	34.65	.006827
The information is clear*	4.16	5.92	42.31	.000072
It is easy to find the information I needed*	4.00	5.48	37	.000706
The information is effective in helping me complete the tasks and scenarios*	4.20	5.68	35.24	.000675.
The organization of information on the system screens is clear*	4.44	5.92	33.33	.000184
The interface of this system is pleasant*	4.28	6.08	42.06	.00002
Like using the interface*	4.20	6.12	45.71	.000014
This system has all the functions and capabilities I expect it to have*	4.08	5.72	40.2	.000168
Overall, I am satisfied with this system*	4.16	5.92	42.31	.000029

Table 7.3: Post Study System Usability Questionnaire (PSSUQ).

which consist of mental load and physical load as shown in Table 7.2. The grading scale was assigned between 0 (low) - 7 (High). We compare the mean of the systems in all six parameters with independent T two tailed test. In all aspects, subjects experienced lower task load using the MCSI. Subjects claimed more success in accomplishing the task using the MCSI. All factors were not significantly different. As a result, we are unable to reject the null hypothesis. We have observe the mean difference between both the systems on these factors was more than 11%. Changes in the percentage of all parameters conclude that the user experienced less subjective mental workload while using the MCSI.

7.4.2 Software Usability

Usability is an important evaluation metric of interactive software. The IBM Computer Usability Satisfaction Questionnaires are a Psychometric Evaluation for soft-

ware from the perspective of the user [28] known as the Post-Study System Usability Questionnaire (PSSUQ) Administration and Scoring. The PSSUQ was evaluated using four dimensions: overall satisfaction score (OVERALL), system usefulness (SYSUSE), information quality (INFOQUAL) and interface quality (INTERQUAL), which include fifteen parameters. In each dimension, the MCSI outperformed the CSI. The grading scale was assigned between 0 (low) - 7 (High). We compare the mean difference of both systems in all parameters. In all aspects, subjects experienced less task load when using the MCSI, as shown in Table 7.3 where * indicates statistically significant results.

1. **H0: User Psychometric Evaluation for the MCSI and CSI has no significant difference** An independent T test was conducted. It was found that for all the parameters the MCSI outperformed the CSI. The null hypothesis was rejected and the H1 hypothesis was accepted, which is that the MCSI performs better than the conventional search interface.

Topic	Pre-Task	Post Task	P Value
DQual (1-3)*	0.32	1.56	.00005
DCrit (0-1)	0	0.32	.0026
DIntrp (0-2)*	0	0.84	.00005

Table 7.4: Comparison of Pre-search and Post-search summary for the CSI (Change in Knowledge).

Topic	Pre-search	Post search	P Value
DQual (1-3)*	0.52	2.12	<.00001
DCrit (0-1)*	0.12	0.72	<.00001
DIntrp (0-2)*	0.28	1.36	<.00001

Table 7.5: Comparison of Pre-search and Post-search summary for the MCSI (Change in Knowledge).

7.4.3 Knowledge Expansion

Knowledge expansion was investigated using a comparison of pre-search and post search summaries written by the participant, based on a number of parameters, as

Parameters	**Influence (Chatbot)**	**P value (< 0.5)**
Increase in Critique	87%	.048153
Increase in Quality	29%	.299076
Increase in Interpretation	22%	.312712.

Table 7.6: Comparison of (change in knowledge structure) using the CSI and the MCSI.

shown in Table 4.5, while using both the systems, we divide the hypothesis into two sub-parts as follows:

1. Comparison of pre-search and post search summaries: This is to verify the knowledge expansion after each task, independent of the search interface used by the participant.

2. Comparison of the mean difference between pre-search and post-search summary for each interface: This is to verify which interface supported users better in gaining knowledge.

The knowledge gain was examined by analyzing the pre-search and post search summaries. We asked subjects to write a short summary of the topic before the search and after the search. The summary was analyzed based on three criteria as described in [96]. The criteria are: Quality of Facts (DQual), Interpretations (DInterp) and Critiques (DCrit), as shown in Table 4.5. The summary was scored against these three factors by two independent analysts with the Kappa coefficient (Approx 0.85) [7]. We conducted hypothesis dependent T testing on tasks completed using both the CSI and the MCSI .

1. **H0: No significant difference in the increase of the knowledge after completing the search task in both settings**: As shown in Tables 7.4 and 7.5, the pre-search score and post-search score for all three factors were statistically significant in both the search settings. This implies that subjects expand their knowledge while carrying out the search. This rejects the null hypothesis which leads to the alternative hypothesis which concludes that users

experienced significant increase in their knowledge after completing the search task in both search settings (CSI and MCSI).

After concluding the alternative hypothesis, it was important to investigate whether one system was better in providing support to users in expanding their knowledge. We purposed and tested the following hypothesis.

1. **H0: Knowledge gain during the search is independent of the interface designed:** In this test, we compared the Mean of the difference in the score for pre-search and post-search summaries in both settings. An independent T test was conducted on the change of the three parameter scores as discussed above for the hypothesis testing, as shown in the Table 7.6. It was found that in the MCSI interface setting, the subjects scored higher in the change of critique, quality and interpretation. This implies that the subjects learned more while using the MCSI. The difference in critique score was statistically significant, while the other two parameters were not statistically significant. The quality and interpretation increased more than 20% while using the MCSI. Although there is a rise in all knowledge parameters while using the MCSI, only one parameter was statistically significant. This means that it fails to reject the null hypothesis.

7.4.4 Search Experience

Based on the search as learning Vakkeri model [121], the user search experience can be evaluated on 15 parameters including the relevance of the search result, the quality of the text presented by the interface, and understanding of the topic in both the search settings, via the pre-search and post-search questionnaires.

1. **H0: Subjects find no significant difference between both the interfaces while using the system for the information seeking activity:** The independent T test was conducted using all 15 parameters. It was found

Parameters	**CSI Mean**	**MCSI Mean**	**Percentage Change**	**P value**
Difficulty in finding the information needed to address this task?	4.64	3.16	-35.25	.002168
Quality of text presented with respect to your information need and query?	4.52	5.64	21.55	.010465
How useful were the search results in the whole search task?	4.04	5.12	23.08	.029826
How useful was the text shown in the whole search task in satisfying the information need?	4.08	5.36	27.62	.010245
Did you find yourself to be cognitively engaged while carrying out the search task? *	3.92	5.92	42.31	.000015
Did you expand your knowledge about the topic while completing this search task?	4.84	6	20	.005026
I feel that I now have a better understanding of the topic of this search task.	4.56	5.88	25.64	.002094
How would you grade the success of your search session for this topic?	4.48	5.72	24.35	.005937
How do you rate your assigned search setting in terms of understanding your inputs?	3.72	5.40	39.18	.003121.
How do you rate your assigned search setting in the presentation of the search results?*	3.84	5.76	45.45	.00001
How do you rate the suggestion(s) skills of your assigned search setting?*	3.72	5.56	54.44	.000053

Table 7.7: Characteristics of the search process [121] by the change in knowledge structure where * indicates statistically significant results.

that the null hypothesis was rejected, as shown in Table 7.7. Subjects search experience was statistically significantly better with the MCSI. In the pre-search questionnaire, subjects were asked to anticipate the difficulty level of the search before starting the search and in the post-search questionnaire, they were asked to indicate the difficulty level they actually experienced. It was observed that pre-search anticipated difficulty level and the post-search actual

difficulty level increased in the CSI (16%) and decreased in the case of MCSI interface search task (14%).

7.4.5 Interactive User Experience

Negative	Positive	Scale	CSI_Mean	MCSI_Mean	P_Values
obstructive	supportive	P	3.44	5.60	2.96e-08
complicated	easy	P	3.40	5.76	7.84e-09
inefficient	efficient	P	2.88	4.40	1.69e-05
confusing	clear	P	3.40	5.48	2.31e-06
boring	exciting	H	2.64	5.44	8.88e-16
not interesting	interesting	H	2.48	5.48	9.76e-15
conventional	inventive	H	2.36	6.28	1.17e-14
usual	leading edge	H	1.96	5.20	8.95e-12

Table 7.8: UEQ-S score based on CSI and MCSI where 'P' stands for Pragmatic Quality and 'H' stands for Hedonic Quality (statistically significant).

To ensure a conversational search system provides reasonable User Experience (UX), it is critical to have a measurability which defines user insights about the system. A UX questionnaire for interactive products is the User Experience Questionnaire (UEQ-S) [66, 132, 150]. This questionnaire also enables us to analyse and interpret outcomes by comparing with a benchmarks of larger dataset of outcomes for other interactive products [150]. This questionnaire also provides the opportunity to compare interactive products with each other. For specified purposes, a brief version (UEQ-S) was prepared which had only 8 parameters to be considered [150]. UEQ-S was preferred for the MCSI based on the practical scenarios. For an example, users filled the experience questionnaire after finishing the search task, if there were too many questions, a user may not complete the answers fully or even refuse to answer (as they have finished the search task and are in the process of leaving or starting the next task, so the motivation to invest more time on feedback may be limited). The UEQ-S contains two meta dimensions Pragmatic and Hedonic quality. Each dimension contains 4 different parameters, as shown in the Table 7.8. Pragmatic quality explores the usage experience of the search system, while Hedonic quality explores the pleasantness of use of the system.

1. **HO: Users feel a similar interactive experience when using the different interfaces**: Users evaluated the system based on 8 parameters as shown in Table 7.8. The grading scale was assigned between 0 (low) - 7 (High). We compared the mean difference of both systems on all parameters. In all aspects, subjects experience was positive in Pragmatic quality and Hedonic quality when using the MCSI, and statistically significantly different in comparison to the CSI. Subjects felt obstructive, complicated, confusing, inefficient, and boring, while using the CSI with significant difference (P<0.10). This implies that the null hypothesis was rejected on the basis of the user experience. Based on these findings, we can conclude that the user experience was more pleasant and easy while using the MCSI.

7.4.6 Analysis of Study Results

In summary, hypothesis testing showed that the MCSI reduced cognitive load, increased knowledge expansion, increased cognitive engagement and provided a better search experience load. Based on the results of the study, a number of research questions dealing with factors relating to CS, the challenges of conventional search, and user search behaviour can be addressed.

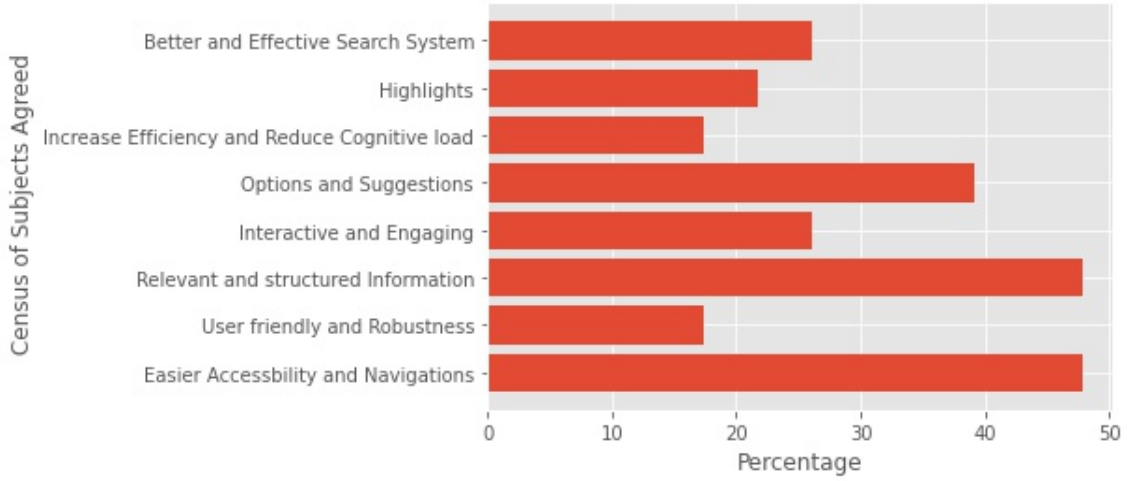


Figure 7.6: Reason provided by the users for preferring MCSI.

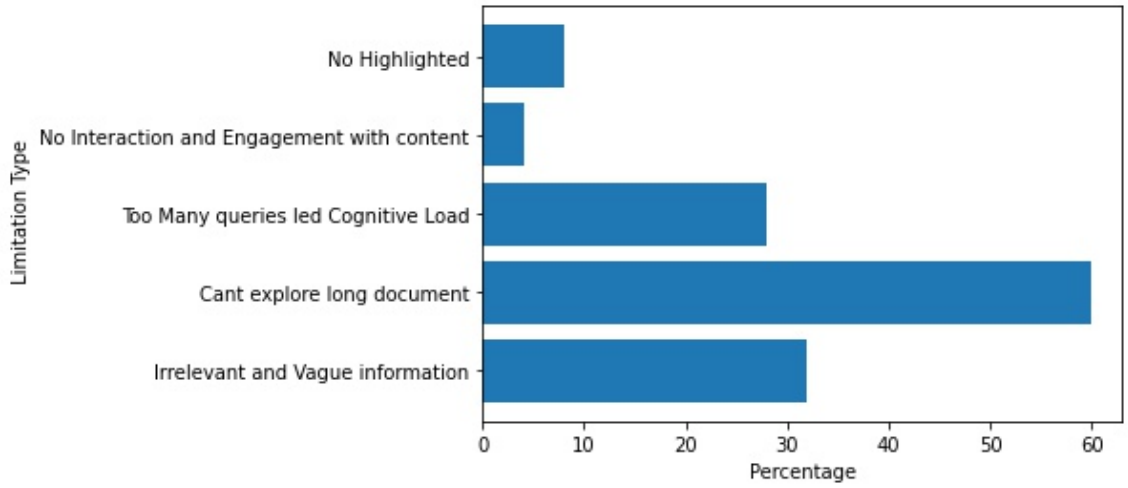


Figure 7.7: Challenges faced by the users while using CSI for information seeking.

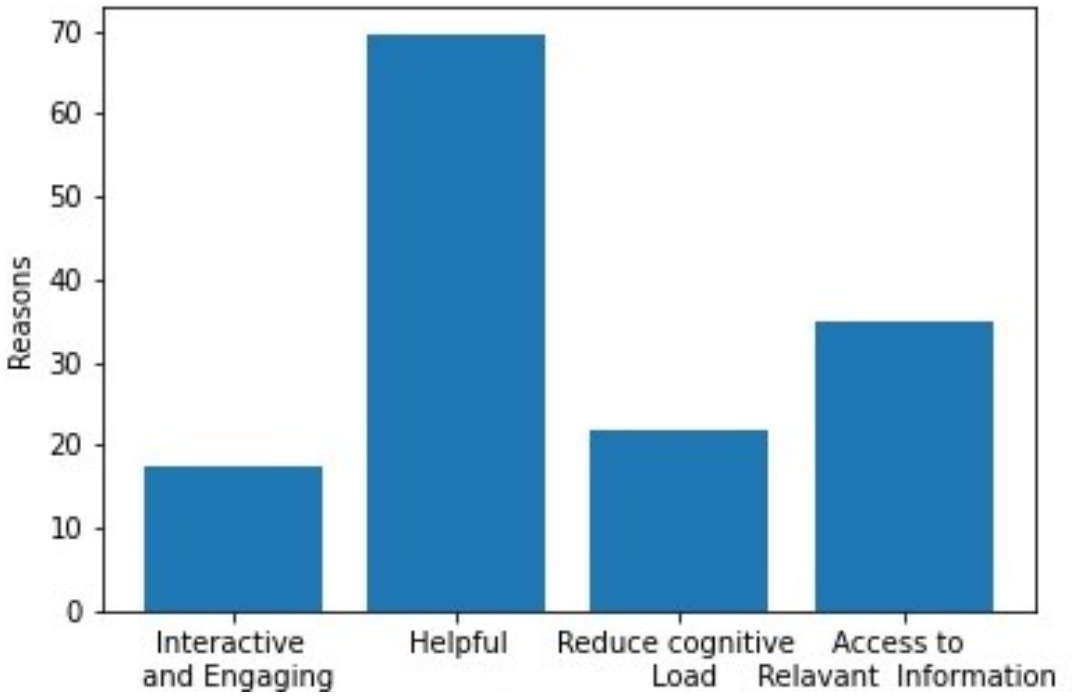


Figure 7.8: Reasons provided by users for preferring the text highlighting option for document interaction where the x axis represents (reasons) and the y axis represents (percentage).

7.4.6.1 RQ1: What are the factors that support search using the MCSI?

Around 92% of the subjects claimed in the post-search interview that the MCSI was better than the conventional search interface. As shown in Figure 7.6, around 48% found that the MCSI allowed them to more easily access the information. A similar view was found in terms of information relevance and its structure as

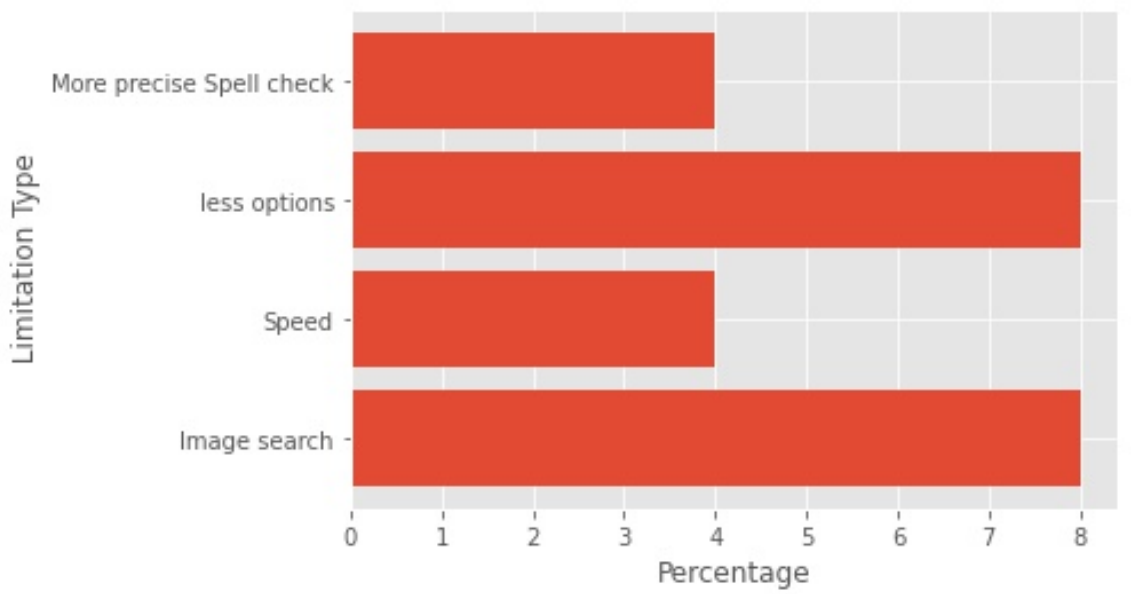


Figure 7.9: Challenges faced by users while using MCSI for information seeking.

presented to the user. Around 38% of subjects were satisfied with the options and suggestions provided by the MCSI. The other reasons for their satisfaction were the highlighting of segments in long documents, finding the search system effective, its being interactive and engaging, and user friendly.

7.4.6.2 RQ2: What are the challenges with the conventional search system?

Subjects found some major challenges in completing the search tasks with the CSI. These limitations were mainly based on the observations from user interactions and feedback after the search task. The limitations can be divided into five broad categories, as shown in Figure 7.7.

Exploration: It was found that the CSI does not support exploration of individual documents. Around 60% of the subjects claimed they found it difficult to explore the content which meant that they were unable to learn through the search process. It was noted that they needed to expend much effort to go through whole documents, which discouraged them from exploring further to satisfy their information need. Another reason was that too much information was displayed to them on the page which confused them during the process of information seeking.

Cognitive Load: Around 28% of subjects experienced issues with cognitive load using the CSI. In a current search systems, a query to the search engine returns the best document in a single shot. The user may need to perform multiple searches by modifying the search query each time to satisfy their information need. There are multiple limitations associated with this single query search approach which put high cognitive load on the user. The following points highlight the limitations and weaknesses of single-shot search [166]:

1. The user must completely describe information need in a single query.

2. The user may not be able to adequately describe their information need.

3. High cognitive load on the user in forming a query.

4. An IR system should return relevant content in a single pass based on the query.

5. The user must inspect returned content to identify relevant information.

Interaction and Engagement: Interaction and engagement with content is a key component of information seeking. Some subjects found difficulty in engaging and interacting with long documents.

Subjects could find content of long documents irrelevant or vague with respect to their specific information need. Using the CSI, 32% of the subjects did not find the long documents precise enough to satisfy their information need. In contrast however, 90% of them were satisfied with the way information was presented to them in the MCSI, although the Wikipedia API and underlying retrieval method was same for both interfaces.

Highlighting: Another issue which was referred by some subjects related to text highlighting. Subjects found that the absence of highlighting in the CSI was frustrating.

7.4.6.3 RQ3: Does Highlighting important segments support users in effective and efficient search?, and Why?

92% of subjects liked the document highlighting options in the MCSI. There is a potential for this highlighting facility to be further extended in the MCSI where the agent could develop an incremental model of the user's developing engagement in the search task. The following reasons were identified for choosing these options as shown in Figure 7.8.

1. Interactive and Engaging: Around 28% of subjects claimed that they were able to engage and interact with documents better by using the highlighting options.

2. Helpful: 68% of the subjects found highlighted documents helpful in information seeking.

3. Reduce the Cognitive Load: Around 24% of the subjects believed that the highlighted documents reduced their cognitive load.

4. Access to Relevant information: 36% of the subjects believed that highlighting documents helped them to more easily access useful information.

7.4.6.4 RQ4: What are the challenges and opportunities to support exploratory search in conversational settings?

The great majority of subjects (92%) claimed that the MCSI was better. The remaining subjects faced some challenges using it as shown in Figure 7.9. Subjects wanted more corresponding sub-sections in the documents to support their exploration, and also wanted support image search. Around 4% of the subjects felt the need for improvement in operational speed and the better incorporation of standard features, such as Spell Checking. Subjects found the chat interface helpful for exploring long documents. They were keen to see the addition of speech as a mode of user interaction and a more refined algorithm for selection of images for presentation to the user.

Subjects appreciated the usefulness of the interface in supporting exploratory search, but suggested that this would be further improved by the incorporation of a question answering facility.

Confidence intervals (p=0.05) per scale							
Scale	Mean (-3 to 3)	Std. Dev.	N	C	C interval		alpha value
P	-0.720	1.349	25	0.529	-1.249	-0.191	0.91
H	-1.640	1.233	25	0.483	-2.213	-1.157	0.92
Overall	-1.180	1.207	25	0.473	-1.653	-0.707	0.91

Table 7.9: CSI confidence intervals on UEQ-S where, 'P' stands for Pragmatic Quality, 'H' stands for Hedonic Quality and 'C' stands for Confidence.

CSI Confidence intervals (p=0.05) per scale							
Scale (-3 to 3)	Mean (-3 to 3)	Std. Dev.	N	C	C Interval		alpha value
P	1.310	0.596	25	0.234	1.076	1.544	0.79
H	1.600	0.559	25	0.219	1.381	1.819	0.79
Overall	1.455	0.519	25	0.203	1.252	1.658	0.79

Table 7.10: MCSI confidence intervals on UEQ-S, where 'P' stands for Pragmatic Quality,'H' stands for Hedonic Quality and 'C' stands for Confidence.

7.4.6.5 RQ5: How does user experience vary between search settings in comparison to each other?

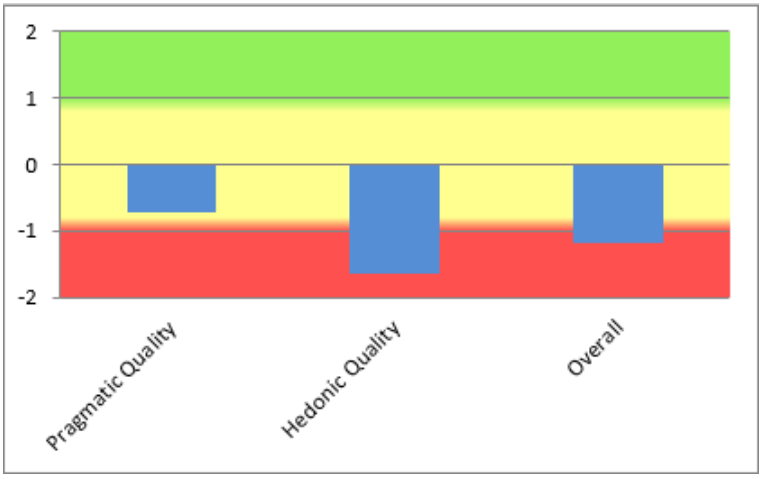


Figure 7.10: Means of the CSI score scales Pragmatic quality and Hedonic quality where green indicates positive, yellow indicates neutral and red indicates bad where the x axis represents 'Quality' and the y axis represents 'scale'.

1. **Observing the Pragmatic and Hedonic properties of CSI**: The users provided feedback based on their experience using the CSI. As shown in Figure 7.10, the CSI score is negative with respect to both Pragmatic and Hedonic properties and the overall score is also negative from which we can infer that the user's experience of the CSI system is neither effective nor efficient as shown in the Table 7.9. From Table 7.9, we can calculate mean range after data transformation for UEQ-S where is -3 too negative and +3 is too positive. Table 7.9 shows the confidence interval and confidence level. The smaller the confidence interval the higher is the precision. The confidence interval and confidence level confirm our analysis that all the dimensions of Pragmatic and Hedonic properties were negatively experienced by the users. Generally, items belonging to the same scale should be highly correlated. To verify the user consistency, alpha-coefficient correlation was calculated using the UEQ-S toolkit. As per different studies, alpha value > 0.7 is considered sufficiently consistent [150]. This shows that user marking of the conventional system is consistent. The UEQ-S tool kit also provides an option to detect random and non-serious answers by the users. This is carried out by checking how much the best and worst evaluation of an item in a scale differ. Based on this evaluation, the users' feedback does not show any suspicious data.

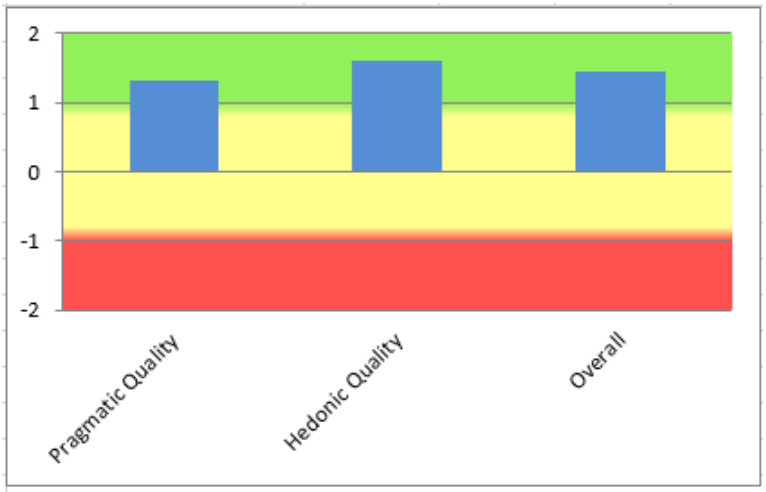


Figure 7.11: Means of the MCSI score scales Pragmatic quality and Hedonic quality where green indicates positive, yellow indicates neutral and red indicates bad where the x axis represents 'Quality' and the y axis represents 'scale'.

2. **Observing the Pragmatic and Hedonic properties of MCSI**: As shown in Figure 7.11, the MCSI scored positive in Pragmatic, Hedonic and Overall score from which infer we can that the user's experience of the MCSI is good in general and with good ease of use. Table 7.10 shows the confidence interval and confidence level. The confidence interval and confidence level confirms our analysis that all the dimensions of pragmatic and hedonic scores were positively experienced by the users. Alpha-coefficient correlation [150] value confirms that the marking of MCSI by the users is consistent. The UEQ-S toolkit also provides an option to detect random and non-serious answers by users. This is conducted by checking how much the best and worst evaluation of an item in a scale differ. Based on this evaluation, the users' feedback does not detect any suspicious data.

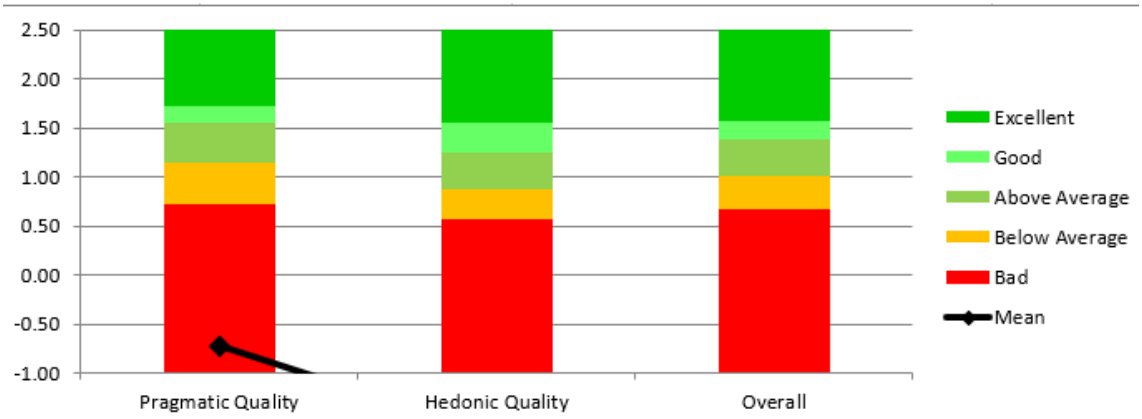


Figure 7.12: Comparison of benchmark with CSI where the x axis represent 'Quality' and the y axis represents 'Scale'.

7.4.6.6 RQ6: How does user experience vary for both search settings in comparison to a standard benchmark?

1. **Comparison of the CSI with the standard benchmark**: This benchmark was developed based users on feedback on 21 interactive products [150]. Based on the comparison from the benchmark, the CSI UX is far below the mean of the interactive products (Pragmatic Quality < 0.4, Hedonic Quality < 0.37 and overall < 0.38) as shown in Figure 7.12. This signifies that the UX with

the CSI needs major improvement on Pragmatic and Hedonic sectors. In the comparison to the benchmark, the CSI rates as a low quality of user experience and lies in the range of worst 25% of the products.

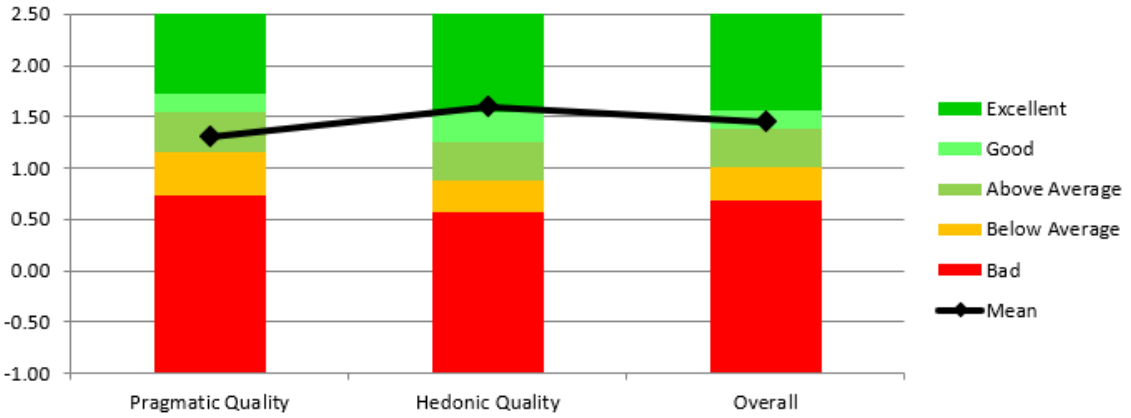


Figure 7.13: Comparison of benchmark with MCSI where the x axis represent 'Quality' and the y axis represents 'Scale'.

2. **Comparison of the MCSI with the standard benchmark**: Based on the comparison from the benchmark [150], the MCSI UX is far above the mean of the interactive products (Pragmatic Quality > 0.4, Hedonic Quality > 0.37 and overall > 0.38) as shown in Figure 8.10. This signifies the UX of the MCSI compared to other interactive products (benchmark) is very high and is of excellent level, and lies in the range of 10% best results.

7.4.6.7 RQ7: What are the user search behaviours and experiences with the conversational system in an exploratory search setting?

We also investigated user behaviour patterns in the conversational setting based on feedback from the pre- and post- questionnaires, the number of queries, and the number of unique documents viewed by the users as shown in Figures 7.14, 7.15 7.16 7.17 7.18 7.19. We do not observe very strong correlation among the variables. As per our observations of individual users, some users had less knowledge about a topic, but were interested to learn more about the topic which led them to engage with limited content with the least interaction. They were happy with the highlighted text in the long documents. While some other users reported less knowledge about the

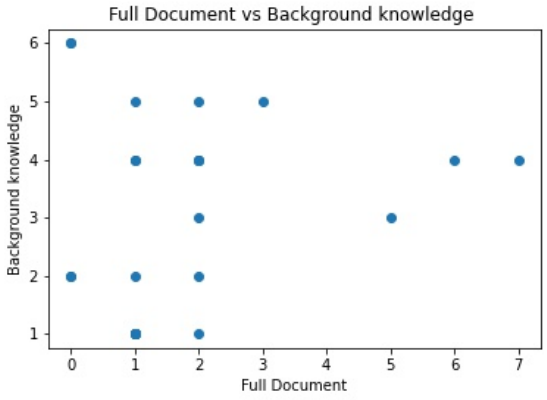


Figure 7.14: Correlation graph of completion background knowledge vs full document open (Correlation coefficient 0.16)

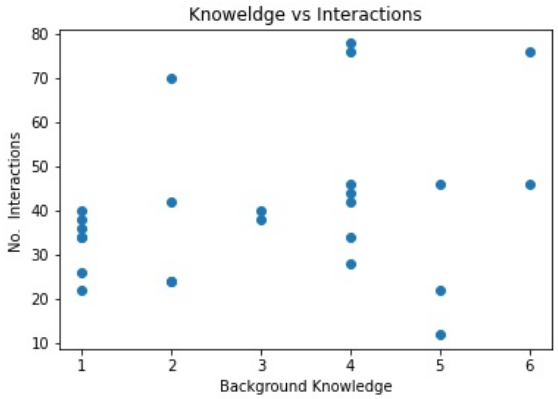


Figure 7.15: Correlation graph of completion background knowledge Vs interactions (Correlation coefficient 0.29)

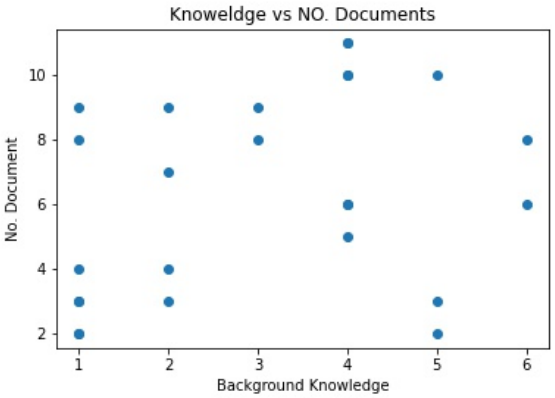


Figure 7.16: Correlation graph of completion background knowledge vs total documents (Correlation coefficient 0.31)

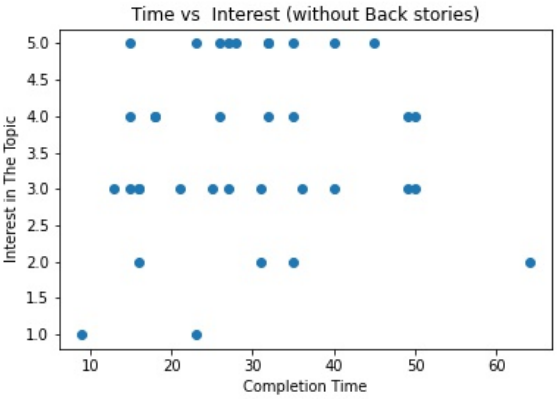


Figure 7.17: Correlation graph of Interest Vs Interactions (Correlation coefficient 0.32)

topic, but being highly interested in the topic which apparently led them to engage more with the snippets, sections and subsections with the multiple interactions, than those with who reported less interest in the topic. Other users reported only average knowledge about the topic, but their interest in the topic was highest which led them to engage more than the users listed above with the snippets, sections, subsections and full documents using multiple interactions.

Other users reported low background knowledge of a topic with their interest in it being average, which apparently led them to engage less with the snippets, sections, subsections and full documents. For these users, their pre-search anticipation of

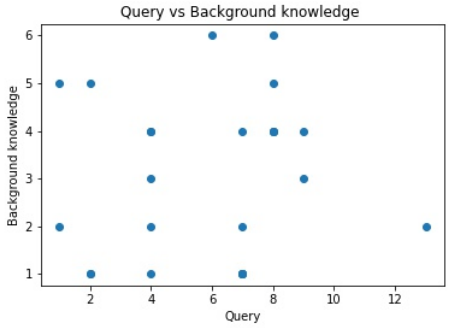


Figure 7.18: Correlation graph of completion background knowledge Vs No. Query (Correlation coefficient 0.07)

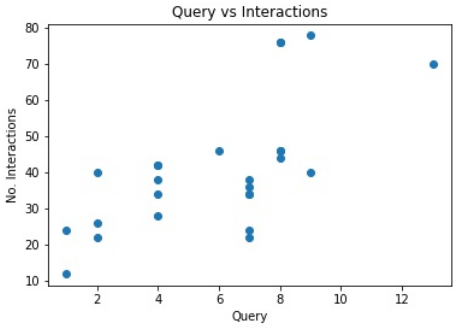


Figure 7.19: Correlation graph of total interactions Vs No. Query (Correlation coefficient 0.29)

difficulty level was consistent with their post-search experience of the difficulty level. It was found in the study that in the pre-search questionnaire users anticipated higher difficulty when using the MCSI, whereas the actual difficulty level reported post-search after using it was much lower. This indicates that the users' experience with the MCSI was intuitive and pleasant.

7.5 Conclusions and Observations

This chapter has reported a study indicating that subjects found our MCSI more helpful than a closely matched CSI. We also observed types of user behaviour while using MCSI which are different to those when using a CSI [137, 152]. Most previous studies of user behaviour in CS have used Wizard-of-Oz type agents [173], in contrast, we studied use of an automated search support agent. Using our agent-based system, we observe the natural expectations of user search in conversational settings. We observe that subjects do not encounter any difficulty in using the new interface because it seems to be similar to the standard search interface with the additional capabilities of conversation. We also observe the information space and its structure is a key component in information seeking. Subjects found highlighting important segments in long documents enables them to access information more easily. The MCSI made the search process less cognitively demanding and more

cognitively engaging.

The current prototype is limited with the use of multi-modality and uses only text modality. This could be further developed to use mixed modalities such as images and spoken responses to satisfying the information need of the user.

In comparison to current state-of-the-art conversational assistants. e.g Amazon Echo Show using Alexa, our prototype can be completely customized and explored further to provide a richer exploratory search experience to the user, while third party applications like Alexa are limited to open source customization and are not designed to support exploratory search, while our prototype provides the opportunity to explore sources and satisfy user information needs.

Clearly our existing rule-based search agent can be extended in terms of functionality. In the next chapter we investigate a extended version of the MCSI revised based on feedback from participants in the experiments reported in this chapter.

Chapter 8

A Machine-learning Based Agent for a Conversational Search Assistant

In this chapter we examine an alternative approach to the assistant agent in the Multiview Conversational Search Interface (MCSI) where the rule-based agent approach examined in Chapter 7 is replaced with a machine learning approach. Revision to the MCSI also includes the addition of new functionalities including image search, input and output audio support, and a more advanced full document option. These functionalities were developed based on suggestions provided by the subjects in the investigation described in Chapter 7.

This chapter evaluates the machine learning based MCSI and corresponding CSI based on the IECSI introduced in Chapter 4. Moreover, we statistically compare the mean of the IECSI for the rule-based system MCSI in Chapter 7 with the IECSI machine learning based system.

In this chapter, we investigate partially RQ3, RQ4 and partially RQ5 as presented in Section 1.5.3, Section 1.5.4 and Section 1.5.5 of Chapter 1. The answers to these questions obtained in this Phd study are summarized in details in Section 9.2.3, Section 9.2.4 and Section 9.2.5 of Chapter 9

8.1 Motivation

In Chapter 7, we observed that the rule-based MCSI can enhance user experience in complex information seeking. However, the rule-based system is defined by manually constructed rules, which restrict its functionality. The user must enter the correct instruction to operate the agent for information seeking, limiting its flexibility. The rule-based MCSI demonstrated that the interface can be effective and efficient for the dialogue-based information-seeking tasks. However, it also suggested the potential to investigate the system using an alternative machine learning based agent. Machine learning can enhance flexibility and provide a more robust solution for assistant agents, and may also offer the opportunity to automate agent learning based on user engagement.

A significant issue for machine learning in conversational search is the lack of training data. To overcome the issue of the limited data and provide a functional MCSI system to the user, we propose to combine the rule-based strategy with machine learning techniques to enable each process to compensate for the limitation of the other method. We refer to this hybrid combined strategy for MCSI as our machine learning approach. The intent and entities are extracted by the machine learning techniques and the rule-based method selects the action to execute after the intent identification. Our combined method is implemented using the python tool kit named RASA[1]. Intents are used to specify what the chatbot needs to respond to when it detects a user's intent. Entities are logically organized chunks of data within a user's message, which act like a query for the search system. For example, "tell me something about Ireland" where "Ireland" represents an entity (country). To enable our model to correctly classify entities, we created a small corpus of examples which contain entities and intentions to train our model in information seeking activities. Details of this are provided later in this chapter.

Another motivation for this investigation is incorporate suggestions provided by the partcipants in the study, described in Chapter 7. Users suggested extending the

[1]https://rasa.com/

original MCSI to enable image search, audio input and output support, handling fact-based questions, and intelligent navigation for entire documents.

Another motive for this study is to compare the mean of all the evaluation parameters of IECSI for both the settings; the rule-based MCSI discussed in the previous chapter with the machine learning approach based on the MCSI study discussed in this chapter.

The whole study is evaluated based on the IECSI described in Chapter 4. Our prototype supports interaction with the interface using both text and speech. The MCSI supports both text and image search, details of its functionalities are given in Section 8.2.

The structure of the chapter is as follows: Section 8.2 describes the MCSI and CSI functionality, implementation and dialogue taxonomy; Section 8.3 explains the methodology followed during the study which includes the experimental setup and procedures, Section 8.4 explains hypothesis testing, and reports analysis and investigation of the Research Questions (RQ), Section 8.5 provides discussion, and Section 8.6 gives conclusions and considerations for further study.

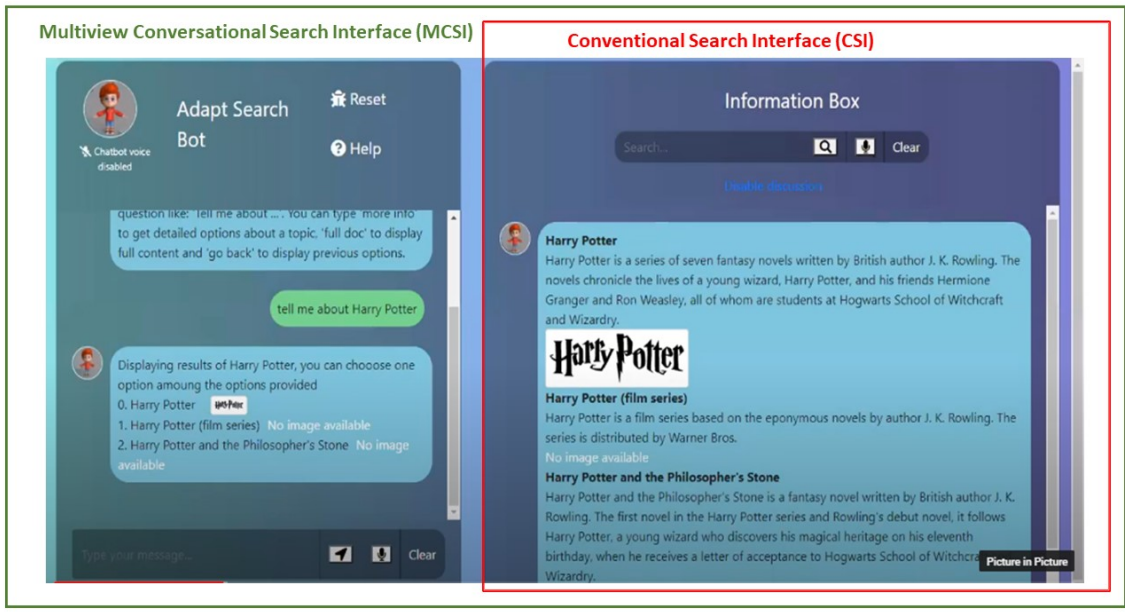


Figure 8.1: Conversational search interface incorporating: chat display, chat box, information box, query box, help-button and retrieved snippets and documents. The green outline indicates the MCSI setting and red block indicates the CSI setting.

8.2 Conversational Search Prototype System

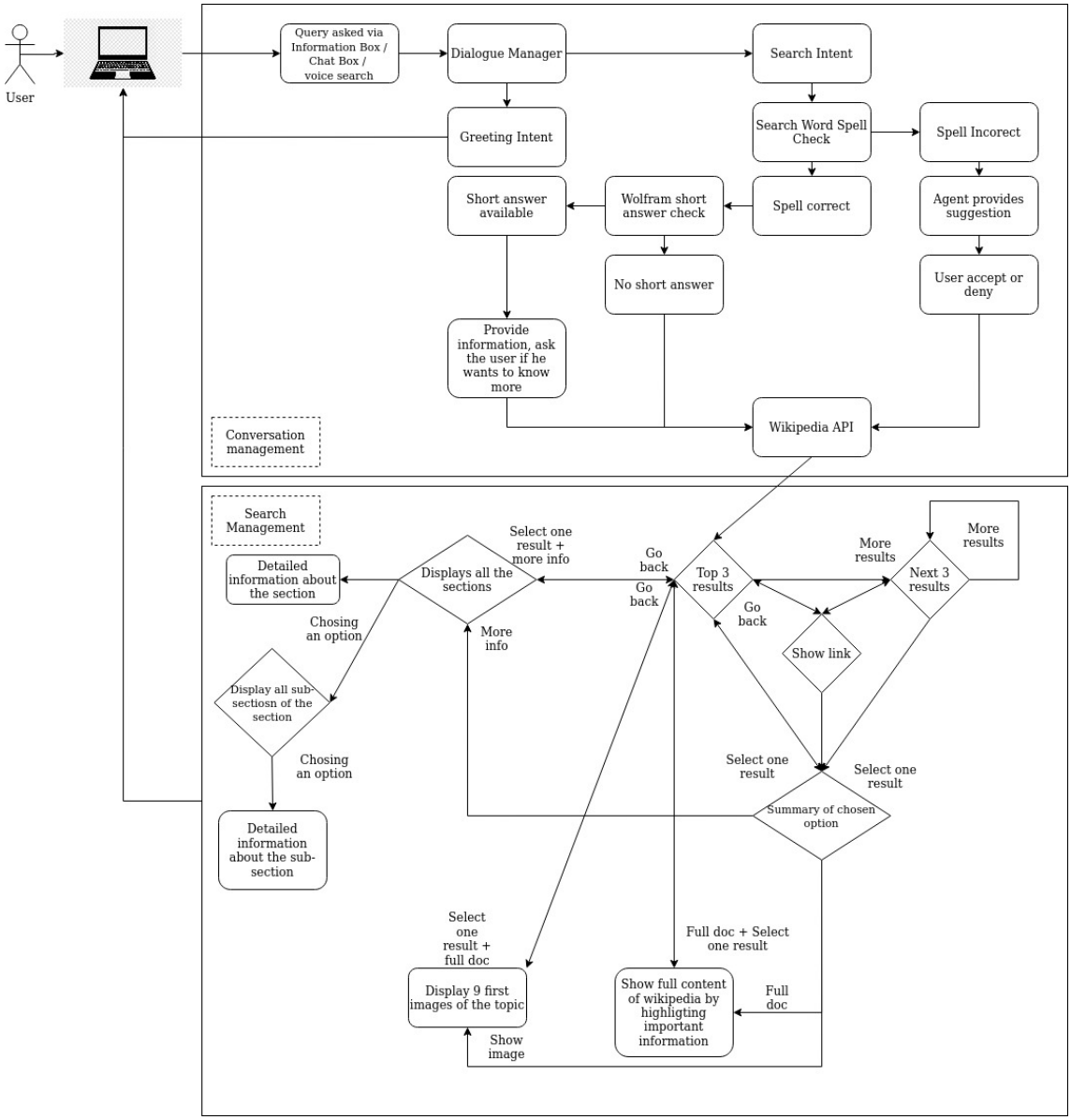


Figure 8.2: Workflow of the Conversational Model.

The conceptual design of the interface for our prototype conversational search (CS) system is shown in Figure 8.1. This is same as the introduced one in Chapter 7, but there is a refinement of the functional aspect after the feedback from the study in Chapter 7 as shown at link[2]. This interface includes the following components:

1. Help Button: Displays information of interface commands and functionality.

2. Chat Display: Shows dialogue between the agent and the user.

[2]`https://tinyurl.com/y5phk9f9`

3. Chat Box: Enables the user to engage with the agent.

4. Information Box: Displays significant features of retrieved items.

5. Query Box: Enables use of standard search interaction by entering a query.

6. Action Button: Enables the following operations:

   (a) Enter: Enter text or query in Chat Box or Search Box.

   (b) Clear: Clear the text from the Chat Box.

   (c) Audio input: Allow the user to enter audio[3] input (represented by the dark colour microphone sign).

   (d) Reset: Clear chat history from the Chat Box.

   (e) Chatbot voice enabled/disabled: Enable/disable audio response from the search agent.

   (f) Disable discussion: Disable the search agent, information panel then functions like a conventional search system .

As in the previous version of the MCSI, the search agent communicates with the user via the Chat Box; input and output interaction with the Chat Box can now be either typed or spoken. The agent again performs various actions, including seeking clarification of ambiguous queries, suggesting words for use in revised queries, or showing key details from retrieved items. The Information Box shows key information from retrieved items and using this, the user's attention can be directed to key details in retrieved content.

8.2.1 System Implementation

Similar to the previous MCSI system, the machine learning based MCSI is divided into two sections: a Web Interface and a Logical System.

[3]https://blog.rasa.com/webkitspeechrecognition-and-rasa/

8.2.1.1 Web Interface

Similar to Chapter 7, this is constructed using the lightweight web server gateway interface (WSGI) web application python framework[4], with HTML, CSS, and JS toolkit known as bootstrap[5].

8.2.1.2 Logical System

This is responsible for conversational engagement with the searcher and search management. The RASA toolkit[6] powered by a fined tuned BERT model [145, 178] and spacy[7] is used to identify intent and entities to understand the user need, and to access the spell checking API[8]. The Wikipedia API is used for the backend revised search process. We have refined the search component in this CS system by encapsulating wolfram alpha (fact based answers)[9] to support the search process.

8.2.1.3 Operations

We have refined the commands used in the Chapter 7. The search assistant can accept the following user commands:

1. Open {*Document or topic number or title*}: Show the snippet of selected document or topic.

2. More Info {*Document number or title*}: Display all available metadata from the document, this allows users to explore a topic in depth.

3. Show Image {*Document or option number or title*}: Show images from the current document.

4. Go Back: Displays the previous results.

5. More Results: Displays more search results.

[4]https://palletsprojects.com/p/flask/
[5]https://getbootstrap.com/
[6]https://rasa.com/
[7]https://spacy.io/
[8]https://pypi.org/project/pyspellchecker/
[9]https://www.wolframalpha.com/

6. Search {*Query keyword*}: Starts search for new query, displays highest scoring results for the query.

7. Search Image {*Query keyword*}: Shows top scoring images from documents for the query using image captions.

8. Full Doc {*Document number or title*}: Shows full document with summary of highlighted top scoring sentences (based on custom algorithm introduced in Chapter 6) at top of document. Clicking on a highlighted sentence navigates to corresponding paragraph.

9. Show Links: Display links to complete documents

8.2.2 Dialogue Strategy and Taxonomy

In this approach, we follow the dialogue strategy and taxonomy introduced in Chapter 7 which is divided into three phases with multiple states [187]. The three phases are: identification of user information need, presentation of results in the chat system, and continuation of the dialogue until the user is satisfied or aborts the search. Example of states: are the user or agent request information, the agent or user offer information, the agent answers, the user or agent accept, and the user or agent accept to refuse. The states were outlined in detail in Chapter 7.

In this section, we introduce the RASA toolkit used for the machine learning (ML) MCSI approach, the architecture of the working ML model and the system workflow of the MCSI. Rasa is an open source machine learning-based computational intelligence architecture for automatic text and voice-based interactions where it can connect to messaging channels and APIs, understand messages and be trained for complex chit-chat systems [178].

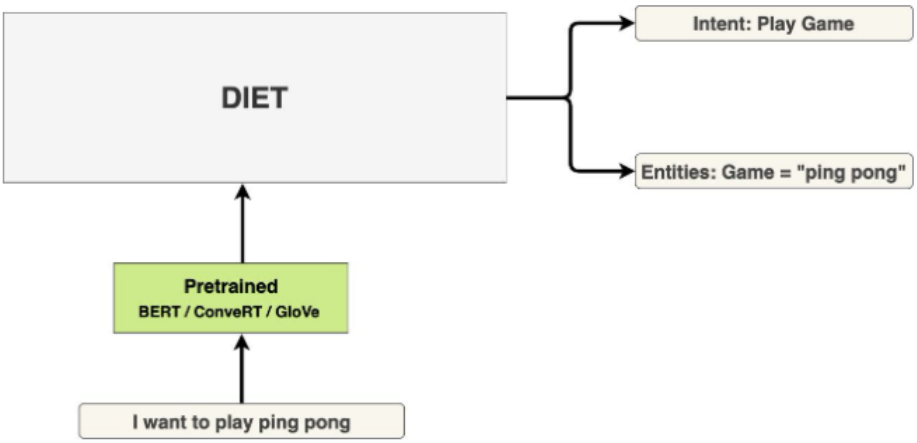


Figure 8.3: Illustration of high level of BERT DIET [178].

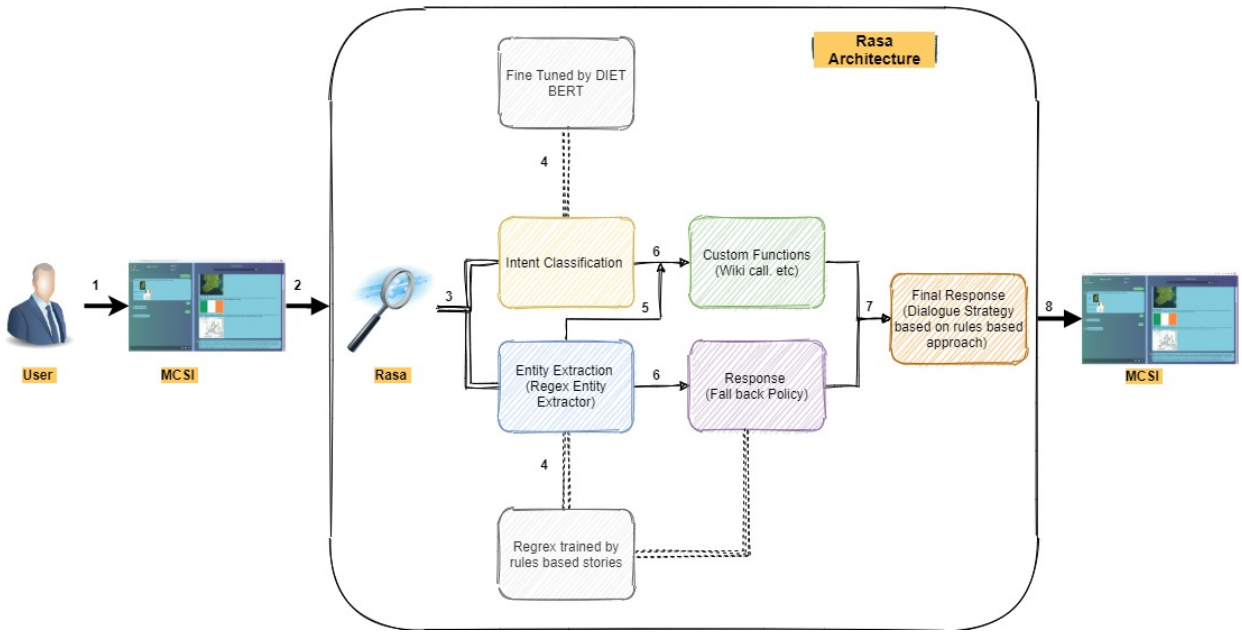


Figure 8.4: Illustration of RASA architecture in information seeking.

8.2.2.1 The RASA Toolkit and its Components

The ML based MCSI is trained using the RASA toolkit[10]. The overall RASA functionalities are divided into three steps, as shown in the Figure 8.5. The details are as follows:

- **Natural language understanding data**: For an agent to recognize what a searcher is saying, we need to provide samples of expected user utterances which can be used to train the model to identify the idea or the intention of

[10]https://rasa.com/

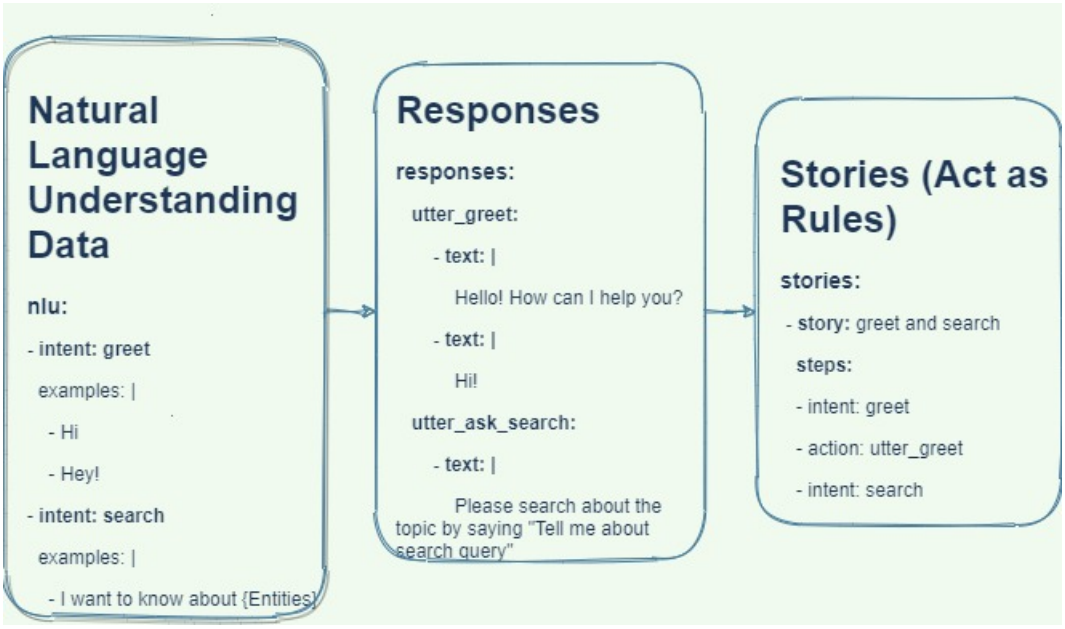


Figure 8.5: Flow of the RASA toolkit.

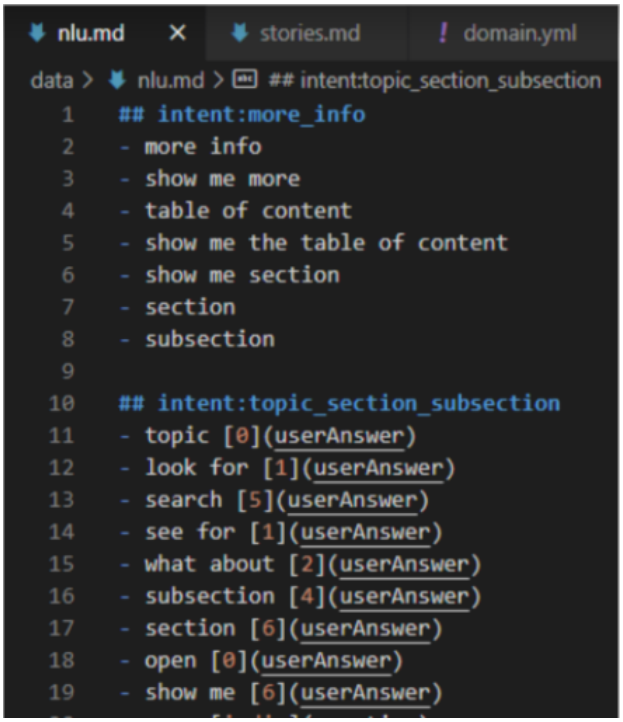


Figure 8.6: Sample of training data where ## sign define intents and ( ) reflect users defined entities.

the user input correctly. We group the collected utterance samples according to our purposes, vision, goal or intention; the utterances are referred to as intents.

For example, in the Figure 8.5, we have two intents “Greet” and “Search”.

The 'Greet' contains example messages like "Hi", "Hey". In the second intent 'Search', we define entities that generally signify a noun. For example, name, place, animal and thing. Examples of intents and entities are shown in Figure 8.6.

The extracted entities are used as a query object. The queries are extracted by the *Regex Entity Extractor* component of RASA which extracts the entity using regular expressions based on the patterns learned from the training data.

Training data was prepared from observations obtained from our user studies on our rule-based agents, as described in Chapter 7. The training data includes 18 intents, including search, full document, more information. Each intent is trained using around 80 utterance, around 3 entities and 18 custom actions for fine tuning the model. In total, we have around 4000 different combinations of utterances and entities for training the model (example shown in Figure 8.6).

Since we have only a limited amount of training utterances, we decided to fine-tune the pre-trained language model DIET BERT (Pre-training of Deep Bidirectional Transformers with Dual Intent and Entity Transformer) provided by RASA [146, 175] for intent and entity classification. Fine tuning is considering our custom defined intents for the MCSI as shown in Figure 8.3.

Epoch is a hyper-parameter controlling the iterations of data passing forward and backward through the model during training. One can vary the epoch size to explore the accuracy of the model for variations in training. To select the most suitable epoch, we used the standard RASA evaluation process. For development purposes, the training data was split using the train and split method, where 80% is eval training data, and 20% is eval testing data. The data is shuffled to avoid any bias or variance effect in training and to prevent the chances of over-fitting and under-fitting errors. Multiple variations of epoch were tested on the dataset to fine tune the model. The model in this PhD research was trained on multiple variations of epochs (50, 100, 150, 200, 250, 300) using training data. 100 epoch was food to be the most suitable for

our investigations based on the highest F1 score for correct intent identification and correct entity extraction for our system. We used the default learning rate and optimizer set by RASA for obtaining the best results [178].

- **Responses**: Once the agent understands a user's intent by classifying the utterance, the agent sends an answer back to the user as shown in Figure 8.5. For each intent, there is a predefined response and actions. For example, if the user greets an agent then the agent must reply with a greeting response. As can be seen in Figure 8.5, there are two responses for any greet intent utterances; the best response is selected based on the fall back policy of RASA. The memorization policy fall back was used in RASA, which is responsible for selecting the best response of each utterance and to handle any mistakes or input provided by the users, by selecting the best fall back response or action from the list of available responses or actions. The response is selected based on the confidence score provided by the model. This confidence score is a prediction probability score. Suppose for an utterance, if we have three responses, then based on the user utterance, the trained RASA model predicts confidence scores for each response. The response with the best confidence score is selected.

  RASA provides flexibility to encapsulate custom functions in their framework. The actions shown in Figure 8.5 are python functions written to perform operations, as introduced in the above sections. In RASA, we can provide a response or action that is triggered once the RASA identifies intent. For example, if my intent is to search from Wikipedia, then RASA will trigger the python functions assigned the search. For example, a custom function to call the Wikipedia API to search or a function to open a full document.

- **Stories**: Stories are examples of conversations that are used to train an agent in RASA based on a fall back policy to answer what the user said previously correctly.

As shown in Figure 8.5, the story format shows the user's intent, followed by the agent's action or response. The stories guide the conversation flow where the agent helps the user to accomplish their goal. The contents of the stories are from Chapter 7.

8.2.2.2 Architecture of RASA Processes

As shown in Figure 8.4, users express their information need to RASA via the MCSI. RASA extracts the entity and identifies the intention of the user based on the trained model. The intent and entity are identified by the trained model. In parallel, the *Regex Entity Extraction* component extracts the entities from the user utterance.

The next stage is divided into two subprocesses, which run in parallel and are related to each other. One calls the user-defined function (known as the action) based on the intent, for example, calling the Wikipedia search API. The other predicts the response utterance to the user through fall back policy. Every intent has multiple response utterances. These responses are selected based on the prediction of the fall back policy as explained in the above section 8.2.2.1. There are multiple fallback policies, and some are trained based on user stories or rules while other based configuring custom policies. The second last stage is to combine the output from the action or function and the response, which again comes from the rule-based approach of the dialogue strategy from Chapter 7. The final stage is to show the combined results to the users. For example, Wikipedia search API triggered by the RASA, provided you three search results on the query of "Mahatma Gandhi" and the best response utterance selected by RASA is "Here are the". The final utterance would be "Here are the *Result list (three documents)*".

8.2.2.3 Workflow of Machine Learning based MCSI

The system workflow is divided into two sections: Conversation Management and Search Management, as shown in Figure 8.2. For the workflow, RASA components are mainly responsible for conversational management, and custom functions are

accountable for the search management. These custom functions are associated with the intents. As soon as RASA identifies an intent, it triggers the custom function in the search management based on the specific purpose, For example, if the user's input is to open a full document, then RASA identifies the intent of the user (conversation management) and triggers the custom function which is responsible for opening the full document of the topic with text highlighting (search management). Additionally, the appropriate response of user utterance is also provided in the chatbox (conversational management).

The details of the conversational management and search management are described below. Conversation management includes: a Dialogue Manager, spell checker and an image search API. The Dialogue Manager validates the user input and decides whether to send to search management or not.

Search Management is responsible for search and displaying the top ranking search results to the user in the Information Box. It can give detailed information about a selected item or display more results. There is also an option to look at all available information about an item. In user engagement, the user commences a search from the Information Box, the assistant initiates a dialogue to guide them in the search process. The system also provides support to a user who wants to examine a specific retrieved item.

In comparison to Chapter 7, this workflow of the ML based MCSI is very different. In the rule-based MCSI, we have to code the utterance, query extraction patterns based on the rules, and work based on the correct user input and unable to entertain any unknown information. So, the conversation management of rule-based is not very robust to errors while the ML based MCSI learned from the rules provided and can easily handle complex user input.

Additionally, ML based MCSI has a more robust mechanism for query extraction compared to the rule-based MCSI, which can lead to better search management. Moreover, the ML based MCSI, as discussed in Section 8.2.2.1, can learn from the stories or sample conversational cases and enable the system to identify the patterns,

which allows a better entity extraction and smarter responses to the user. The ML based MCSI also has extended functionalities to provide better user experience. These extended functionalities are discussed in details, in Section 8.1 and Section 8.2 of this chapter.

8.3 User Study Methodology

In this section we describe the details of our user study. The study aims to enable us to observe and better understand and contrast the behaviour of searchers using the CSI and the machine learning based search agent in the MCSI.

To enable direct comparison between the MCSI and the CSI, the CSI is again formed by using the conversational interface with the agent panel removed. The searcher enters their query in the Query Box, and document summaries, etc are returned by the Wikipedia API in the previous investigation described in Chapter 7.

For the investigation of our revised MSCI, we again use the UQV100 backstories [106] used for the previous study described in Chapter 7.

Experiment	Age	No. Male (M)	No. Female (F)	Ratio (F/M)
Chapter 8 (MCSI ML based Agent both settings)	18-25 26-35 36-45	5 3 0	4 4 6 0	4:5 4:3 NA
	Total	8	8	1:1

Table 8.1: Details of age distribution throughout PhD project.

8.3.1 Experimental Procedure

Participants in this study were again required to complete a search session consisting of multiple search tasks based on selected backstories. As a part of their search session they again had to complete questionnaires before and after undertaking each search task. The experimental setup and procedure is same as Chapter 7. The data provided in the Table 8.1 is for those participants whose data has been used for analysis. Approval was again obtained from the DCU Research Ethics Committee prior to the data collection.

The questionnaire again used a Google form divided into three sections, as discussed in Chapter 4 and 7.

Similar to Chapter 7, a pilot study with two undergraduate students in Computer Science was conducted using two additional backstory search tasks to check the usability for the extended functionalities.

For the main study, as in the previous study, each participant was assigned two of the 12 search task backstories with the expectation that as in the previous study their overall session would last around one hour. The pairs of backstories for each session were again arranged using a Latin square procedure [3].

8.3.2 User Study Details

In total, 17 subjects participated in this study (excluding the pilot study), we examined the data of 16 subjects, since 1 subject was found not to have followed the instructions correctly. The study was conducted in two phases, similar to Chapter 7. As before, after completing the questionnaires, the participant also attended a semi-structured interview after completion of the whole task in for both interfaces. The interviews were thematically labelled by two independent analysts. The modified MCSI and CSI were again compared using our implicit evaluation method to examine the dimensions of: user experience, cognitive load, software usability, search experience and knowledge expansion.

8.4 Hypothesis & Research Question Analysis

This section introduces the hypothesis and research questions examined in this study based on the framework presented in Chapter 6.

8.4.1 User Experience

Following Chapter 7, we again tested the hypothesis relating to interactive user experience for the machine learning based MCSI and CSI using a UX questionnaire for interactive products known as User Experience Questionnaire (UEQ-S) [66, 132, 150] similar to Chapter 7.

Negative	Positive	Scale	CSI_Mean	MCSI_Mean	P_Values
obstructive	supportive	P	3.25	6.56	1.09E-07
complicated	easy	P	3.12	6.68	3.98E-08
inefficient	efficient	P	2.56	6.18	7.87E-09
confusing	clear	P	3.68	6.43	2.51E-05
boring	exciting	H	2.31	6.31	3.40E-12
not interesting	interesting	H	2.12	6.37	9.81E-13
conventional	inventive	H	1.75	6.31	6.62E-12
usual	leading edge	H	1.75	6.12	3.37E-11

Table 8.2: UEQ-S score for the CSI and the MCSI, where 'P' stands for Pragmatic Quality and 'H' stands for Hedonic Quality.

1. **HO: Users feel a similar interactive experience during the search with either of the interfaces**: Similar to Chapter 7, the user evaluated the system based on 8 parameters as shown in Table 8.2, with a grading scale between 0 (low) - 7 (High). We compare the mean difference of both systems on all parameters. These results show that the user experienced a more pleasant experience and ease of use with the MCSI.

8.4.2 Cognitive Dimensions

Similar to Chapter 7, the user was asked to evaluate the CSI and MCSI in 6 dimensions [15] of cognitive load as shown in Table 8.3.

NASA Task Load	CSI Mean	MCSI Mean	P Values
How mentally demanding was the task?	5.62	2.93	2.67E-05
How physically demanding was the task?	4.12	2.06	0.0052
How hurried or rushed was the pace of the task?	3.75	2.00	0.0047
How successful were you in accomplishing what you were asked to do?	3.12	6.31	1.16E-07
How hard did you have to work to accomplish your level of performance?	5.93	2.18	1.48E-10
Overall, I am satisfied with how easy it was to use this system	2.62	6.75	1.24E-09

Table 8.3: NASA task load score for CSI and MCSI.

1. **HO: Users experience a similar task load during the search with either of the interfaces**: Similar to Chapter 7, the null hypothesis ($P<0.10$) was rejected on the basis of the Task Load index. This means that the user experienced less subjective mental workload when using the MCSI.

8.4.3 Software Usability

Similar to Chapter 7, software usability was evaluated using 15 parameters. For each dimension, the machine learning based MCSI outperformed the CSI as shown in Table 8.4.

1. **H0: User Psychometric Evaluation for the conversational interface and conventional search has no significant difference:** Similar Chapter 7, the null hypothesis was rejected and the H1 hypothesis was accepted, which is that the MCSI performs better than the CSI.

8.4.4 Search Experience via Learning

Based on the search as learning Vakkeri model [121], the user search experience can be evaluated in a similar way to Chapter 7.

1. **H0: Subjects find no significant difference between using either of the interfaces:** Similar to Chapter 7, it was found that the null hypothesis

System Usability (IBM)	CSI Mean	MCSI Mean	P Values
It was simple to use this system	3.06	6.81	2.05E-07
I can effectively complete my work using this system	2.00	6.56	1.97E-13
I am able to complete my work quickly using this system	1.93	6.50	1.31E-11
I am able to efficiently complete my work using this system	1.87	6.37	1.05E-12
I feel comfortable using this system	2.25	6.50	7.01E-12
Whenever I make a mistake using the system, I recover easily and quickly	2.43	6.062	3.76E-08
The information (such as online help, on-screen messages) provided with this system is clear	2.625	6.56	1.69E-10
It is easy to find the information I needed	2.12	6.62	1.55E-14
The information is effective in helping me complete the tasks and scenarios	2.37	6.62	4.25E-12
The organization of information on the system screens is clear	3.37	6.62	1.01E-08
The interface of this system is pleasant	3.12	6.12	1.20E-07
I like using the interface of this system	2.68	6.312	4.93E-10
This system has all the functions and capabilities I expect it to have	2.00	6.25	3.49E-13
Overall, I am satisfied with this system	2.25	6.37	2.68E-13

Table 8.4: System Usability Score for CSI and MCSI.

Search Mode	User Search Experience	CSI Mean	MCSI Mean	P Value
Pre-search	Pre-existing knowledge about search topic	1.93	2.87	0.089
	Interest about this topic	6.25	6.25	1
	Anticipated difficulty to find the information needed	3.625	3.8125	0.71
Post-search	Actual difficulty to find the information needed	4.93	1.68	1.25E-08
	Helpful in satisfying information need	2.73	6.40	5.71E-09
	Quality of Text presented	3.25	6.125	9.44E-08
	Usefulness of search results	2.56	6.31	9.50E-08
	Usefulness of text shown	2.25	6.25	2.60E-07
	Cognitively engaging	2.5	6.56	2.50E-10
	Expansion of knowledge about the topic	4.06	6.25	4.15E-05
	Understanding of the topic after search task.	3.93	5.93	3.03E-05
	Grade the success of your search session	2.93	6.25	4.38E-08
	Rating search setting in terms of understanding user inputs	3.31	6.12	5.97E-05
	Rating assigned search setting in the search results presentation	2.81	6.25	3.93E-06
	Rating the Image search options	1.31	6.00	3.40E-08
	Rating the voice search options	0.81	6.18	1.78E-12
	Rating the Navigation System for scanning full documents	0.81	6.62	2.55E-14
	Rating WH Questions/or fact Based Question response system	0.87	6.31	1.88E-13
	Rating the suggestion(s) skills	1.46	6.26	0.001

Table 8.5: Searching as learning parameter scores for CSI and MCSI.

was rejected, as shown in Table 8.5. Subjects search experience was statistically significantly better with the MCSI. In the pre-search questionnaire, subjects were asked to anticipate the difficulty level of the search before starting the search and in post search questionnaire, they were asked to indicate the difficulty level they actually experienced. It was observed that subjects pre-search anticipated difficulty level and the actual difficulty level after conducting the search increased for the CSI and decreased in the case of the MCSI significantly as shown in Table 8.6.

Confusion Metrics	Anticipated Difficulty	Actual Difficulty	P value
Mean_CSI	3.62	4.93	0.023
Mean_MCSI	3.812	1.68	2.23E-06
P value	0.71	1.25E-08	NA

Table 8.6: Anticipated Difficulty vs Actual Difficulty for CSI and MCSI.

8.4.5 Knowledge Expansion

Similar to Chapter 7, knowledge expansion (Change in Knowledge) was investigated using a comparison of pre-search and post-search summaries based on a number of parameters, as shown in Table 4.5 of Chapter 4.

The summary was scored against these three factors as shown in Table 4.5 of Chapter 4, by two independent analysts with the Kappa coefficient (Approx 0.85) [7].

CSI	Mean_Pre-Search	Mean_Post-Search	P value
DQual (0-3)	0.43	1.81	2.317E-7
DIntrp (0-2)	0.43	1.18	0.003
DCrit (0-1)	0.31	0.68	0.019

Table 8.7: Knowledge expansion (Change in Knowledge) score based on CSI.

MCSI	Mean_Pre-Search	Mean_Post-Search	P value
DQual (0-3)	0.62	2.43	2.317E-7
DIntrp (0-2)	0.62	1.75	0.003
DCrit (0-1)	0.18	0.81	0.019

Table 8.8: Knowledge expansion (Change in Knowledge) score based on MCSI.

Comparison CSI	Diff_Mean _CSI	Diff_Mean _MCSI	P value	Change Percentage (approx)
Diff-Pre-Post -Summ_DQual (0-3)	1.37	1.81	.061	15%
Diff-Pre-Post -Summ_DIntrp (0-2)	0.75	1.12	.189	19%
Diff-Pre-Post (0-1) -Summ_DCrit	0.37	0.62	.167	25%

Table 8.9: Knowledge expansion (Change in knowledge) for the CSI vs the MCSI.

1. **H0: No significant difference in the increase of the knowledge (Change in Knowledge) after completing the search task when using either setting**: As shown in Tables 8.7 and 8.8, the pre-search score and post search score for all three factors as shown in Table 4.5, were statistically significant for both of the search settings (T dependent). This implies that subjects expand their knowledge while carrying out the search.

2. **H0: Knowledge gain during the search is independent of the interface setting:** After concluding the alternative hypothesis, it was important to investigate whether one system was better in providing support to users in expanding their knowledge. We purposed and tested the following hypothesis. In this test, we compared the mean of the difference in the score for pre-search and post-search summaries for both settings. An independent T test was conducted on the change of the three parameter scores as discussed above for the hypothesis test as shown in Table 8.9. It was found that in the MCSI interface setting, subjects scored higher in change of critique, quality and interpretation. This implies that the subjects learned more while using the MCSI. The difference in quality score was statistically significant, while the other two parameters were not statistically significant due to low difference in the margin range. The quality and interpretation increased more than 15% while using the MCSI.

8.4.6 Research Questions

The details of the research questions are as follows:

CSI Confidence intervals (p=0.05) per scale							
Scale	Mean (-3 to 3)	Std. Dev.	N	Confidence	C interval		Alpha value
P	2.469	0.706	16	0.346	2.123	2.815	0.87
H	2.281	0.645	16	0.316	1.965	2.597	0.85
Overall	2.375	0.617	16	0.303	2.072	2.678	0.86

Table 8.10: MCSI confidence intervals on UEQ-S where 'P' stands for Pragmatic Quality, 'H' stands for Hedonic Quality and 'C' stands for Confidence.

Confidence intervals (p=0.05) per scale							
Scale	Mean (-3 to 3)	Std. Dev.	N	Confidence	C interval		Alpha value
P	-0.844	1.546	16	0.758	-1.601	-0.086	0.89
H	-2.016	1.270	16	0.622	-2.638	-1.393	0.94
Overall	-1.430	1.310	16	0.642	-2.072	-0.788	0.91

Table 8.11: CSI confidence intervals on UEQ-S where 'P' stands for Pragmatic Quality, 'H' stands for Hedonic Quality, and 'C' stands for Confidence.

8.4.6.1 RQ1: How does user experience vary between search settings in comparison to each other?

1. **Observing the pragmatic and hedonic properties of CSI**: Similar to Chapter 7, a shown in Figure 8.7, the CSI score is negative in pragmatic, hedonic, and the overall which implies that the user's experience of the CSI is neither effective nor efficient, as shown in the Table 8.11.

2. **Observing the pragmatic and hedonic properties of MCSI**: Similar to Chapter 7, a shown in Figure 8.8 and Table 8.10, the MCSI scored positive in pragmatic, hedonic, and the overall score which inference that the user's experience about MCSI is good and ease to use.

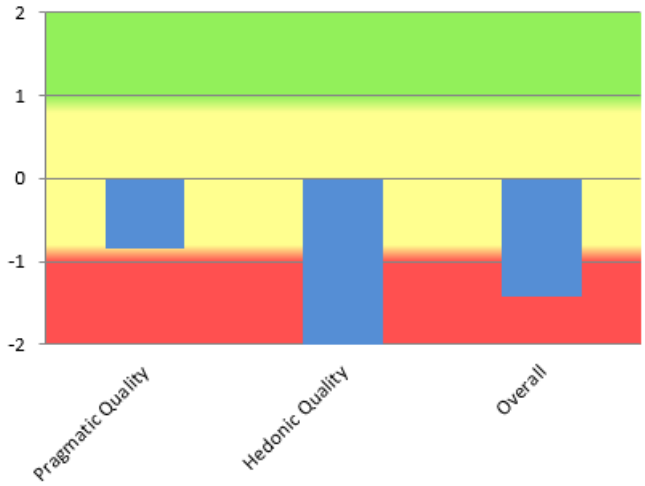


Figure 8.7: Means of the CSI score scales pragmatic quality and hedonic quality, where green indicates positive, yellow indicates neutral and red indicates negative, where the x axis represents 'Quality' and the y axis represents 'Scale'.

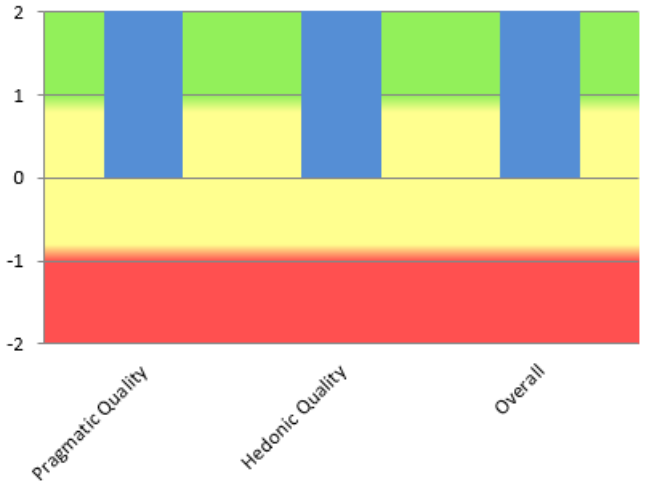


Figure 8.8: Means of the MCSI score scales pragmatic quality and hedonic quality where green indicates positive, yellow indicates neutral and red indicates negative, where the x axis represents 'Quality' and the y axis represents 'Scale'.

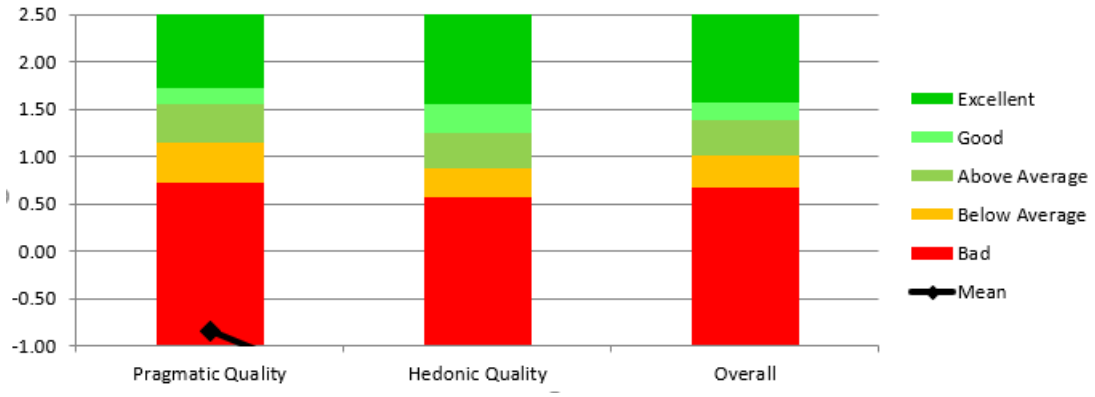


Figure 8.9: Comparison of benchmark with CSI where x axis represents the 'Quality' and the y axis represents 'Scale'.

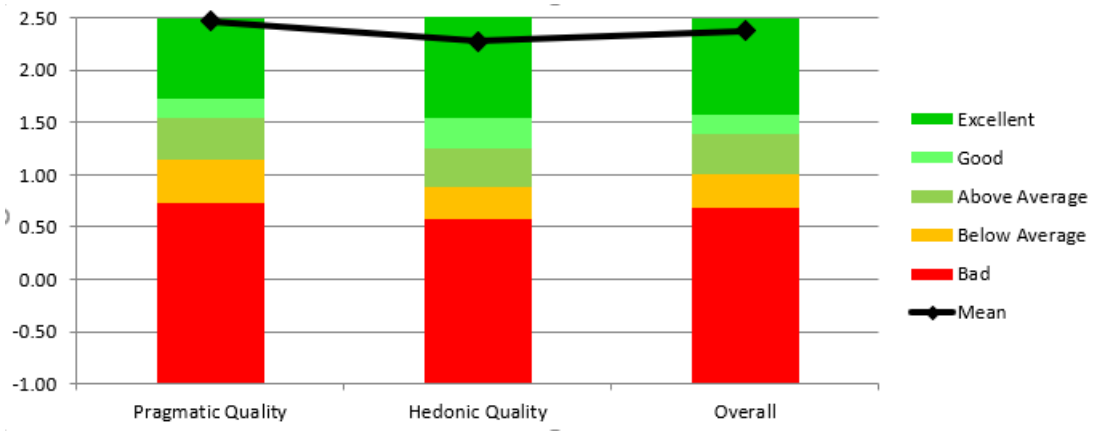


Figure 8.10: Comparison of benchmark with MCSI where the x axis represents 'Quality' and the y axis represents 'Scale'.

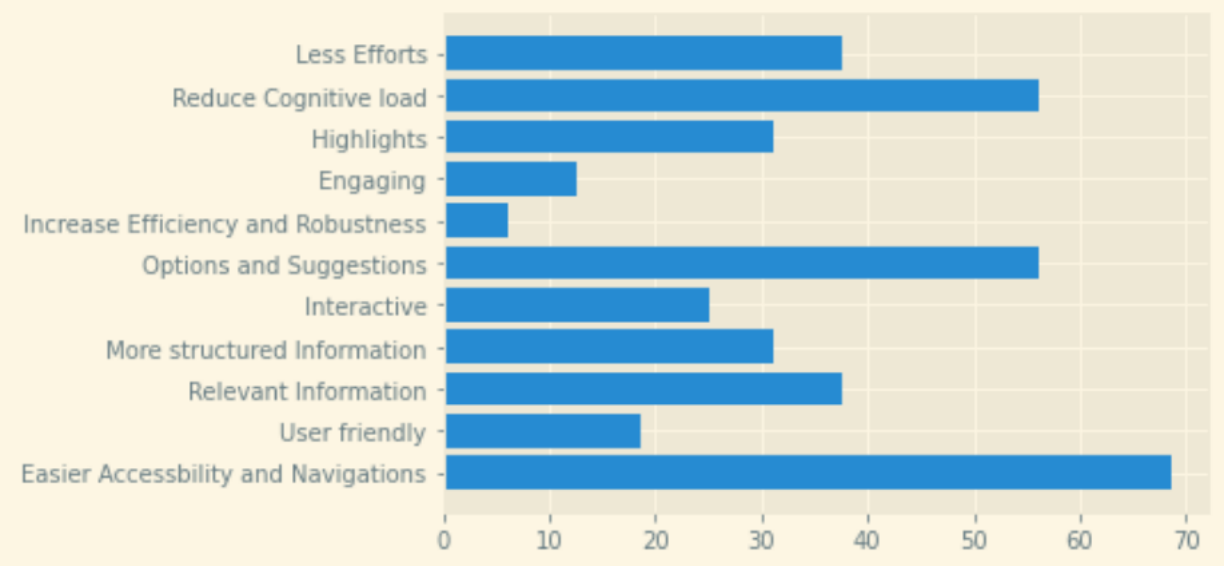


Figure 8.11: Users reason to prefer MCSI where x axis represents the 'Quality' and y axis represents the 'Scale'.

8.4.6.2 RQ2: How does user experience vary for both search settings in comparison to a standard benchmark?

1. **Comparison of the CSI with a standard benchmark**: Similar to Chapter 7, the comparison from the benchmark [150], the CSI UX is far below the mean of the interactive product (Pragmatic Quality < 0.4, Hedonic Quality < 0.37 and overall < 0.38) as shown in Figure 8.9. This signifies that the UX for the CSI needs major improvement on pragmatic and hedonic sectors. With the comparison of the benchmark, the CSI marked with the low quality of user experience lies in the range of worst 25% of products.

2. **Comparison of the MCSI with a standard benchmark**: Similar to Chapter 7, the MCSI UX is far above from the mean of the interactive product benchmark (Pragmatic Quality > 0.4, Hedonic Quality > 0.37 and overall > 0.38) as shown in Figure 8.10 which signifies the UX on MCSI compared with other interactive products (benchmark) is very high and its excellent level lies in the range of 10% best results.

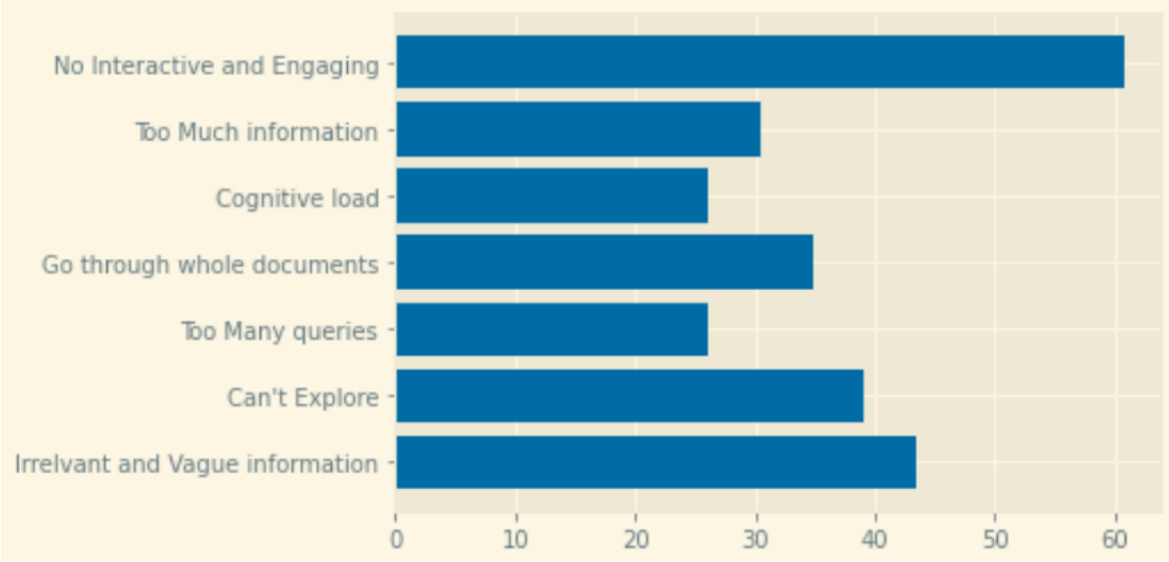


Figure 8.12: User challenges with the CSI where x axis represents 'Quality' and y axis represents 'Scale'.

8.4.6.3 RQ3: What are the factors that support search using the machine learning based MCSI?

Around 87% of the subjects claim in the post-search interview that the MCSI was better than the CSI. Similar to Chapter 7, users suggested reasons for preferring the MCSI as shown in Figure 8.11.

8.4.6.4 RQ4: What are the challenges with the conventional search system?

Similar to Chapter 7, subjects found some major challenges in completing the search tasks with the CSI. The limitations are mainly associated with the observations from

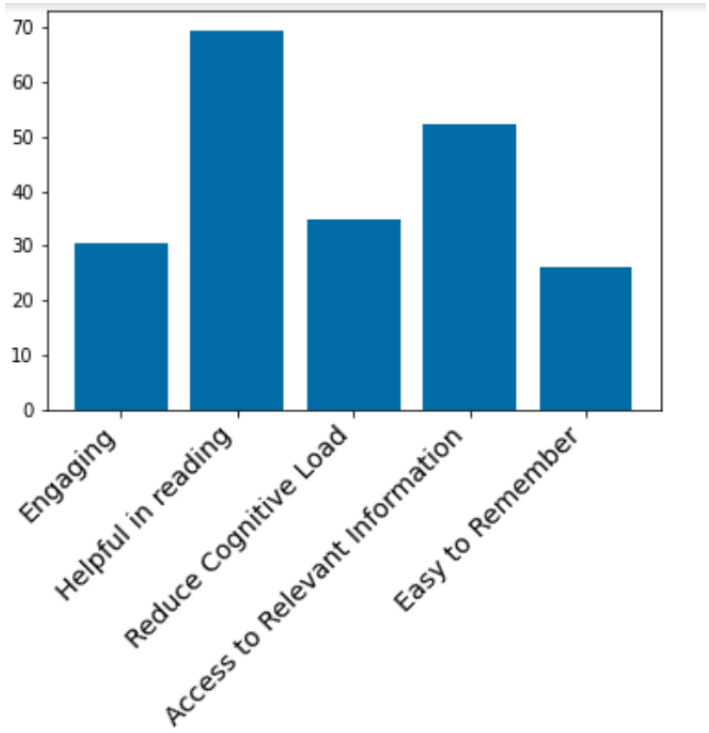


Figure 8.13: User reasons to prefer highlighting where y axis represents 'Percentage' and x axis represents 'Reasons'.

user interactions and feedback after the search task. The limitations can be divided into five broad categories, as shown in Figure 8.12.

Exploration: Around 40% of the subjects claimed they found it difficult to explore the content. The other reason was that too much information was displayed to them on the page, which confused them during the process of information seeking (40%).

Cognitive Load: Around 30% of subjects experienced issues with cognitive load using the CSI. As discussed in Chapter 7, there are multiple limitations associated with this single query search approach which put high cognitive load on the user. Around 15% claimed to have cognitive load due to multiple query formulation.

Interaction and Engagement: Similar to Chapter 7, subjects (60%) found difficulty in engaging and interaction with long documents. As per UX hypothesis testing, users ranked CSI negative in hedonic property in comparison to the MCSI and with the standard benchmark.

Irrelevant and Vague Information: Similar to Chapter 7, 35% of the subjects

did not find the long documents precise enough to satisfy their information need when using the CSI. In contrast however, 90% of them were satisfied with the way information was presented to them in the MCSI, although the Wikipedia API and underlying retrieval method was the same for both interfaces.

8.4.6.5 RQ5: Does Highlighting important segments support users in effective and efficient search and why?

87% of subjects liked the document highlighting options in the MCSI. Similar to Chapter 7, highlighting was found to be helpful in reading, access to relevant information and reduce cognitive load, as shown in Figure 8.13.

8.4.6.6 RQ6: What are the challenges and opportunities to support exploratory search in conversational settings?

8% of subjects faced some challenges using the machine learning based MCSI, such subjects needed more sections and subsections to be identified in the documents to support their exploration, and an interactive learning tutorial with the interface, which is out scope of this investigation. Around 4% of the subjects felt the need for improvement in operational speed.

Negative	Positive	Scale	Rule based Mean	ML based Mean	P_Values
obstructive	supportive	P	5.60	6.56	0.00038
complicated	easy	P	5.76	6.68	0.0000383
inefficient	efficient	P	4.40	6.18	0.000002
confusing	clear	P	5.48	6.43	0.000119
boring	exciting	H	5.44	6.31	0.000184
not interesting	interesting	H	5.48	6.37	0.000163
conventional	inventive	H	6.28	6.31	0.0901255
usual	leading edge	H	5.50	6.12	0.001735

Table 8.12: UEQ-S score based on rule-based and machine learning where 'P' stands for Pragmatic Quality and 'H' stands for Hedonic Quality (statistically significant).

NASA Task Load	Rule Mean	ML Mean	P Values
How mentally demanding was the task?	3.68	2.93	0.176
How physically demanding was the task?	2.76	2.06	0.147
How hurried or rushed was the pace of the task?	2.76	2.00	0.037
How successful were you in accomplishing what you were asked to do?*	5.32	6.31	0.004
How hard did you have to work to accomplish your level of performance?*	3.96	2.18	0.00008

Table 8.13: Comparison of NASA task load score based on rule based MCSI and machine learning based MCSI where * indicates statistical significant.

8.4.6.7 RQ7: How does the rule-based MCSI differ from the machine learning based MCSI?

We next compare our findings from the rule-based MCSI study in the previous chapter and machine learning based MCSI study in this chapter. We compare the mean results (T independent) of rule-based MCSI and the machine learning based MCSI based on IECSI described in Chapter 4.

1. **User interactive experience:** As shown in the Table 8.12, users reported having a better interactive experience with the machine learning based MCSI, which arises from its extended capabilities, such as improved questions handling, image search, audio input and output. The difference between parameters of UEQ of rule-based and machine learning based the MCSI is statistically significant.

2. **Task Load:** As shown in the Table 8.13, out of the 5 parameters, 3 are statistically significant. Users reported that use of the revised interface was less hurried, easier to accomplish task and achieved greater task success, which are statistically significant. For the other two parameters, mentally demanding and physically demanding, the machine learning based MCSI, scored less well than the rule-based MCSI, but these two parameters are not statistically signif-

System Usability (IBM)	Rule Mean	ML Mean	P_Values
Overall, I am satisfied with how easy it was to use this system to	5.96	6.75	0.002
It was simple to use this system	5.92	6.81	0.003
I can effectively complete my work using this system	5.64	6.56	0.012
I am able to complete my work quickly using this system	5.76	6.50	0.02
I am able to efficiently complete my work using this system	5.76	6.37	0.07
I feel comfortable using this system	5.88	6.50	0.046
Whenever I make a mistake using the system, I recover easily and quickly	5.44	6.06	0.11
The information (such as online help, on-screen messages) provided with this system is clear	6.08	6.56	0.134
It is easy to find the information I needed	5.48	6.62	0.0008
The information is effective in helping me complete the tasks and scenarios	5.68	6.62	0.008
The organization of information on the system screens is clear	5.92	6.62	0.008
The interface of this system is pleasant	6.08	6.12	0.87
I like using the interface of this system	6.12	6.31	0.55
This system has all the functions and capabilities I expect it to have	5.72	6.25	0.09
Overall, I am satisfied with this system	5.92	6.37	0.112

Table 8.14: System usability score for rule based and machine learning based MCSIs.

Search Mode	**User Search Experience**	**Rule Mean**	**ML Mean**	**P Value**
Pre-search	Pre-existing knowledge about search topic	3.04	2.87	0.734
	Interest about this topic	5.56	6.25	0.088
	Anticipated difficulty to find the information needed	3.68	3.81	0.75
Post-search	Actual difficulty to find the information needed	3.28	1.68	0.000075
	Helpful in satisfying information need	5.52	6.40	0.017
	Quality of Text presented	5.64	6.12	0.14
	Usefulness of search results	5.12	6.31	0.005
	Usefulness of text shown	5.36	6.25	0.02
	Cognitively engaging	5.92	6.56	0.010
	Expansion of knowledge about the topic	6.00	6.25	0.47
	Understanding of the topic after search task.	5.88	5.93	0.83
	Grade the success of your search session	5.72	6.25	0.09
	Rating search setting in terms of understanding user inputs	5.40	6.12	0.09
	Rating assigned search setting in the search results presentation	5.76	6.25	0.081
	Rating the suggestion(s) skills	5.56	6.26	0.002

Table 8.15: Searching as learning parameter scores for rule-based and machine learning based MCSIs.

Comparison	Diff_Mean Rule	Diff_Mean ML	P value	Change Percentage (approx)
Diff-Pre-Post -Summ_DQual (0-3)	1.60	1.81	.457	12.5%
Diff-Pre-Post -Summ_DIntrp (0-2)	1.08	1.12	.853	3.7%
Diff-Pre-Post (0-1) -Summ_DCrit	0.60	0.62	.876	3.3%

Table 8.16: Knowledge expansion (Change in Knowledge) for the rule-based vs the machine learning based MCSI.

icantly different, which indicates that both the interfaces are equally mentally and physically demanding.

3. **Software Usability:** As shown in the Table 8.14, out of 15 parameters, 12 are statistically significant. In relation to these parameters, users reported better software usability with the machine learning based MCSI arising from its extended capabilities, including extended question handling, image search, audio input, and output. For the other 3 parameters of the machine learning based MCSI, it scored more than the rule-based MCSI, but these 3 parameters were not statistically different, as the users experienced similar usability experience with both interfaces.

4. **Search Experience via Learning:** In Chapter 7, 15 parameters were analysed based on the IECSI, this was further developed into 18 parameters to account for the extended capabilities added of the machine learning based MCSI. To distinguish the systems, we compare the common 15 parameters of the two systems. As shown in Table 8.15, of the 15 parameters, 14 are statistically insignificant. Based on these 14 parameters, users report better learning experience, knowledge expansion, understanding of the topic, relevancy of the text, help in satisfying the information need and cognitive engagement with the machine learning based MCSI. The difficulties encounted by users while completing their search task using the machine learning based MCSI were less than those with the rule-based system. The only significant parameter is the user's claimed pre-existing knowledge about the search topic before the task, which is expected since it is unrelated to the interface used. The overall conclusion of this analysis is that the user reported better search support for learning when using the machine learning based MCSI but the difference is statistically not significant which implies both the MCSI interfaces provide similar search experience.

5. **Knowledge expansion:** When comparing the knowledge gain on the three

factors from the IECSI using different interfaces, users-reported more knowledge gain when using the machine learning based MCSI in comparison to the rule-based MCSI. Although the change in knowledge varied from 3% to 12% in all three parameters, as shown in Table 8.16, the knowledge expansion is not statistically different.

Overall, based on the T test conducted on data collected in studies of Chapter 7 and Chapter 8, users found the machine learning based MCSI interface a better platform to support conversational search.

8.5 Discussion

As described earlier in this chapter and in Chapter 3, there is very little existing work related to the examination of CS interfaces [22]. The evaluation of such systems has either focused on search satisfaction [152, 143] or software usability [28]. Our study has evaluated a machine learning based MCSI using our IECSI framework to study the effect on the searcher of this alternative mode of search interaction. Our study indicates that subjects find our MCSI more helpful than a closely matched CSI. We also observed varied user information seeking behaviour while using MCSI in comparison to CSI [152, 137]. Some existing studies conducted to understand user behaviour in CS use Wizard-of-Oz type studies [173]; in contrast, we study using an automated search support agent. Using our agent-based system, we observe the natural expectations of user search in conversational settings. We observe that subjects do not encounter any difficulty in using the new interface because it seems to be similar to the standard search interface with additional conversation capabilities. We also observed the information space and its structure is a key component in information seeking. Subjects found highlighting important segments in long documents enables them to access information much easily. The MCSI made the search process less cognitively demanding and more cognitively engaging.

In this study, we compared our rule-based MCSI and machine learning based

MCSI using IECSI dimensions, and found that the machine learning based MCSI provided a significantly better experience to the user in four dimensions out five dimensions of the IECSI. Although, the dimension of knowledge expansion is not statistically different, but knowledge varies from 3% to 12% while using machine learning based MCSI.

8.6 Conclusions

Based on user feedback from the rule-based MCSI, we have introduced a machine learning based MCSI with extended search functionality and evaluated it based on our CS interface implicit evaluation framework (IECSI). Our study indicates that the MCSI system is less cognitively demanding and more cognitively engaging than the CSI and rule-based MCSI systems. In the next chapter, we discuss the detailed RQs and their corresponding answers for this PhD project.

Chapter 9

Conclusions and Future work

This chapter gives a summary of our work reported in this thesis, and revisits the research questions introduced in Chapter 1. Our analysis describes the conclusions that can be made and limitations of our investigation, and suggests directions for future work.

9.1 Key Findings

In this PhD dissertation, we describe our work on the introduction of conversational processes into IR services. We examine the challenges and opportunities for conversational search, user search behaviour in traditional IR and conversational settings and our exploration of an implicit evaluation framework for CS interfaces. The contributions of this are reviewed below:

9.1.1 Exploring the Challenges for Search in Current Conversational Systems

We investigated the challenges of CS on a smart AI speaker conversational system. This investigation highlights the major problems faced by users while attempting to undertake search tasks on these systems. The results of this study were used to purpose and evaluate an extended application to support exploratory search using

these platforms. We have also introduced a new evaluation framework covering different aspects such as software usability, search experience, user usability, cognitive load and knowledge expansion during search process. Statistical test are used to validate hypotheses and compare alternative search interfaces.

9.1.2 Multi-view Conversational Search Interface (MCSI)

We introduced and explored a Multiview Interface for Conversational Search (MCSI). The MCSI encapsulates a conventional search system with an assistant agent interacts with the user via a chat type dialogue strategy. Dialogue Strategies for the agent which were built based on two major techniques: rule based and machine learning. These dialogue strategies were each embedded in the MCSIs. To the best of our knowledge, this is the first CS interface which examines a non-human agent to support exploratory search. This agent is built on the basis of the results of our investigation of user search behaviour carried as a part of this PhD research.

The MCSI and a corresponding conventional search interface (CSI) were compared and evaluated for complex back story based information seeking activities. Results measuring the cognitive load reduction, knowledge expansion gain and change user experience indicate that the MCSI is more effective and efficient for the user in our exploratory search tasks than the CSI.

9.2 Research Questions Addressed in this Thesis

We investigated the research questions introduced in Chapter 1. The research questions examined user search behaviour, user interaction behaviour with conversational agents, challenges and possible areas to support users in CS, investigating and developing dialogue strategies and the evaluation of CS. In this section we revisit these questions and summarize our findings from this thesis.

9.2.1 RQ1: Based on analysis of user search behaviour with an existing search systems, what are the challenges and opportunities for conversational search paradigms in exploratory search settings?

RQ 1 focuses on current state-of-the-art search and conversational systems. Our first investigation studied user search behaviour in a current web search engines.

The investigation of this research question is explained in Chapter 5. We explored a conventional search system, analysed user search behaviour, and found potential opportunities for inclusion of conversational support in the search process.

1. **What are the current search behaviours of users?**

   We conducted a user study with a standard online web search engine. We studied the behaviour of non-expert users of web search engines.

   We identified five types of search behaviour using this study:

   (a) Type A: The user enters one query and selects one document from the retrieved list. In our study, users showing this behaviour spent on average of 22 minutes studying this single document to address an assigned information need. This document was able to entirely satisfy the information need. There was no opportunity to study learning behaviour from the examination of multiple documents.

   (b) Type B: Users enter one query and open multiple documents as they seek to fulfill their information need. The number of documents opened was on average 2.5. The average time spent viewing each document was approximately 9 minutes.

   (c) Type C: The user performs more than one iteration of query and inspection of results in order to fulfill their information need. The average time spent on each search task was 34.6 minutes, with an average of 2.6 queries and 4.4 documents opened per search task. The average time spent on

viewing each document was approximately 7.8 minutes.

(d) Type D: Users perform more than one search interaction, but only opened a single document. We did not observe this behaviour in our study,

(e) Type E: Users perform more than one search, but do not open any document to fulfill their information need. In this special case, users may fulfill their information need by reading the titles and snippets of retrieved documents. We did not observe this behaviour in our study.

2. **What are the possibilities to support current user search behaviours in a conversational system?**

Analysis of these results led us to suggest the following opportunities for the introduction of conversation into the search process.

(a) Type A: Conversational search could support user exploration of individual retrieved documents to reduce the time and complexity of accessing potential relevant information.

(b) Type B: A conversational agent might assist users in identifying useful content found across multiple documents in an integrated way to improve efficiency of satisfying their information need.

(c) Type C: A conversational agent could assist a user in reformulating queries and displaying potentially relevant retrieved content.

9.2.2 RQ2: Based on analysis of user search behaviour with conversational applications, what are the challenges and opportunities for conversational search paradigms in exploratory search settings?

RQ 2 focuses on current state-of-the-art conversational systems in exploratory search settings. Our first investigation studied user search behaviour using a smart Speaker system. The investigation of this research question is described

in Chapter 6. We explored a state-of-the-art conversational system, analysed user search behaviour, and found potential opportunities to explore CS. There are a number of commercial conversational agents currently available including Amazon Alexa, Google Assistant, Apple Siri and Microsoft Cortana. After analyzing the parameters (accuracy, awareness, multitasking and customization) based on past investigations of conversational agents [191, 151, 153]. We selected the Alexa Echo Show for our investigation. The details of our methodology for this investigation were described in Chapter 5.

(a) **What are the opportunities and challenges for incorporating interactive search capabilities into a current conversational system?**

Challenges:

Attempting to use the Default Alexa assistant to address exploratory information needs expressed in a set of scenario backstories led to considerable user frustration with poor success in addressing the information need using the standard Alexa application.

As described in Chapter 6, the limitations can be classified into four broad categories:

i. Poor knowledge representation: From the responses to queries given by Alexa, it became clear that Alexa only represents either fact based answers or simply starts reading from the beginning of a long Wikipedia document. This was noted by around 18% of participants in our study who claimed that Alexa had poor knowledge representation.

ii. Poor speech recognition and high error rate: While not directly related to its search capabilities, around 52% of our participants noted that they experienced frustration arising from poor speech recognition and high word error rates while interacting with Alexa.

iii. Difficulty in asking questions: This was the most important reason identified by participants, 75% of whom indicated that they had difficulty in creating queries. Participants were unable to search effectively since they had no background knowledge about the subject.

iv. Others: A number of other factors were identified, the key ones were "Interruptions" (2%) and "Cognitive Load" (4%). Participants observed that they were unable to complete queries due to interruptions by Alexa which commenced the search process before query entry was complete.

Opportunities:

Our investigations using the standard Alexa application highlighted some critical areas which provide opportunities to improve exploratory search using Alexa.

i. Background knowledge support and effective knowledge representations: The standard Alexa application provides fact based answers, but does not support the user to learn about the topic and refine the search. After conducting our study, we proposed that the user should be provided with relevant information related to their search query based on facts which could help them to explore and analyse the available information space, to improve their ability to form effective follow on queries.

ii. Priming, dialogue-driven approach and interactive search process: For the standard Alexa application, we observed that the application was not able to actively engage the user in the search process. Moreover, the efforts made by the searcher to identify and explore documents to satisfy their information need led to considerable frustration. To reduce this frustration and enhance search effectiveness, we proposed to introduce a dialogue driven approach into search process.

(b) **What are the characteristics which prevent existing conversational systems from becoming effective tools for complex information seeking tasks?**

- **Fact-based Search**

  Current conversational assistants such as Alexa are designed to answer of fact-based questions, like a traditional question answering system, in contrast to traditional IR systems, which enable users to perform exploratory search operations.

- **Poor information seeking strategy**

  Our investigation found that Alexa did not support exploratory information seeking. This was reported by more than 45% of participants, reported in Chapter 6. As shown in Figure 6.6, the three major reasons for poor information seeking in standard Alexa are reported as being lack of background knowledge measuring that user are unable to create the right query, poor recognition of the user's query by Alexa, and poor representation of knowledge by Alexa.

(c) **What are the main user expectations of conversational search systems?**

From our study we found five major expectations of our participants for search in conversational systems.

- **Exploratory**

  A CS system should provide a broad information space to the user to provide the opportunity for them to explore a space of relevant information and to narrow the exploration to focus on addressing their information need.

- **Content selection**

  In general, the standard Alexa system presents very limited information to the searcher in response to a query. The information returned by Alexa was often not suitable, and in many cases the searcher had

to issue many queries in order to even begin to address an exploratory information need.

Our investigation showed that the important variables with respect to the user experience while attempting to satisfy an information need are as follows: the average number of interactions, the number of successful interactions, the number of unsuccessful interactions, the average time to complete a search process and the quality of the presented text. Based on our results, we can conclude that the average interaction failure rate was around 62%, which is very high. For the total of 438 interactions, the average total time of interactions by a user was found to be 5.6 minutes for the default settings.

We can see from these figures that this engagement is very inefficient, as a result of this the user struggles to learn since most of the interactions result in failure, leading to demotivation and frustration of the user.

- **Content interactions**

  Searcher interactions include use of multiple Alexa skills including navigation skills, presentation skills and speech skills of conversational agents.

  i. Navigation skills: A conversational search agent should support the user in navigation through the information space or the documents.

  ii. Presentation skills: More than 85% of users considered that all three dimensions (text, speech and images) are required enable the best mode of information presentation to be available.

  iii. Speech skills: The speech skills can be classified based on multiple parameters, including speech speed, speech recognition, interruptions, speech content and speech length.

- **Information representations**

In our study, 18% of the searchers reported that the information represented during the search process in Alexa default setting was poor. The searchers were not satisfied with the combination of text and images. They also found difficulty in maintaining the contextual information flow while undertaking a multi-step search task.

- **Conversational properties**

  Our study indicates properties that every CS system should have.

  i. On boarding: This is the initial interaction in which the user is introduced to the system, in which it explains its competencies.

  ii. System as teacher: The user expects a system should ease their interactions by revealing its capabilities and essentially teach the user how to use the system. In our study we observed that users who have previous experience with conversational application interacted with it for longer (7.2 minutes), than users who were using it for the first or second time (5.8 minutes).

  iii. Incite: We observed that most of the interactions were one way with the system unable to engage in useful dialogue with the user. However, ideally a conversational application should engage the user in back and forth dialogue to assist them in reaching their goal.

  iv. Diverge flow and course corrections: A conversational application should be robust. As such, it should be able to handle any unexpected entries from the user, and use this to guide the user towards their goal.

9.2.3 RQ3: How might dialogues be embedded in search settings for improved satisfaction of user information needs?

This RQ is divided into the following sub questions.

9.2.3.1 Dialogue Strategy to Support Content Engagement

Based on our investigation in Chapter 6, we focused on designing and implementing a new dialogue strategy for the Alexa Assistant on the Amazon Echo Show with respect to the default search setting, which we refer to as *Alexa custom search*. The new dialogue strategy was designed to enable users to search and explore through long documents, and to facilitate two way interaction between Alexa and the user via a dialogue. The details of the setup and methodology of Alexa custom search are explained in Chapter 6.

1. **Can we develop dialogue strategies which engage users in information seeking activities?**

   The investigations reported in Chapters 5, 7, 8 shows that a dialogue strategy can engage the user in information seeking. These studies, investigated this question by looking at three different dialogue strategies using three different platforms: Alexa on the Amazon Echo show, a MCSI with rule-based and a MCSI with machine learning-based. In all these investigations, users claimed that the dialogue strategy based system allowed them to access information effectively. On average users were satisfied with the way options and suggestions were provided. The users found the dialogue-based search agent effective, interactive, engaging, and user friendly. They were also satisfied with the suggestions and options provided by the search agents embedded within the dialogue strategy. They also reported satisfaction with the speed of the search agent and found the dialogue-based agent satisfactory.

   The top five reasons to prefer dialogue-based agents were: navigation and

directed search, relevant and more informative, options and suggestions, increased efficiency and reduced cognitive load, and better and more effective systems. Overall, the users found the agent informative and well-directed, and provided options for exploration throughout the search process.

Based on the hypotheses examined in Chapters 6, 7 and 8 relating to cognitive load, we found that users experienced less cognitive load while using the dialogue supported search system. Similarly, they gained more knowledge while using the dialogue-based system. Additionally, they claimed to have a better interactive user experience. Hence we can conclude that dialogue strategies can be developed to support effective information seeking.

2. **What factors affect dialogue strategies for user information seeking?**

   Multiple factors affect dialogue strategies for user information seeking, these were investigated in Chapters 6, 7 and 8. The factors can be assigned to the following categories:

   - Structure of representation information: This is an important criteria on which can be broken down into two sub parts: content of the document and representation of the documents. Based on studies conducted in this PhD, approximately 85% of users were satisfied with the content shown via the dialogue-based agent. Moreover, around 80% of the users were satisfied with text representation with the dialogue-based agent.

   - Exploration: It was found that dialogue strategies allow the user to explore the documents. Exploration supports the user in the information seeking process and allows them to learn via the search process.

     A poor dialogue strategy in information seeking can require the user to expend much effort to go through whole documents, discouraging them from exploring further to satisfy their information needs, similar to the situation observed for conventional search systems.

- Interaction and engagement: Interaction and engagement with the content is a key component of information seeking. As discussed in Chapters 6, 7, 8, subjects find difficulty engaging and interacting with long documents in conventional search systems.

  As per the study reported in this PhD, dialogue strategies can support the user's interaction and engagement with documents, which enable them to satisfy their information need effectively. Based on our hypothesis testing on interactive user experience, users have a significantly better interactive experience than the standard benchmark and conventional search system. Similarly, the hypothesis on searching as learning proved that users have more content engagement with the dialogue-based systems.

3. **How is the user's search gain in topical knowledge affected by the use of dialogue strategies in the search process?**

   As per the studies described in Chapters 6, 7, 8, we observe that users find dialogue-based systems efficient and effective in seeking relevant information. To verify the users' claims, we further investigated the CSI and MCSI systems to examine knowledge gain and search as learning. This hypothesis testing concluded that dialogue-based systems provide better support to users in expanding their knowledge.

   In the dialogue-based setting based on knowledge gain, the subjects scored higher in change of critique, quality, and interpretation factors. All three parameters increased by more than 15% while using a dialogue-based system. In dialogue-based search, the users reported a better learning experience in all three studies.

9.2.4 RQ4:Can enhanced content engagement support improve satisfaction of user information needs?

This RQ is divided into the following sub questions.

1. **Can highlighting potentially relevant important sentences within long documents help in the search process by reducing the work of the user in examining documents to satisfy their information needs?** In the MCSI systems explored in Chapter 7 and Chapter 8, the "full document" option opens a full document highlighting important segments. Around 90% of the users liked the document highlighting option in MCSI.

   There is potential for this highlighting facility to be further extended in the MCSI, where for example the agent could develop an incremental model of the user's developing search engagement. The following reasons were identified for liking the highlighting option:

   - Interactive and engaging: Around 30% of users claimed that they could better engage and interact with documents by using the highlighting options.
   - Helpful: Around 70% of the subjects found highlighted documents helpful in information seeking.
   - Reduce the cognitive load: Around 35% of the subjects believed that the highlighting in documents reduced their cognitive load by highlighting important sentences, which helped them satisfy their information needs.
   - Access to relevant information: Around 55% of the subjects believed that highlighted documents helped them easily access useful information, saving them time and effort.

9.2.5 RQ5: Can we demonstrate improved search effectiveness when using a conversational search tool?

This RQ is divided into the sub question.

9.2.5.1 How does a conversational search interface compare to a comparable conventional search system and why?

Based on the evaluation framework discussed in Chapter 6, we compared the MCSI with a CSI. As described in Chapter 7 and Chapter 8, the MCSIs were compared to a conventional system in five dimensions, which address the user's expectations of the search system.

Concerning all aspects of cognitive load, subjects experienced less task load when using the conversational interface. They were also more success in accomplishing the task using the conversational interface, and the results for accomplishing their search task were found not to be statistically significantly different. Subjects felt less insecure, discouraged, irritated, stressed, and annoyed while using the conversational interface, again with not a significant difference between the interfaces. For user interactive experience aspects, the subjects' experience was positive in pragmatic quality and hedonic quality using the MCSI, and statistically significantly different from that with the CSI. Subjects felt obstructed, complicated, confused, inefficient, and bored while using the CSI again with significant difference from the MCSIs.

For software usability, PSSUQ was evaluated using four dimensions: overall satisfaction score (OVERALL), system usefulness (SYSUSE), information quality (INFOQUAL) and interface quality (INTERQUAL), which include fifteen parameters. The MSCI outperformed the CSI, and the users had a better software usability sense in all dimensions. Based on the search as learning Vakkeri model [121, 152], various parameters including the relevance of the search result, the quality of the text presented by the interface, and understanding the topic in both the search settings via pre-search and post-search questionnaires were compared. The users found that the MCSI provided better learning and searching experience, and improved knowledge expansion. These factors were evaluated using three parameters: Quality of Facts (DQual), In-

terpretation (DIntrp) and Critiques (DCrit).

In all three parameters, the MCSI scored higher than the CSI, which led to conclude that users better understand and learn while using the MCSI. We further examined the MCSI with respect to two search agent approaches: a) rule-based b) machine learning based. Based on the MCSI results obtained from Chapter 7 and Chapter 8, we conducted an independent T test to compare the means of rule-based and machine learning based MCSIs, and found the machine learning based MCSI with extended capabilities, is better than the rule based approach in IECSI and statistically significant. The current results indicate the MCSI system is less cognitively demanding and more cognitively engaging than the CSI systems.

9.2.6 Limitations and Opportunities

In this section, we illustrate the limitations of the work described in this thesis, and examine opportunities for future work:

- **Subjects Population**

  As outlined in Chapter 2, user-based evaluation is complicated and complex. We carried out multiple preparatory pilot studies before the main user studies with a substantial number of participants. We decided to control sequencing effects by rotating each group's topics and using the Latin square method to arrange tasks. However, while this study produced interesting findings; the sample size varied between 16 - 32 depending upon the subjects' availability. Furthermore, it can be influenced by the characteristics of the sample population, such as education, age group, gender, etc. Following analysis of the results, we recommend that the investigation should be repeated with a larger number of participants with various demographics, with more complex backstories, to see if the trends reported in this work are consistent across more participants and have a wider range of backstories.

- **Retrieval and Ranking**

  In this PhD research, our focus was on the exploration of the potential of CS and to understand users' expectations of it. This exploratory research into CS allowed us to understand the user perspective and their requirements, but we have not explored the effects of the ranking and retrieval models on user reaction to CS. Future work should include the connection of the dialogue-based information seeking models with study of retrieval and ranking models.

- **Self-learning Model**

  In this PhD, we investigated the MCSI based on rule-based and machine learning agents for a dialogue-based system, but these systems were based on offline learning and could not learn from the user's interaction during a search session. Future work should focus on building of an automatic adaption of the dialogue-model based on user interactions, e.g. using reinforcement learning.

- **Highlighting model**

  In the MCSI system, the full document option opens the full document with important segments highlighted. These highlightes can direct the user more easily to access relevant information. To extract important segments from the full documents, we developed a custom algorithm using an unsupervised approach. The objective in this PhD research was to explore the potential of a highlighting option in full documents and how it can help users in information seeking. Our method could be further developed, evaluated and compared for standard summarization benchmarks to potentially to improve the method and to provide better support to the user in information seeking.

9.2.7 Future Directions

There are a number of potential directions for further work arising from this thesis. Suggested possibilities include modelling conversational information seeking in a model to reduce cognitive load and increase cognitive engagement using reinforce-

ment learning. This could investigate the potential to understand user engagement in the search process and design a framework for conversational information seeking models, which can be autonomic in nature by using reinforcement learning. This information seeking framework could be evaluated based on the evaluation framework introduced in Chapter 4. The scope of the search could be increased by connecting the CS interfaces to the API of a web search service which will enable examination of more diverse types of information needs, and will require more complex summarization and highlighting of retrieved content.

The functionality of the agent could be extended to proactively seek to support and anticipate user information needs. For example, by extracting queries from content attended to by the searcher, and locating related content which may be of interest to the searcher.

Moreover, we could further investigate MCSI by evaluating of different complexity of the tasks targeting a specific age group of participants. Furthermore, we could also extend MCSI by embedding with the capability of multimedia search.

The prototype MCSI explored in this thesis introduced limited image search capabilities based on image captions. This could be extended to exploration of content-based search of images and video archives, and audio collections such as podcasts and audio focused content such as massive open online courses (MOOCs).

Search for video and speech archives would require engagement with the agent to be extended to include facilities for interaction with temporal media.

Appendix A

A.1 Retrieval Models and Evaluations

This section contains an overview of the various retrieval model used in IR and the standard evaluation techniques adopted by the IR and IIR communities. We have adopted few techniques to draft our CS framework discussed in Chapter 4.

A.1.1 Retrieval Models

We now briefly introduce the details these retrieval models.

1. Boolean model: The Boolean model was the first model of information retrieval. This model is based on exact matching of the keyword in the document and can use the boolean logic in queries, i.e, documents are either retrieved or not, and the retrieved documents are not ranked. An advantage of this model is that it gives the searcher a sense of control over the system. It enables user to understand why a document has been retrieved given a query. The main disadvantage of this model is that it does not provide a ranking of retrieved documents [73, 83]. If the resulting document set is either too small or too big, it is directly clear which operators will produce respectively a bigger or smaller set. For untrained users, the model has a number of clear disadvantages.

2. Vector space model: The vector space model follows a statistical method based on the bag of words where the cosine similarity is calculated between the query vector and documents vector for searching the information. The cosine sim-

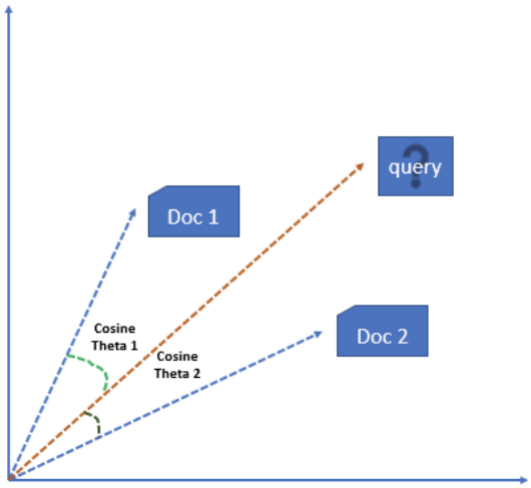


Figure A.1: Vector space model [179].

ilarity is the angle between query vector and document vector as shown in Figure A.1. Luhn [2] was the first to suggest the use of statistical methods for searching. The main idea behind the concept is to find out the statistical similarity between indexed documents and query asked by the user. To calculate the similarity between the indexed documents and query, the indexed document and query need to be converted into the vectors of equal length. The similarity score between the query and the indexed documents would rank the retrieved documents. The greater the score, the better would be the rank.

3. Probabilistic approaches: The goal of the probabilistic model of IR is to rank the retrieved documents based on their predicted probability of being relevant to the user's search query [5, 83].

4. Language models: The idea of language models came from language modeling in automatic speech recognition. Ponte and Croft [39] were the first from IR the community propose the use of language modeling in IR. For IR, language models are built for each document based on the query likelihood model. A separate language model is associated with each document in an index. The documents are ranked based on the probability of the search query in the document's language model [73, 83].

5. Neural models: Neural ranking models for IR employ shallow or deep neural

networks to populate a SERP in response to a query. Neural models bridge the gap between query and document representations by learning language representation (into vectors) from raw text. These AI-based methods are data-hungry, needing significant amounts of training data before being implemented. For neural-based approaches have been divided into two broad categories, i.e, shallow neural, (ii) deep neural. Shallow neural techniques use pre-trained neural embeddings (vectors) instead of learning the IR problem from start to finish. While deep neural techniques are concentrated on contemporary models, which have deep architectures, i.e. the deep neural network (DNN) contains multiple neural network layers [127].

A.1.2 IR Evaluation and methods

In this section, we briefly review the existing IR evaluation and methods. IR systems have traditionally been evaluated in terms of *precision* (the proportion of retrieved items which are relevant) and *recall* (the proportion of the available relevant documents that have been retrieved). In laboratory-based IR research, these metrics are used with predefined text collections of documents, representative search queries and corresponding relevance data, to examine the retrieval effectiveness of different IR methods.

IR metrics are also evaluated in interactive settings. These metrics focus on recording the characteristics of search in terms of demographic information and search experience. This is then combined with details of the interactions between the search system and the searcher. This includes features of the user's interactive search behaviour, such as the length of each query, the number of queries searched, and the number of the documents read [75]. Interactive evaluation can also record and examine features such as the user's subjective opinion of their search experience. This type of information is also referred to as the "usability" of the search system.

Conversational systems, in general, are typically evaluated using multiple dimensions such as quality of user interactions (depth of communication funnel which

means how deep the participants are going into the conversation and the number of dialogue turns), response speed, functionality and performance-based measures (effectiveness and efficiency of an agent to complete the task) [139]. Performance-based evaluation of conversational systems is still at an early stage and investigation of user engagement with conversational systems is an important topic for further analysis. The absence of existing work on evaluation means that this PhD work will need to explore methods for the evaluation of the use of conversational engagement in search.

We divide evaluation into three areas: evaluation of algorithm IR, IIR and conversational systems. Evaluation of conversational systems is discussed in Chapter 3, in this section, we briefly review existing methods for algorithmic IR and IIR evaluation.

A.1.3 Evaluation of Current Search System

Evaluation is a way to measure the quality of the search results which enable the user to satisfy their information need. The evaluation criteria in current IR can be divided into two major dimensions: online metrics and offline metrics [75].

- Online metrics: These metrics are based on user interactions with a search system. The following are some of the most widely used online evaluation metrics:

    1. Session abandonment rate: The proportion of search sessions which are abandoned without the user information need being satisfied.
    2. Click-through rate: The proportion of users who click on a particular document link compared to the total number of the users who see the link.
    3. Session success rate: The proportion of user sessions which lead to success in satisfying the user information need. This aspect can be ambiguous, as success depends upon the nature of the task.

4. Zero result rate (ZRR): This is the proportion of search engine results pages (SERP) which return zero relevant results. The metric either indicates a recall issue or that the information being searched for is not in the index. For instance, out of a total 10 search attempts, if 7 search results give zero results, then the ratio will be 7/10. A higher ratio indicates the documents are not indexed or that there are recall issues.

- Offline metrics: These metrics are responsible for evaluating the relevance of the result produced by the search engine appearing in the SERP as a whole. The major online metric components are: Recall, Precision, F1 score, Fall-out, Average precision [83]. The standard definitions of these metrics are as follow. The definitions of abbreviations are given in Table A.1.

Terms	Full Forms
RR	Relevant retrieved documents
TR	Total number of relevant document
TRR	Total number of relevant retrieved documents
NRR	Non-Relevant retrieved documents
TNR	Total number of non-relevant documents

Table A.1: Formula abbreviation.

  - Recall: The fraction of known relevant documents that are retrieved [83].

$$Recall(R) = RR/TR \tag{A.1}$$

  - Precision: The fraction of retrieved documents that are relevant [83].

$$Precision(P) = RR/TRR \tag{A.2}$$

  - F1 Score: The harmonic mean of precision and recall. The standard definition equally weights precision and recall but this can be varied [83].

$$F1 = (2 * P * R)/P + R \tag{A.3}$$

– Fall-out: The ratio of retrieved non-relevant documents to all non-relevant documents available in the collection index.

$$AvgP = \int_0^1 P(r)dr \tag{A.4}$$

– Average precision: The computation of precision at every position in the ranked list of documents. The average precision computes the average value of P (r) over the interval from r = 0 to r = 1 [57].

A.1.4 Evaluation Methods in Interactive Information Retrieval

Evaluation methods for IIR can be classified into four major classes: contextual, interaction, performance and usability [75].

A.1.4.1 Contextual

This focuses on the context in which search and interaction activity occurs. These measures are used to characterize the subject and their information need. Characterization of subjects includes age, sex, search experience, etc. Characterization of information needs focuses on information seeking situations such as subject background knowledge, subject familiarity with the search topic. These measures basically describe the context in which the information search occurs [54, 60, 75]. Ingwersen *et al.* [60] performed a comprehensive study of context in information seeking and retrieval. Dourish [54] addresses the theoretical notion of evaluating, in which context and action are intertwined and form "embodied interaction." Contextual measures often are used in association with other types of measures discussed below.

Measure	Description
Search outcome	-Number of relevant document saved -Precision and Recall -Average Mean Precision (AMP) -Discount Cumulative Gain, etc.
Time Based	-Completion time of search -User document reading time, etc. -capabilities and demerit to the user to enable in understanding the potential usage the system.
Informativeness	-Rank the document based on the its relevant content
Cost Based Measure	-Expenditure and utility in software development -Expenditure in maintaining search system, etc.

Table A.2: Performance measure subcategories.

A.1.4.2 Interaction

Interaction focuses on characterization of the interactions between the search system and the subject. It also includes user interactive search behaviour, such as the length of each query, the number of queries searched, and the number of the document reads [75]. This is one of the most important components with respect to conversational information retrieval

A.1.4.3 Performance

Performance focuses on measuring the results obtained from user interaction with systems, such as presentation of relevant documents, precision, mean average precision, and recall.It also includes time-based measures and informativeness measure. The criteria to select the performance measures for a particular study depend on the requirements of the IIR task.

Examples of performance categories are shown in Table A.2 [4, 75].

Measure	Description
Effectiveness	Accuracy and completeness with which subject achieve the specified task.
Efficiency	Subjects to fulfill their tasks with minimum expense and effort
Satisfaction	freedom from discomfort, and positive attitudes of the user to the product

Table A.3: Key dimension of usability.

A.1.4.4 User-feedback based Evaluation

This measure investigates subjects feelings and experiences arising from their interactions with a search system. These measures are also referred to as of relating to "usability" and can be divided into multiple dimensions as shown in Table A.3 [38, 130, 139].

1. Effectiveness : The *"accuracy and completeness with which the subject achieves a specified task"* [75]. In general, a system is effective if it helps the subject to complete their search tasks.

2. Efficiency: The resources expended to complete a search task using the search system. In general, a system is efficient if it enables subjects to fulfill their tasks with minimum expense and effort.

3. Satisfaction: The *"freedom from discomfort, and positive attitudes of the user to the product"* [38]. Satisfaction is the feeling of contentment that the subject experiences after finishing the search task.

The choice of any measure and its interpretation depend on the nature of the task and the user's expectations of the search system. For instance, the user needs to explore the topic, then the major focus will be on the number of interactions. In contrast, if the user is looking the specific answers, then finding one or a small number of highly relevant documents will be key, and system effectiveness will be measured based on precision of the results to answer the specific question.

Appendix B

The appendix includes the questionnaires, interview questions and backstories used in the studies conducted in the thesis.

B.1 Questionnaires

The following section includes the questionnaires of Chapters 5, 6, 7 and 8.

B.1.1 Questionnaire Based on Chapter 5

This questionnaire is from Chapter 5, to investigate the user search behaviour in standard search engine.

Section		**Questions**
Basic Information Survey	Q	User Id
	Q	Search Task id
	Q	Occupation
	Q	Age
	Q	Gender (M/F)
	Q	For how many years have you been using Web search engines?
	Q	On average, how many Web searches do you make each week?
Pre-Search Questions	Q	How much do you know about the subject of this search task? (low (1) - high (5))

Pre-Search Questions	Q	How interested are you to learn more about this topic? (low (1) - high (5))
	Q	What do you understand about the information needed from the search task description? (low (1) - high (5))
	Q	How difficult do you think it will be for you to find the information needed to address this task using a Web search engine? (low (1) - high (5))
	Q	Please write 3-4 sentences outlining what you know about the topic.
In Process Search Questions		**Query to the search engine after understanding the user stories (With each interaction)**
	Q	What search query did you enter?
	Q	Why did you choose these words?
	Q	Which results look useful to you? (please enter the document names and links)
	Q	How did you decide which results are useful?
	Q	If you opened any documents, Why and what did you learn from them?
	Q	Did this query enable you to satisfy the search task? If not, what will be your strategy to refine the query for the next iteration?
Post Search Questions	Q	While carrying out the search. Did you recall anything about the subject of you search from your previous experience of this subject? (Yes - No)
	Q	Which you didn't remember when you started the search?
	Q	If yes, then please give brief details of what triggered you to remember these details, and say briefly what you recalled.
	Q	Did you learn anything new why carrying out the search?
	Q	If yes, then please give brief details.

	Q	Did you find yourself to be cognitively engaged while carrying out the search task? (low (1) - high (5))
Post Search Questions	Q	Did you expand your knowledge about the topic covered in this search task? (low (1)- high (5))
	Q	I feel that I now have a better understanding of the topic of this task. (low (1)- high (10))
	Q	How would you grade the success of your search session for this topic? (low (1)- high (10))
	Q	Having completed the search task, please write a short 4-5 sentence summary describing your knowledge about the topic you were investigating which relates directly to what you learned by carrying out the search task

Table B.1: Questionnaire for Chapter 5

B.1.2 Questionnaire Alexa

This questionnaire is from Chapter 6 to investigate the user search behaviour in the standard conversational system (Alexa echo show).

Section		**Questions**
	Q	User Id
	Q	Search Task id
	Q	Occupation
	Q	Age
	Q	Gender (M/F)
	Q	For how many years have you been using Web search engines?
	Q	Which search setting of Alexa you are using? (Default - Custom Alexa)
	Q	Did you ever use any conversational agents (Siri, Alexa, google assistant and etc.) for any task such as Phone calls, setting calendar, listening to music etc. (Yes - No)?
Basic	Q	if Yes, How often you used it in a day? (Answers in digits) and Hows your experience with it?
Information	Q	Did you ever search with help of conversational systems? (Yes - No)
survey	Q	If Yes, How often you used it in a day? (Answers in digits) and Hows your experience with it?
	Q	How will you grade your English speaking fluency and pronunciation? (low (0) - high (5))
	Q	On average, how many Web searches do you make each week?
Pre-Search	Q	How much do you know about the subject of this search task? (low (1) - high (5))
Question	Q	How interested are you to learn more about this topic? (low (1) - high (5))

Pre-Search Questions	Q	What do you understand about the information needed from the search task description? (low (1) - high (5))
	Q	How difficult do you think it will be for you to find the information needed to address this task using a Web search engine? (low (1) - high (5))
	Q	How difficult do you think it will be for you to find the information to address this task using a Conversational agent? (low (1) - high (5))
	Q	How much you know about the Alexa? (low (1) - high (5))
	Q	Please write 2-3 lines what you know about Alexa
	Q	Please write 3-4 sentences outlining what you know about the topic.
Post Search Questions	Q	While carrying out the search. Did you recall anything about the subject of you search from your previous experience of this subject which you didn't remember when you started the search? (Yes - No)
	Q	If yes, then please give brief details of what triggered you to remember these details, and say briefly what you recalled.
	Q	Did you learn anything new why carrying out the search?
	Q	If yes, then please give brief details.
	Q	Does Alexa guide you to navigate through the document? (Yes - No)
	Q	If Yes, How will you grade the skipping skills of Alexa? (Yes - No)
	Q	Did Alexa's method of presentation of the answer (through speech) help you during the whole process and how?
	Q	How useful the answer (through speech or audio output) in the whole search task?
	Q	Is image shown in screen enhance your user experience during whole search process and how?

Post Search Questions	Q	How useful the images shown in the whole search task in satisfying the Information need? (low (1) - high (5))
	Q	How will you grade the navigation skills of Alexa? (low (1) - high (5))
	Q	How you rate the quality of text presented by Alexa with respect to your Information need and query?
	Q	Does Alexa enable you to read and skip the specific sections? (Yes - No)
	Q	Did you find yourself to be cognitively engaged while carrying out the search task? (low (1) - high (5))
	Q	Did you expand your knowledge about the topic covered in this search task? (low (1)- high (5)
	Q	I feel that I now have a better understanding of the topic of this task. (low (1) - high (5))
	Q	How would you grade the success of your search session for this topic? (low (1) - high (5))
	Q	How will you rate the Alexa skills in understanding your voice input ? (low (1) - high (5))
	Q	How will you rate the Alexa skills in the presentation of the search results? (low (1) - high (5))
	Q	Did Alexa suggest or recommend some options as a query while search? (Yes - No)
	Q	How many times Alexa stop in the middle of the search session (Approximately) and if you have to start your search process again?
	Q	What you think, would be the reason of Alexa stop working from the following? (Unable to understand your voice - Alexa is too fast to act - You are too fast - other - None)

	Q	How would you rate the suggestion skills of Alexa ? (very poor (1) - very good (5))
	Q	What would be the best way in which Alexa should present you the result? (Speech only - Speech and Text - Text only - Speech, Images and Text - Speech and Images)
	Q	How will you grade the Alexa for the current way of showing result with mix-multimodality in speech and text both? (very poor (1) - very good (5))
	Q	Which mode (text or speech or images) help you most to satisfy your information need and why?
	Q	Having completed the search task, please write a short 4-5 sentence summary describing your knowledge about the topic you were investigating which relates directly to what you learned by carrying out the search task

Table B.2: Questionnaire for Chapter 6 (Work with Alexa)

B.1.3 Questionnaire Multi-view Conversational Interface

This questionnaire is from Chapter 7 to investigate the user search behaviour in CSI and rule based MCSI interfaces for measuring search experience. This questionnaire is based on our study [152] and inspired by vakkeri [121] concept of searching as learning.

Section		**Questions**
Basic Information Survey	Q	User Id
	Q	Search Task id
	Q	Occupation
	Q	Age
	Q	Gender (M/F)
	Q	For how many years have you been using Web search engines?
	Q	Which search setting are you are using? (Default - MICS System)
	Q	On average, how many Web searches do you make each week?
	Q	Have you ever searched with the help of conversational systems? (Yes - No)
	Q	If Yes, How's your experience with it?
	Q	Do you use a conversational search tool regularly? if Yes, how many times per day do you generally use this system? (Answers in digits)
	Q	How will you grade your English speaking fluency and pronunciation? (low (1) - high (7))
	Q	How much do you know about the subject of this search task? (low (1) - high (7))
Pre Search Questions	Q	How interested are you to learn more about this topic? (low (1) - high (7))
	Q	What do you understand about the information needed from the search task description?

Pre Search	Q	How difficult do you judge it was for you to find the information needed to address this task using your assigned search setting (Default Setting or Chat Search interface Setting)? (low (1) - high (7))
Questions	Q	Please write 3-4 sentences outlining what you know about the topic.
Post Search Questions	Q	How difficult do you judge it was for you to find the information needed to address this task using your assigned search setting (Default Setting or Chat Search interface Setting)? (low (1) - high (7))
	Q	Did you see the full document with the highlighted text? (Yes-No)
	Q	While carrying out the search. Did you recall anything about the subject of you search from your previous experience of this subject? Which you didn't remember when you started the search? (Yes - No)
	Q	If yes, was it helpful to you in satisfying your information need?
	Q	If yes, then please give brief details of what triggered you to remember these details, and say briefly what you recalled.
	Q	Did you learn anything new why carrying out the search?
	Q	If yes, then please give brief details.
	Q	How do you rate the quality of text presented by your assigned search setting with respect to your Information need and query? (low (1) - high (7))
	Q	How useful were the search results in the whole search task? (low (1) - high (7))
	Q	How useful was the text shown in the whole search task in satisfying the Information need? (low (1) - high (7))
	Q	Did you find yourself to be cognitively engaged while carrying out the search task? (low (1) - high (7))

Post Search Questions	Q	Did you expand your knowledge about the topic while completing this search task? (low (1) - high (7))
	Q	I feel that I now have a better understanding of the topic of this task. (low (1) - high (7))
	Q	How would you grade the success of your search session for this topic? (low (1) - high (7))
	Q	How do you rate your assigned search setting in terms of understanding your inputs? (low (1) - high (7))
	Q	How do you rate your assigned search setting in the presentation of the search results? (low (1) - high (7))
	Q	Did your assigned search setting suggest or recommend some options as a query while searching? (Yes - No)
	Q	How do you rate the suggestion(s) skills of your assigned search setting? (very poor (1) - very good (7))
	Q	Having completed the search task, please write a short 4-5 sentence summary describing your knowledge about the topic you were investigating which relates directly to what you learned by carrying out the search task

Table B.3: Questionnaire for Chapter 7 (Rule based MCSI)

B.1.4 Questionnaire NASA TASK Load Index

This questionnaire is used in the study of Chapter 6, Chapter 7 and Chapter 8 to measure the cognitive load on the user while using the system.

Section	**Questions (very low (1) - very high (7))**
Q	How mentally demanding was the task?
Q	How physically demanding was the task?
Q	How hurried or rushed was the pace of the task?
Q	How successful were you in accomplishing what you were asked to do?
Q	How hard did you have to work to accomplish your level of performance?
Q	How insecure, discouraged, irritated, stressed, and annoyed were you?

Table B.4: NASA TASK Load Index Questionnaire for Chapter 7 (Rule based MCSI) and Chapter 8 (Rasa based MCSI)

B.1.5 Software Usability Satisfaction Questionnaires

This questionnaire is used in the study of Chapter 7 and Chapter 8 to measure the software usability experience of the users while information seeking inspired by the study [28].

Section	Questions (very low (1) - very high (7))
Q	Overall, I am satisfied with how easy it was to use this system
Q	It was simple to use this system
Q	I can effectively complete my work using this system
Q	I am able to complete my work quickly using this system
Q	I am able to efficiently complete my work using this system
Q	I feel comfortable using this system
Q	Whenever I make a mistake using the system, I recover easily and quickly
Q	If it was difficult to recover from any mistake, please comment on the problems you faced. (Descriptive Answer)
Q	The information (such as online help, on-screen messages, and other documentation) provided with this system is clear
Q	It is easy to find the information I needed
Q	The information is effective in helping me complete the tasks and scenarios
Q	The organization of information on the system screens is clear
Q	if any information was not clear, what difficulties did you face? (Descriptive Answer)
Q	The interface of this system is pleasant
Q	I like using the interface of this system
Q	This system has all the functions and capabilities I expect it to have
Q	If not, what functions and capabilities would you like to see in this system? (Descriptive Answer)
Q	Overall, I am satisfied with this system

Table B.5: Software Usability Questionnaire for Chapter 7 (Rule based MCSI) and Chapter 8 (Rasa based MCSI)

B.1.6 Questionnaire User Interactive Experience

Questionnaire of Chapters 7 and 8 to measure user interactive experience [150].

Section	Questions (1-7)
Q	While using the System your experience is like (obstructive - supportive)
Q	While using the System your experience is like (complicated - easy)
Q	While using the System your experience is like (inefficient - efficient)
Q	While using the System your experience is like (confusing - clear)
Q	While using the System your experience is like (boring - excited)
Q	While using the System your experience is like (not interesting - interesting)
Q	While using the System your experience is like (conventional - inventive)
Q	While using the System your experience is like (Usual - Leading edge))

Table B.6: User Interactive Experience Questionnaire for Chapter 7 (Rule based MCSI) and Chapter 8 (Rasa based MCSI)

B.2 Example of Back stories

This section focus on the back stories used during the studies.

B.2.1 Back Stories used in Chapter 5, 7 and 8

The following are the backstories for search tasks in the study discussed in Chapters 5, 7 and 8 to understand the user search behaviour in standard search engine and explore the potential for CS in CIS and MCSI interfaces.

S.no	Backstories
B1	You are finalizing the plans for your new house, and one of spaces you are really looking forward to is a well-designed home theater. You decide to read some online advice about how that room should be arranged and set out, so that you get maximum listening and viewing pleasure.

B2	You are baffled by an upcoming event called "Earth Day". However, your nephew's school has asked all kids to take part in Earth Day activities. You therefore decide to find information about what kinds of fun activities children can do to celebrate Earth Day.
B3	You are making choices about studying at college now that high school is almost over. Your career counsellor told you about occupational therapists, and you would like to find out more about the requirements to become one.
B4	You need a second car to cope with the logistics of managing the kids and their different sports. You are considering a Ford Edge, but have heard that maybe it has some problems with it. Find out more.
B5	While you were out gardening, you felt a small sting or bite on your arm where you were pulling out weeds. You think you see a spider in the dirt nearby but it is hard to tell. You want to see if there is a way of identifying whether or not it was one that bit you.
B6	It is late, but you can't get to sleep because a sore throat has taken hold and it is hard to swallow. You have run out of cough drops, and wonder if there are any folk remedies that might help you out until the morning
B7	You are thinking of taking a job in Marshall County, West Virginia, and you'd like to find out more about living there. In particular, you'd like to find any information on the schools there.
B8	Halloween is coming up, and you would like to find some suitable activities for your middle-school-aged children.
B9	You've been having strange, recurring dreams recently. Although you don't really believe they mean anything, you'd be interested to know more about how dreams are generally interpreted.
B10	When you were younger, you always dreamed of being an astronaut. You recently heard that NASA is conducting interplanetary missions. What interplanetary missions has the organization implemented, or planned?

B11	Electronic medical records are now in widespread use, and make health treatments more consistent between providers. But it hasn't always been this way, and you become interested in knowing more about the evolution of electronic health records, and how they have changed through the years to reach their current state.
B12	It seems like ten times a day that you say to your daughter "get off facebook". But recently she has been retorting that social networking is both important as a communications tool, and as an educational medium, and that by spending time at such sites, she is learning. You decide to find out whether there is any evidence to support her claim that such sites have educational benefits.

Table B.7: UQV Backstories used in Chapters 5, 7 and 8

B.2.2 Back Stories For Alexa Task

The following are the backstories for search tasks in the study discussed in Chapter 6 to explore the potential for CS.

S.no	**Backstories**
B1	Mahatma Gandhi is known as father of Nation of India. You have to find the information about the personality using alexa skills (as per the search setting) and based on your Information gain, you have to write a short summary (in questionnaire) about the personality and fill the questionnaire accordingly.
B2	Sourav Ganguly (a legendary cricketer and former captain) who helped India to win many cricket series. You have to find the information about the personality as per the search setting and based on your Information gain, you have to write a short summary (in questionnaire) about the personality and fill the questionnaire accordingly.
B3	Bill Gates is an American business magnate, investor, author, philanthropist, and humanitarian. He is best known as the principal founder of Microsoft Corporation. You have to find the Information about the personality using alexa skills (as per the search setting) and based on your Information gain, you have to write a short summary (in questionnaire) about the personality and fill the questionnaire accordingly.
B4	Steve Jobs was the chairman, chief executive officer (CEO), and co-founder of Apple Inc. You have to find the information about the personality as per the search setting and based on your Information gain, you have to write a short summary (in the questionnaire) about the personality and fill the questionnaire accordingly.

B5	Robert Walpole known between 1725 and 1742 as Sir Robert Walpole, was the first Prime Minister of Great Britain. You have to find the information about the personality using as per the search setting) and based on your Information gain, you have to write a short summary (in the questionnaire) about the personality and fill the questionnaire accordingly.
B6	Donald Trump is the 45th and current president of the United States. You have to find the Information about the personality using alexa skills (as per the search setting) and based on your Information gain, you have to write a short summary (in questionnaire) about the personality and fill the questionnaire accordingly.
B7	Mother Teresa honoured in the Roman Catholic Church as Saint Teresa of Calcutta. She was the first woman to get a Nobel peace prize. You have to find the Information about the personality using alexa skills (as per the search setting) and based on your Information gain, you have to write a short summary (in questionnaire) about the personality and fill the questionnaire accordingly.
B8	Annie Besant was a British socialist, orator, and supporter of both Irish and Indian self-rule. You have to find the Information about the personality using alexa skills (as per the search setting) and based on your Information gain, you have to write a short summary (in questionnaire) about the personality and fill the questionnaire accordingly.
B9	Nelson Mandela was a South African anti-apartheid revolutionary who served as President of South Africa from 1994 to 1999. You have to find the Information about the personality using alexa skills (as per the search setting) and based on your Information gain, you have to write a short summary (in questionnaire) about the personality and fill the questionnaire accordingly.

B10	Angela Merkel is a German politician serving as Chancellor of Germany since 2005. You have to find the Information about the personality using alexa skills (as per the search setting) and based on your Information gain, you have to write a short summary (in questionnaire) about the personality and fill the questionnaire accordingly.
B11	Scott Morrison Scott is an Australian politician serving as Prime Minister of Australia and Leader of the Liberal Party since 2018. You have to find the Information about the personality using alexa skills (as per the search setting) and based on your Information gain, you have to write a short summary (in questionnaire) about the personality and fill the questionnaire accordingly.
B12	Emmanuel Macron is a French politician serving as President of the French Republic and ex officio Co-Prince of Andorra since 2017. You have to find the Information about the personality using alexa skills (as per the search setting) and based on your Information gain, you have to write a short summary (in questionnaire) about the personality and fill the questionnaire accordingly.

Table B.8: Backstories used for Alexa task in Chapter 6

B.3 Interview Questions

The following are the interview questions based on the studies in this thesis.

B.3.1 Interview Questions for Alexa Study (Chapter 6)

The following are the interview questions asked after completing the search session.

Section	Questions
Q	Which Setting would be better with respect to text? and Why? (Default Alexa - Custom Alexa)
Q	Which Setting would you prefer in seeking information ? and Why (Reading Long document - Reading long document with highlighted important segments - Custom Alexa
Q	Concerning Custom Alexa, on what basis you selected the topic from display to seek information?
Q	What is your feedback regarding the whole task?
Q	What are the challenges with task 1?
Q	What are the challenges with task 2?
Q	Which search setting was more useful in satisfying your information need? And why?
Q	What are your suggestions for both the search setting and the whole experiment?

Table B.9: Interview Questions for Chapter 6

B.3.2 Interview Questions for MCSI and CSI (Chapter 7 and Chapter 8)

The following are the interview questions asked after completing the search session.

Section	Questions
Q	Which system was better for you, "Chatbot MCSI" or "CSI" search? And why?
Q	Do you like the full doc (Long document with highlighted text) in the Chat interface? Why?
Q	What are the challenges in searching with traditional search?
Q	Do you find any challenges in the chat interface? if yes, and what
Q	Any suggestion for the chat-interface?

Table B.10: Interview Questions for Chapter 7 and Chapter 8

B.3.3 Sample of Thematic Coding for Alexa

As it is difficult to accommodate all the thematic tables in appendix. Therefore, we demonstrated a sample of our thematic analysis of the interview questions conducted in the Alexa study chapter 6, as shown in table B.11. Here is an example of the thematic coding for the interview questions `"Which search setting was more useful in satisfying your information need? And why?"` where participants prefer custom setting.

Themes	**Count by 1st Analyst**	**Count by 2nd Analyst**	**Agreed coding**
Helpful	20	20	20
Interactive	8	8	8
Provide Suggestions	5	5	5
Interesting	1	4	4
Easy to understand	4	4	4
Structured	2	2	2
Lot of information	2	2	2
Comfortable speed	1	1	1

Table B.11: Sample of thematic coding on interview question on chapter 5.

B.3.4 Sample of Thematic Coding for MCSI

As it is difficult to accommodate all the thematic tables in appendix. Therefore, we demonstrated a sample of our thematic analysis of the interview questions conducted in the MCSI study Chapter 7 and Chapter 8, as shown in table B.12. Here is an example of the thematic coding for the interview questions `"Which system was better for you, "Chatbot MCSI" or "CSI" search? And why??"` where participants prefer MCSI setting.

Themes	Count by 1st Analyst	Count by 2nd Analyst	Agreed coding
Easier Accessibility and Navigation	11	12	11
User friendly and Robustness	4	4	4
Relevant and structured Information	11	11	11
Interactive and Engaging	6	5	6
Options and Suggestions	9	10	9
Increase Efficiency and Reduce Cognitive load	4	4	4
Highlights	5	6	5
Better and Effective Search System	6	6	6

Table B.12: Sample of thematic coding on interview question on chapter 7.

Appendix C

C.1 Publications

- Kaushik A and Jones G.J.F. Exploring current user web search behaviours in analysis tasks to be supported in conversational search. In SIGIR 2nd International Workshop on Conversational Approaches to Information Retrieval (CAIR'18) at SIGIR 2018.

- Kaushik A. Dialogue-Based Information Retrieval. In European Conference on Information Retrieval (pages. 364-368). Springer, 2019.

- Arora P., Kaushik A. and Jones G.J.F. DCU at the TREC 2019 Conversational Assistance Track. Proceeding of TREC 2019. 2019

- Kaushik A, Ramachandra V. and Jones G.J.F. An Interface for Agent Supported Conversational Search. In Conference on Human Information Interaction and Retrieval (CHIIR 2020),2020

- Arora P., Shterionov D.,Moriya Y., Kaushik A., Dzendzik D. and Jones G.J.F.. Openclir 2019 task report. OpenCLIR 2019 Challenge hosted by National Institute of Standards and Technology (NIST), 1:10, 2019.

- Kaushik A., Loir N., Jones G.J.F. Multi-view Conversational Search Interface Using a Dialogue-Based Agent. In: Hiemstra D., Moens MF., Mothe J., Perego R., Potthast M., Sebastiani F. (eds) Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science, vol 12657. Springer, Cham.

- Kaushik A, Jones G.J.F. A Conceptual Framework for Implicit Evaluation of Conversational Search Interfaces. arXiv preprint arXiv:2104.03940. 2021 Apr 8.

- Kaushik A, Bhat Ramachandra V, Jones G.J.F. DCU at the FIRE 2020 Retrieval from Conversational Dialogues (RCD) task.

Bibliography

[1] Benjamin S Bloom et al. "Taxonomy of educational objectives. Vol. 1: Cognitive domain". In: *New York: McKay* 20 (1956), p. 24.

[2] Hans Peter Luhn. "A statistical approach to mechanized encoding and searching of literary information". In: *IBM Journal of research and development* 1.4 (1957), pp. 309–317.

[3] James V Bradley. "Complete counterbalancing of immediate sequential effects in a Latin square design". In: *Journal of the American Statistical Association* 53.282 (1958), pp. 525–528.

[4] Tefko Saracevic. "Relevance: A review of and a framework for the thinking on the notion in information science". In: *Journal of the American Society for information science* 26.6 (1975), pp. 321–343.

[5] Stephen E Robertson and K Sparck Jones. "Relevance weighting of search terms". In: *Journal of the American Society for Information science* 27.3 (1976), pp. 129–146.

[6] J Richard Landis and Gary G Koch. "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers". In: *Biometrics* (1977), pp. 363–374.

[7] J Richard Landis and Gary G Koch. "The measurement of observer agreement for categorical data". In: *biometrics* (1977), pp. 159–174.

[8] Robert N Oddy. "Information retrieval through man-machine dialogue". In: *Journal of documentation* (1977).

[9] James F Allen and C Raymond Perrault. "Analyzing intention in utterances". In: *Artificial intelligence* 15.3 (1980), pp. 143–178.

[10] N. Belkin. "Anomalous states of knowledge as a basis for information retrieval". In: *The Canadian Journal of Information Science* 5 (1980), pp. 133–143.

[11] N.J. Belkin, R.N. Oddy, and H.M. Brooks. "ASK FOR INFORMATION RETRIEVAL: PART I. BACKGROUND AND THEORY". In: *Journal of Documentation* 38.2 (1982), pp. 61–71.

[12] Penny J. Daniels, Helen M. Brooks, and Nicholas J. Belkin. "Using problem structures for driving human-computer dialogues". In: *Proceedings of RIAO 1985*. 1985.

[13] Nicholas J. Belkin, Helen M Brooks, and Penny J Daniels. "Knowledge elicitation using discourse analysis". In: *International Journal of Man-Machine Studies* 27.2 (1987), pp. 127–144.

[14] W Bruce Croft and Roger H Thompson. "I3R: A new approach to the design of document retrieval systems". In: *Journal of the american society for information science* 38.6 (1987), pp. 389–404.

[15] SG Hart and LE Staveland. *Development of NASA-TLX (Task Load Index): Results and theoretical research, Human Mental Workload*. 1988.

[16] Willem JM Levelt. "Speaking: form intention to articulation". In: *Cambridge. MA: The MIT Press* 8 (1989), p. 13.

[17] Gerard Salton and Chris Buckley. "Improving retrieval performance by relevance feedback". In: *Journal of the American society for information science* 41.4 (1990), pp. 288–297.

[18] Carol Collier Kuhlthau. "Inside the search process: Information seeking from the user's perspective". In: *Journal of the American Society for Information Science* 42.5 (1991), pp. 361–371.

[19] Peter Ingwersen. *Information retrieval interaction*. Vol. 246. 1992.

[20] Stefan Sitter and Adelheit Stein. "Modeling the illocutionary aspects of information-seeking dialogues". In: *Information Processing & Management* 28.2 (1992), pp. 165–180.

[21] Shan-Ju Chang and Ronald E Rice. "Browsing: A multidimensional framework." In: *Annual review of information science and technology (ARIST)* 28 (1993), pp. 231–76.

[22] Adelheit Stein and Ulrich Thiel. *A conversational model of multimodal interaction*. GMD, 1993.

[23] Mette Krogsæter, Reinhard Oppermann, and Christoph G Thomas. "A user interface integrating adaptability and adaptivity". In: *Adaptive User Support. Ergonomic Design of Manually and Automatically Adaptable Software* (1994), pp. 97–125.

[24] Pattie Maes. "Agents That Reduce Work and Information Overload". In: *Communications of the ACM* 37.7 (July 1994), pp. 30–40. ISSN: 0001-0782. DOI: 10.1145/176789.176792. URL: http://doi.acm.org/10.1145/176789.176792.

[25] George A Miller. "The magical number seven, plus or minus two: Some limits on our capacity for processing information." In: *Psychological review* 101.2 (1994), p. 343.

[26] James Allen. *Natural Language Understanding (2nd Ed.)* USA: Benjamin-Cummings Publishing Co., Inc., 1995. ISBN: 0805303340.

[27] Nicholas J Belkin et al. "Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems". In: *Expert systems with applications* 9.3 (1995), pp. 379–395.

[28] James R Lewis. "IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use". In: *International Journal of Human-Computer Interaction* 7.1 (1995), pp. 57–78.

[29] Gary Marchionini. *Information Seeking in Electronic Environments*. Cambridge Series on Human-Computer Interaction. Cambridge University Press, 1995. DOI: 10.1017/CBO9780511626388.

[30] Adelheit Stein and Elisabeth Maier. "Structuring collaborative information-seeking dialogues". In: *Knowledge-Based Systems* 8.2-3 (1995), pp. 82–93.

[31] Herbert H. Clark. *Using Language*. 'Using' Linguistic Books. Cambridge University Press, 1996. DOI: 10.1017/CBO9780511620539.

[32] Peter Ingwersen. "Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory". In: *Journal of documentation* 52.1 (1996), pp. 3–50.

[33] Katia Sycara et al. "Distributed intelligent agents". In: *IEEE expert* 11.6 (1996), pp. 36–46.

[34] Micheline Beaulieu. "Experiments on interfaces to support query expansion". In: *Journal of documentation* (1997).

[35] Tefko Saracevic. "The stratified model of information retrieval interaction: Extension and applications". In: *Proceedings of the annual meeting-american society for information science*. Vol. 34. LEARNED INFORMATION (EUROPE) LTD. 1997, pp. 313–327.

[36] Graham Cooper. *Research into cognitive load theory and instructional design at UNSW*. 1998.

[37] Anne De Roeck et al. "YPA-an intelligent directory enquiry assistant". In: *BT Technology Journal* 16.3 (1998), pp. 145–154.

[38] W Iso. "9241-11. Ergonomic requirements for office work with visual display terminals (VDTs)". In: *The international organization for standardization* 45.9 (1998).

[39] Jay Michael Ponte. "A language modeling approach to information retrieval". PhD thesis. University of Massachusetts at Amherst, 1998.

[40] Craig Silverstein et al. *Analysis of a very large AltaVista query log*. Tech. rep. Technical Report 1998-014, Digital SRC, 1998.

[41] Paul Jen-Hwa Hu, Pai-Chun Ma, and Patrick YK Chau. "Evaluation of user interface designs for information retrieval systems: a computer-based experiment". In: *Decision support systems* 27.1-2 (1999), pp. 125–143.

[42] Justine Cassell. "Embodied conversational interface agents". In: *Communications of the ACM* 43.4 (2000), pp. 70–78.

[43] Pertti Vakkari. "Relevance and contributing information types of searched documents in task performance". In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. 2000, pp. 2–9.

[44] James F Allen et al. "Toward conversational human-computer interaction". In: *AI magazine* 22.4 (2001), pp. 27–27.

[45] Lorin W Anderson, Benjamin Samuel Bloom, et al. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman, 2001.

[46] Jonathan Back and Charles Oppenheim. "A model of cognitive load for IR: implications for user relevance feedback interaction". In: *Information Research* 6.2 (2001), pp. 6–2.

[47] Monika R. Henzinger, Rajeev Motwani, and Craig Silverstein. "Challenges in Web Search Engines". In: *SIGIR Forum* 36.2 (Sept. 2002), pp. 11–22. ISSN: 0163-5840. DOI: 10.1145/792550.792553. URL: https://doi.org/10.1145/792550.792553.

[48] David R Krathwohl. "A revision of Bloom's taxonomy: An overview". In: *Theory into practice* 41.4 (2002), pp. 212–218.

[49] Jennifer Lai and Nicole Yankelovich. "Conversational speech interfaces". In: *The human-computer interaction handbook: Fundamentals, evolving Technologies and emerging applications*. 2002, pp. 698–713.

[50] Artemio Ramirez Jr et al. "Information-seeking strategies, uncertainty, and computer-mediated communication: Toward a conceptual model". In: *Human communication research* 28.2 (2002), pp. 213–228.

[51] Darwin P Hunt. "The concept of knowledge and how to measure it". In: *Journal of intellectual capital* (2003).

[52] Mikko Pennanen and Pertti Vakkari. "Students' conceptual structure, search process, and outcome while preparing a research proposal: A longitudinal case study". In: *Journal of the American Society for Information Science and Technology* 54.8 (2003), pp. 759–770.

[53] Pertti Vakkari, Mikko Pennanen, and Sami Serola. "Changes of search terms and tactics while writing a research proposal: A longitudinal case study". In: *Information processing & management* 39.3 (2003), pp. 445–463.

[54] Paul Dourish. "What we talk about when we talk about context". In: *Personal and ubiquitous computing* 8.1 (2004), pp. 19–30.

[55] James Lester, Karl Branting, and Bradford Mott. "Conversational agents". In: *The Practical Handbook of Internet Computing* (2004), pp. 220–240.

[56] Michael K Tanenhaus. "On-line sentence processing: past, present and, future". In: *On-line sentence processing: ERPS, eye movements and beyond* (2004), pp. 371–392.

[57] Mu Zhu. "Recall, precision and average precision". In: *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo* 2 (2004), p. 30.

[58] Elena Barsky and Judit Bar-Ilan. "From the search problem through query formulation to results on the web". In: *Online Information Review* (2005).

[59] Lei Gao. "Latin squares in experimental design". In: (2005).

[60] P Ingwersen. *Järvelin. K.(2005b). The turn: Integration of information seeking and retrieval in context*. 2005.

[61] Alejandro Jaimes and Nicu Sebe. "Multimodal human–computer interaction: A survey". In: *Computer Vision and Image Understanding* 108.1 (2007). Special Issue on Vision for Human-Computer Interaction, pp. 116–134. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2006.10.019. URL: https://www.sciencedirect.com/science/article/pii/S1077314206002335.

[62] Avi Arampatzis and Jaap Kamps. "A Study of Query Length". In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. Singapore, Singapore: Association for Computing Machinery, 2008, pp. 811–812. ISBN: 9781605581644. DOI: 10.1145/1390334.1390517. URL: https://doi.org/10.1145/1390334.1390517.

[63] Nicholas J. Belkin. "Some(What) Grand Challenges for Information Retrieval". In: *SIGIR Forum* 42.1 (June 2008), pp. 47–54. ISSN: 0163-5840. DOI: 10.1145/1394251.1394261. URL: https://doi.org/10.1145/1394251.1394261.

[64] Yu-Ching Chan. "Elementary school EFL teachers' beliefs and practices of multiple assessments". In: *Reflections on English language teaching* 7.1 (2008), pp. 37–62.

[65] Diane Kelly, David J Harper, and Brian Landau. "Questionnaire mode effects in interactive information retrieval experiments". In: *Information processing & management* 44.1 (2008), pp. 122–141.

[66] Bettina Laugwitz, Theo Held, and Martin Schrepp. "Construction and evaluation of a user experience questionnaire". In: *Symposium of the Austrian HCI and usability engineering group*. Springer. 2008, pp. 63–76.

[67] Heather L O'Brien and Elaine G Toms. "What is user engagement? A conceptual framework for defining user engagement with technology". In: *Journal of the American society for Information Science and Technology* 59.6 (2008), pp. 938–955.

[68] Yan Qu and George W. Furnas. "Model-driven formative evaluation of exploratory search: A study under a sensemaking framework". In: *Information Processing & Management* 44.2 (2008), pp. 534–555.

[69] Matthew Richardson. "Learning about the world through long-term query logs". In: *ACM Transactions on the Web (TWEB)* 2.4 (2008), pp. 1–27.

[70] Ian Ruthven. "Interactive information retrieval". In: *Annual review of information science and technology* 42 (2008), pp. 43–92.

[71] Leif Azzopardi. "Query side evaluation: an empirical analysis of effectiveness and effort". In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009, pp. 556–563.

[72] Samir Chatterjee and Alan Price. "Healthy living with persuasive technologies: framework, issues, and challenges". In: *Journal of the American Medical Informatics Association* 16.2 (2009), pp. 171–178.

[73] Djoerd Hiemstra. "Information retrieval models". In: Wiley Online Library, 2009.

[74] Bernard J Jansen, Danielle Booth, and Brian Smith. "Using the taxonomy of cognitive learning to model online searching". In: *Information Processing & Management* 45.6 (2009), pp. 643–663.

[75] Diane Kelly. "Methods for evaluating interactive information retrieval systems with users". In: *Foundations and Trends in Information Retrieval* 3.1—2 (2009), pp. 1–224.

[76] Giridhar Kumaran and Vitor R. Carvalho. "Reducing Long Queries Using Query Quality Predictors". In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '09. Boston, MA, USA: Association for Computing Machinery, 2009, pp. 564–571. ISBN: 9781605584836. DOI: 10.1145/1571941.1572038. URL: https://doi.org/10.1145/1571941.1572038.

[77] Alain Loisel, Nathalie Chaignaud, and Jean-Philippe Kotowicz. "Modeling human interaction to design a human-computer dialog system". In: *arXiv preprint arXiv:0911.5652* (2009).

[78] Harri Oinas-Kukkonen and Marja Harjumaa. "Persuasive systems design: Key issues, process model, and system features". In: *Communications of the Association for Information Systems* 24.1 (2009), p. 28.

[79] Stephen Robertson, Hugo Zaragoza, et al. "The probabilistic relevance framework: BM25 and beyond". In: *Foundations and Trends® in Information Retrieval* 3.4 (2009), pp. 333–389.

[80] Anne Aula, Rehan M Khan, and Zhiwei Guan. "How does search behavior change as search becomes more difficult?" In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2010, pp. 35–44.

[81] Nicholas J Belkin. "On the evaluation of interactive information retrieval systems". In: *The Janus Faced Scholar* (2010), p. 167.

[82] Jacek Gwizdka. "Distribution of cognitive load in web search". In: *Journal of the American Society for Information Science and Technology* 61.11 (2010), pp. 2167–2187.

[83] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. "Introduction to information retrieval". In: *Natural Language Engineering* 16.1 (2010), pp. 100–103.

[84] Sarah K Tyler and Jaime Teevan. "Large scale query log analysis of re-finding". In: *Proceedings of the third ACM international conference on Web search and data mining*. 2010, pp. 191–200.

[85] Simon Attfield et al. "Towards a science of user engagement (position paper)". In: *WSDM workshop on user modelling for Web applications*. 2011, pp. 9–12.

[86] David R Karger, Sewoong Oh, and Devavrat Shah. "Iterative learning for reliable crowdsourcing systems". In: Neural Information Processing Systems. 2011.

[87] Diane Kelly and Karl Gyllstrom. "An examination of two delivery modes for interactive search system experiments: remote and laboratory". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2011, pp. 1531–1540.

[88] Birgit Kopainsky, Stephen M Alessi, and Pål I Davidsen. "Measuring knowledge acquisition in dynamic decision making tasks". In: *The 29th International Conference of the System Dynamics Society*. System Dynamics Society Albany, NY. 2011, pp. 1–31.

[89] James Allan et al. "Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne". In: *ACM SIGIR Forum*. Vol. 46. 1. ACM New York, NY, USA. 2012, pp. 2–32.

[90] Evangelos Kanoulas et al. "Overview of the trec 2011 session track". In: (2012).

[91] Nitendra Rajput and Amit Anil Nanavati. "Evaluation of mobile and pervasive speech applications". In: (2012).

[92] John R Anderson. *Language, memory, and thought*. Psychology Press, 2013.

[93] Virginia Braun and Victoria Clarke. *Successful qualitative research: A practical guide for beginners*. sage, 2013.

[94] Anita Crescenzi, Robert Capra, and Jaime Arguello. "Time pressure, user satisfaction and task difficulty". In: *Proceedings of the American Society for Information Science and Technology* 50.1 (2013), pp. 1–4.

[95] Kellie Morrissey and Jurek Kirakowski. "'Realness' in Chatbots: Establishing Quantifiable Criteria". In: *International Conference on Human-Computer Interaction*. Springer. 2013, pp. 87–96.

[96] Mathew J Wilson and Max L Wilson. "A comparison of techniques for measuring sensemaking and learning within participant-generated summaries". In: *Journal of the Association for Information Science and Technology* 64.2 (2013), pp. 291–306.

[97] Mathew J. Wilson and Max L. Wilson. "A comparison of techniques for measuring sensemaking and learning within participant-generated summaries". In: *Journal of the American Society for Information Science and Technology* 64.2 (2013), pp. 291–306. DOI: 10.1002/asi.22758. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.22758. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22758.

[98] Mira Dontcheva et al. "Combining crowdsourcing and learning to improve engagement and performance". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2014, pp. 3379–3388.

[99] Carsten Eickhoff et al. "Lessons from the journey: a query log analysis of within-session learning". In: *Proceedings of the 7th ACM international conference on Web search and data mining*. 2014, pp. 223–232.

[100] Luanne Freund, Heather O'Brien, and Rick Kopak. "Getting the big picture: supporting comprehension and learning in search". In: *Proceedings SAL* (2014).

[101] Sameera A. Abdul-Kader and Dr. John Woods. "Survey on Chatbot Design Techniques in Speech Conversation Systems". In: *International Journal of Advanced Computer Science and Applications* 6.7 (2015). DOI: 10.14569/IJACSA.2015.060712. URL: http://dx.doi.org/10.14569/IJACSA.2015.060712.

[102] Peter Bailey et al. "User variability and IR system evaluation". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. 2015, pp. 625–634.

[103] Ankur Joshi et al. "Likert scale: Explored and explained". In: *Current Journal of Applied Science and Technology* (2015), pp. 396–403.

[104] Diane Kelly et al. "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*. 2015, pp. 101–110.

[105] Johanne R Trippas et al. "Results presentation methods for a spoken conversational search system". In: *Proceedings of the first international workshop on novel web search interfaces and systems*. 2015, pp. 13–15.

[106] Peter Bailey et al. "UQV100: A Test Collection with Query Variability". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '16. Pisa, Italy: ACM, 2016, pp. 725–728. ISBN: 978-1-4503-4069-4. DOI: 10.1145/2911451.2914671. URL: http://doi.acm.org/10.1145/2911451.2914671.

[107] Kevyn Collins-Thompson et al. "Assessing learning outcomes in web search: A comparison of tasks and query strategies". In: *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. ACM. 2016, pp. 163–172.

[108] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. "Impacts of time constraints and system delays on user experience". In: *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. 2016, pp. 141–150.

[109] Jacek Gwizdka et al. "Search as learning (SAL) workshop 2016". In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 2016, pp. 1249–1250.

[110] Chang Liu, Xiangmin Zhang, and Wei Huang. "The exploration of objective task difficulty and domain knowledge effects on users' query formulation".

In: *Proceedings of the Association for Information Science and Technology* 53.1 (2016), pp. 1–9.

[111] Ewa Luger and Abigail Sellen. "" 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*. 2016, pp. 5286–5297.

[112] Michael McTear, Zoraida Callejas, and David Griol. "Conversational interfaces: Past and present". In: *The Conversational Interface*. Springer, 2016, pp. 51–72.

[113] Michael McTear, Zoraida Callejas, and David Griol. "Creating a conversational interface using chatbot technology". In: *The conversational interface*. Springer, 2016, pp. 125–159.

[114] Michael McTear, Zoraida Callejas, and David Griol. "The dawn of the conversational interface". In: *The Conversational Interface*. Springer, 2016, pp. 11–24.

[115] Drew Paulin and Caroline Haythornthwaite. "Crowdsourcing the curriculum: Redefining e-learning practices through peer-generated approaches". In: *The Information Society* 32.2 (2016), pp. 130–142.

[116] Rebecca B Reynolds. "Relationships among tasks, collaborative inquiry processes, inquiry resolutions, and knowledge outcomes in adolescents during guided discovery-based game design in school". In: *Journal of Information Science* 42.1 (2016), pp. 35–58.

[117] Jeff Sauro and James R Lewis. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann, 2016.

[118] Ameneh Shamekhi et al. "An exploratory study toward the preferred conversational style for compatible virtual agents". In: *International Conference on Intelligent Virtual Agents*. Springer. 2016, pp. 40–50.

[119] Andree Thieltges, Florian Schmidt, and Simon Hegelich. "The devil's triangle: Ethical considerations on developing bot detection methods". In: *2016 AAAI Spring Symposium Series*. 2016.

[120] Johanne R Trippas. "Spoken conversational search: Speech-only interactive information retrieval". In: *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. 2016, pp. 373–375.

[121] Pertti Vakkari. "Searching as learning: A systematization based on literature". In: *Journal of Information Science* 42.1 (2016), pp. 7–18.

[122] Kevin K Bowden et al. "Combining search with structured data to create a more engaging user experience in open domain dialogue". In: *arXiv preprint arXiv:1709.05411* (2017).

[123] Tongfei Chen and Benjamin Van Durme. "Discriminative Information Retrieval for Question Answering Sentence Selection". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 719–725. URL: https://www.aclweb.org/anthology/E17-2114.

[124] M v Eeuwen. "Mobile conversational commerce: messenger chatbots as the next interface between businesses and consumers". MA thesis. University of Twente, 2017.

[125] Srini Janarthanam. *Hands-on chatbots and conversational UI development: build chatbots and voice user interfaces with Chatfuel, Dialogflow, Microsoft Bot Framework, Twilio, and Alexa Skills*. Packt Publishing Ltd, 2017.

[126] Z Liu et al. "Conversation in ir: its role and utility". In: *SIGIR Workshop on Conversational Approaches to IR*. Vol. 222. 2017.

[127] Bhaskar Mitra and Nick Craswell. "Neural Models for Information Retrieval". In: *CoRR* abs/1705.01509 (2017). arXiv: 1705.01509. URL: http://arxiv.org/abs/1705.01509.

[128] Rodrigo Nogueira and Kyunghyun Cho. "Task-oriented query reformulation with reinforcement learning". In: *arXiv preprint arXiv:1704.04572* (2017).

[129] Filip Radlinski and Nick Craswell. "A theoretical framework for conversational search". In: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. ACM. 2017, pp. 117–126.

[130] Nicole M Radziwill and Morgan C Benton. "Evaluating quality of chatbots and intelligent conversational agents". In: *arXiv preprint arXiv:1704.04579* (2017).

[131] R Ramos. "Screw the Turing Test-Chatbots don't need to act human". In: *VentureBeat. Retrieved on March* 13 (2017), p. 2017.

[132] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. "Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S)." In: *IJIMAI* 4.6 (2017), pp. 103–108.

[133] Abigail See, Peter J. Liu, and Christopher D. Manning. *Get To The Point: Summarization with Pointer-Generator Networks*. 2017. arXiv: 1704.04368 [cs.CL].

[134] Thomas Staven. *What Makes a Good Bot or Not?* 2017. URL: https://www.unit4.com/blog/2017/03/what-makes-a-good-bot-or-not%20accessed%20on%2020-02-2019 (visited on 03/22/2017).

[135] Rohail Syed and Kevyn Collins-Thompson. "Optimizing search results for human learning goals". In: *Information Retrieval Journal* 20.5 (2017), pp. 506–523.

[136] Johanne R Trippas et al. "Crowdsourcing User Preferences and query Judgments for Speech-Only Search". In: *1st SIGIR Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*. 2017.

[137] Johanne R. Trippas et al. "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*. CHIIR

'17. Oslo, Norway: ACM, 2017, pp. 325–328. ISBN: 978-1-4503-4677-1. DOI: 10.1145/3020165.3022144. URL: http://doi.acm.org/10.1145/3020165.3022144.

[138] Alexandra Vtyurina et al. "Exploring Conversational Search With Humans, Assistants, and Wizards". In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2017, pp. 2187–2193. ISBN: 9781450346566. URL: https://doi.org/10.1145/3027063.3053175.

[139] Binny Vyas. *6 key metrics to measure the performance of your chatbot*. 2017. URL: https://chatbotslife.com/6-key-metrics-to-measure-the-performance-of-your-chatbot-5fd0adfd0b5b%20accessed%20on%2020-02-2019 (visited on 11/09/2017).

[140] H James Wilson, Paul Daugherty, and Nicola Bianzino. "The jobs that artificial intelligence will create". In: *MIT Sloan Management Review* 58.4 (2017), p. 14.

[141] Henry Akon. "ECHO SHOW Simplified User Manual: A Simplified Step By Step Amazon Echo Show User Guide That Will Help You Explore The Full Capabilities. Of Amazon Echo Show. 100% Simplified!" In: (2018).

[142] Piyush Arora. "Promoting user engagement and learning in search tasks by effective document representation". PhD thesis. Dublin City University, 2018.

[143] Sandeep Avula et al. "SearchBots: User Engagement with ChatBots during Collaborative Search". In: *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval*. ACM. 2018, pp. 52–61.

[144] Leif Azzopardi et al. "Conceptualizing agent-human interactions during the conversational search process". In: *The second international workshop on conversational approaches to information retrieval*. 2018.

[145] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[146] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[147] Mateusz Dubiel et al. "Investigating how conversational search agents affect user's behaviour, performance and search experience". In: *The second international workshop on conversational approaches to information retrieval*. 2018.

[148] Ahmed Fadhil. "Can a chatbot determine my diet?: Addressing challenges of chatbot application for meal recommendation". In: *arXiv preprint arXiv:1802.09100* (2018).

[149] Debjyoti Ghosh et al. "Assessing the Utility of the System Usability Scale for Evaluating Voice-based User Interfaces". In: *Proceedings of the Sixth International Symposium of Chinese CHI*. ChineseCHI '18. Montreal, QC, Canada: ACM, 2018, pp. 11–15. ISBN: 978-1-4503-6508-6. DOI: 10.1145/3202667.3204844. URL: http://doi.acm.org/10.1145/3202667.3204844.

[150] Andreas Hinderks, Martin Schrepp, and Jörg Thomaschewski. "A Benchmark for the Short Version of the User Experience Questionnaire." In: *WEBIST*. 2018, pp. 373–377.

[151] Matthew B. Hoy. "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants". In: *Medical Reference Services Quarterly* 37.1 (2018). PMID: 29327988, pp. 81–88. DOI: 10.1080/02763869.2018.1404391. eprint: https://doi.org/10.1080/02763869.2018.1404391. URL: https://doi.org/10.1080/02763869.2018.1404391.

[152] Abhishek Kaushik and Gareth J. F. Jones. "Exploring Current User Web Search Behaviours in Analysis Tasks to be Supported in Conversational Search". In: *Second International Workshop on Conversational Approaches*

to Information Retrieval (CAIR'18), July 12, 2018, Ann Arbor Michigan, USA. 2018.

[153] Irene Lopatovska et al. "Talk to me: Exploring user interactions with the Amazon Alexa". In: *Journal of Librarianship and Information Science* (2018), p. 0961000618759414.

[154] Gustavo López, Luis Quesada, and Luis A. Guerrero. "Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces". In: *Advances in Human Factors and Systems Interaction*. Ed. by Isabel L. Nunes. Cham: Springer International Publishing, 2018, pp. 241–250. ISBN: 978-3-319-60366-7.

[155] Martin Schrepp. *User Experience Questionnaire*. 2018. URL: https://www.ueq%5C-online.org/%20accessed%20on%2010-01-2021.

[156] Yueming Sun and Yi Zhang. "Conversational recommender system". In: *The 41st international acm sigir conference on research & development in information retrieval*. 2018, pp. 235–244.

[157] Johanne R. Trippas et al. "Informing the Design of Spoken Conversational Search: Perspective Paper". In: *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. CHIIR '18. New Brunswick, NJ, USA: ACM, 2018, pp. 32–41. ISBN: 978-1-4503-4925-3. DOI: 10.1145/3176349.3176387. URL: http://doi.acm.org/10.1145/3176349.3176387.

[158] Anu Venkatesh et al. "On Evaluating and Comparing Conversational Agents". In: *CoRR* abs/1801.03625 (2018). arXiv: 1801.03625. URL: http://arxiv.org/abs/1801.03625.

[159] Yongfeng Zhang et al. "Towards conversational search and recommendation: System ask, user respond". In: *Proceedings of the 27th acm international conference on information and knowledge management*. 2018, pp. 177–186.

[160] Piyush Arora, Abhishek Kaushik, and Gareth J. F. Jones. "DCU at the TREC 2019 Conversational Assistance Track". In: *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019*. Ed. by Ellen M. Voorhees and Angela Ellis. Vol. 1250. NIST Special Publication. National Institute of Standards and Technology (NIST), 2019. URL: https://trec.nist.gov/pubs/trec28/papers/ADAPT-DCU.C.pdf.

[161] Sandeep Avula et al. "Embedding Search into a Conversational Platform to Support Collaborative Search". In: *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 2019, pp. 15–23.

[162] Matthias Biehl. "Making Money with Alexa Skills". In: (2019).

[163] Daniel Braun and Florian Matthes. "Towards a Framework for Classifying Chatbots." In: *ICEIS (1)*. 2019, pp. 496–501.

[164] Stephan Diederich et al. "Design for fast request fulfillment or natural interaction? insights from an experiment with a conversational agent". In: (2019).

[165] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. "A Survey on Conversational Agents/Chatbots Classification and Design Techniques". In: *Web, Artificial Intelligence and Network Applications*. Ed. by Leonard Barolli et al. Cham: Springer International Publishing, 2019, pp. 946–956. ISBN: 978-3-030-15035-8.

[166] Abhishek Kaushik. "Dialogue-Based Information Retrieval". In: *European Conference on Information Retrieval*. Springer. 2019, pp. 364–368.

[167] Abhishek Kaushik, Vishal Bhat Ramachandra, and Gareth J. F. Jones. "DCU at the FIRE 2020 Retrieval from Conversational Dialogues (RCD) task". In: *FIRE 2020 proceeding*. 2019.

[168] Siva Reddy, Danqi Chen, and Christopher D Manning. "Coqa: A conversational question answering challenge". In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 249–266.

[169] Abhishek Singh, Karthik Ramasubramanian, and Shrey Shivam. "Introduction to Microsoft Bot, RASA, and Google Dialogflow". In: *Building an Enterprise Chatbot*. Springer, 2019, pp. 281–302.

[170] Svitlana Vakulenko et al. "QRFA: A data-driven model of information-seeking dialogues". In: *European conference on information retrieval*. Springer. 2019, pp. 541–557.

[171] Avishek Anand et al. "Conversational Search Dagstuhl Seminar 19461". In: *Dagstuhl Reports* 9.11 (2020). Ed. by Avishek Anand et al., pp. 34–83. ISSN: 2192-5283. DOI: 10.4230/DagRep.9.11.34. URL: https://drops.dagstuhl.de/opus/volltexte/2020/11983.

[172] Theo Araujo. "Conversational Agent Research Toolkit: An alternative for creating and managing chatbots for experimental research". In: *Computational Communication Research* 2.1 (2020), pp. 35–51.

[173] Sandeep Avula and Jaime Arguello. "Wizard of Oz Interface to Study System Initiative for Conversational Search". In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 2020, pp. 447–451.

[174] Krisztian Balog et al. "Common Conversational Community Prototype: Scholarly Conversational Assistant". In: *arXiv preprint arXiv:2001.06910* (2020).

[175] Tanja Bunk et al. "Diet: Lightweight language understanding for dialogue systems". In: *arXiv preprint arXiv:2004.09936* (2020).

[176] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. "TREC CAsT 2019: The Conversational Assistance Track Overview". In: *arXiv preprint arXiv:2003.13624* (2020).

[177] DE Hannover. "Evaluating Conversational Search-An Alternate Perspective". In: *Conversational Search* (2020), p. 32.

[178] Mady Mantha. *Introducing DIET: state-of-the-art architecture that outperforms fine-tuning BERT and is 6X faster to train*. 2020. URL: https://blog.rasa.com/.

[179] Alex Moltzau. *Vector Space for Information Retrieval accessed on 08-03-2021*. 2020. URL: https://alexmoltzau.medium.com/vector-space-for-information-retrieval-d00ad3762210.

[180] Tatwadarshi P Nagarhalli, Vinod Vaze, and NK Rana. "A Review of Current Trends in the Development of Chatbot Systems". In: *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE. 2020, pp. 706–710.

[181] Heather L O'Brien, Jaime Arguello, and Rob Capra. "An empirical study of interest, task complexity, and search behaviour on user engagement". In: *Information Processing & Management* 57.3 (2020), p. 102226.

[182] Filipa Perdigão and Marta Ramos. *How Important is Linguistics in Your Chatbot?* https://www.visor.ai/linguistics-in-your-chatbot/. [Online; accessed 05-11-2020]. 2020.

[183] Stephen Roller et al. "Recipes for building an open-domain chatbot". In: *arXiv preprint arXiv:2004.13637* (2020).

[184] Duc Cash Vo. *Design and control conversation flow*. https://docs.microsoft.com/en-us/azure/bot-service/bot-service-design-conversation-flow?view=azure-bot-service-4.0. [Online; accessed 05-11-2020]. 2020.

[185] Alexa Official Website. *Get Started with the Alexa Skills Kit*. 2020. URL: https://developer.amazon.com/en-US/alexa/alexa-skills-kit/start (visited on 03/26/2020).

[186] Hamed Zamani and Nick Craswell. "Macaw: An extensible conversational information seeking platform". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020, pp. 2193–2196.

[187] Abhishek Kaushik, Nicolas Loir, and Gareth J. F. Jones. "Multi-view Conversational Search Interface Using a Dialogue-Based Agent". In: *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR*

2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II. Ed. by Djoerd Hiemstra et al. Vol. 12657. Lecture Notes in Computer Science. Springer, 2021, pp. 520–524. DOI: 10.1007/978-3-030-72240-1_58. URL: https://doi.org/10.1007/978-3-030-72240-1%5C_58.

[188] Beth Lewis. *Using Bloom's Taxonomy for Effective Learning*. https://www.thoughtco.com/blooms-taxonomy-the-incredible-teaching-tool-2081869. [Online; accessed 05-02-2021]. 2021.

[189] Michael Anderson. "Crowdsourcing higher education: A design proposal for distributed learning". In: ().

[190] Sandeep Avula. "Wizard of Oz: Protocols and Challenges in Studying Searchbots to Support Collaborative Search". In:

[191] Raluca Budiu and Page Laubheimer. *Intelligent assistants have poor usability: A user study of Alexa, Google Assistant, and Siri. 2018*.

[192] Geoffrey Leech and Martin Weisser. "Generic speech act annotation for task-oriented dialogues". In: