arXiv:2201.08808v1 [cs.IR] 21 Jan 2022

Conversational Information Seeking

An Introduction to Conversational Search, Recommendation, and Question Answering

Suggested Citation: Hamed Zamani, Johanne R. Trippas, Jeff Dalton and Filip Radlinski (2022), "Conversational Information Seeking", : Vol. xx, No. xx, pp 1–194. DOI: 10.1561/XXXXXXXXX.

Hamed Zamani
University of Massachusetts Amherst
zamani@cs.umass.edu

Johanne R. Trippas
University of Melbourne
johanne.trippas@unimelb.edu.au

Jeff Dalton
University of Glasgow
jeff.dalton@glasgow.ac.uk

Filip Radlinski
Google Research
filiprad@google.com

This article may be used only for the purpose of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval.

now
the essence of knowledge
Boston — Delft

Contents

Conversational Information Seeking

Hamed Zamani[1], Johanne R. Trippas[2], Jeff Dalton[3] and Filip Radlinski[4]

[1] *University of Massachusetts Amherst; zamani@cs.umass.edu*
[2] *University of Melbourne; johanne.trippas@unimelb.edu.au*
[3] *University of Glasgow; jeff.dalton@glasfow.ac.uk*
[4] *Google Research; filiprad@google.com*

ABSTRACT

Conversational information seeking (CIS) is concerned with a sequence of interactions between one or more users and an information system. Interactions in CIS are primarily based on natural language dialogue, while they may include other types of interactions, such as click, touch, and body gestures. This monograph provides a thorough overview of CIS definitions, applications, interactions, interfaces, design, implementation, and evaluation. This monograph views CIS applications as including conversational search, conversational question answering, and conversational recommendation. Our aim is to provide an overview of past research related to CIS, introduce the current state-of-the-art in CIS, highlight the challenges still being faced in the community. and suggest future directions.

Hamed Zamani, Johanne R. Trippas, Jeff Dalton and Filip Radlinski (2022), "Conversational Information Seeking", : Vol. xx, No. xx, pp 1–194. DOI: 10.1561/XXXXXXXXX.

1

Introduction

1.1 Motivation

Over the years, information retrieval and search systems have become more *conversational*: For instance, techniques have been developed to support queries that refer indirectly to previous queries or previous results; to ask questions back to the user; to record and explicitly reference earlier statements made by the user; to interpret queries issued in fully natural language, and so forth. In fact, systems with multi-turn capabilities, natural language capabilities as well as robust long-term user modeling capabilities have been actively researched for decades. However, the last few years have seen a tremendous acceleration of this evolution.

This has been driven by a few factors. Foremost, progress in machine learning, specifically as applied to natural language understanding and spoken language understanding, has recently surged. Whereas the possibility of a conversational information seeking (CIS) system robustly understanding conversational input from a person was previously limited, it can now almost be taken for granted. In concert with this, consumer hardware that supports and encourages conversation has become common, raising awareness of and the expectation of conversa-

tional support in CIS systems. From the research community, this has been accompanied by significant progress in defining more natural CIS tasks, metrics, challenges and benchmarks. This has allowed the field to expand rapidly. This monograph aims to summarize the current state of the art of conversational information seeking research, and provide an introduction to new researchers as well as a reference for established researchers in this area.

1.2 Guide to the Reader

The intended audience for this survey is computer science researchers in fields related to conversational information seeking, as well as students in this field. We do not assume an existing understanding of conversational systems. However, we do assume the reader is familiar with general concepts from information retrieval, such as indexing, querying and evaluation. As this monograph is not a technical presentation of recent machine learning algorithms, we also assume a basic understanding of machine learning and deep learning concepts and familiarity with key algorithms.

The reader will be provided with a summary of the open CIS problems that are currently attracting the most attention, and many promising current results and avenues of investigation. We will also provide an overview of applications attracting interest in the community, and the resources available for addressing these applications.

When discussing the structure of conversations we adopt terminology used in the speech and dialogue research community. The most basic unit is an *utterance* (analogous to a single query in retrieval). All contiguous utterances from a single speaker form a single *turn* (Traum and Heeman, 1996), with a conversation consisting of multiple turns from two or more participants. For the reader we note that somewhat confusingly, a commonly adopted definition in CIS publications defines a turn as the pair of a user turn and a system response turn (a user query and system answer).

The focus of this work differs from recent related surveys. We draw the reader's attention to the following related work. Gao *et al.* (2019)

presented an overview focused on specific neural algorithmic solutions for question answering, task-oriented and chat agents. Freed (2021) also focused on developing chatbots that are often for customer support. Our focus is more on characterizing the problem space related to information seeking conversations and providing a broad overview of different problems, metrics and approaches. Moreover, the report from the third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018) (Culpepper *et al.*, 2018) provided a broader summary of important open challenges in information retrieval, where various challenges associated with CIS were ranked first. That document provides a briefer overview and reading list, more concretely aimed at summarizing open challenges. A more recent report from the Dagstuhl Seminar on Conversational Search (Anand *et al.*, 2020) reiterated these challenges in more detail.

1.3 Scope

This monograph focuses on a particular class of conversational systems, namely those that exhibit key attributes of human conversation. In particular, we take a cue from Radlinski and Craswell (2017), who propose that a conversational system should incorporate mixed initiative (with both system and user able to take initiative at different times), memory (the ability to reference and incorporate past statements), system revealment (enabling the system to reveal its capabilities and corpus), user revealment (enabling the user to reveal and/or discover their information need) and set retrieval (considering utility over sets of complementary items). Here, we study approaches that exhibit at least some of these properties. In particular, we do not delve deeply into *dialogue systems* that restrict themselves largely to identifying slot/value pairs in back and forth exchanges between the system and user.

Additionally, we focus on *information seeking*, which refers to the process of acquiring information through conversation in order to satisfy the users' information needs. This implies that the conversation should exhibit a clear goal or assist the human user in completing a specific task through finding information. While significant progress has been recently made on chit-chat systems, with a primary goal of keeping users

engaged in realistic conversational exchanges over a prolonged time (for example, Ram *et al.*, 2018), we do not attempt to cover such work in depth. Our focus thus aligns more with traditional search concepts such as the presence of an information need or user agenda that existed before they engaged with the CIS system, and which can be satisfied through a conversation.

On the other hand, we do not make a strong distinction between *search* and *recommendation* tasks. Rather, we cover both types of conversational information seeking interactions. We see these as strongly related tasks that are becoming more closely related as time passes. To illustrate, the same task can be seen as either search or recommendation depending on context and interface: A query for "hotels in London" can be seen as either search or recommendation tasks, depending on devices, context, and interfaces.

Finally, we draw attention to three key aspects of CIS that, while having received significant attention, remain largely unsolved. First, the level of natural language comprehension in conversational systems remains far from human-level, particularly over long sequences of exchanges. Even over adjacent conversational steps, question/answer interpretation remains challenging. Second, robust evaluation of conversational systems remains a critical research challenge: The highly personalized and adaptive nature of conversations makes test collection construction highly challenging. We will cover many of the common approaches, and their limitations. Third, *conversation* is sometimes taken to imply voice or speech interactions. We do not make this assumption, recognizing that conversations can happen in many types of interfaces and modalities. We discuss research of conversations using different types of interfaces and presentations in depth.

There are a number of particularly important aspects of conversational information seeking that despite their importance are not covered in depth here, as they apply broadly across many non-conversational search and recommendation tasks. The first is the question of privacy. Clearly this is an essential aspect of all search tasks – and should be considered in depth in any practical system. We refer readers to Cooper (2008) and Zhang *et al.* (2016) as a starting point for privacy

considerations as applied to logging and log analysis.

Similarly, we do not consider the type of information that a user may request or receive – including information that might be considered offensive or harmful. As this issue is not specific to conversational systems and heavily studied, a detailed consideration of such information access is beyond our scope. We refer readers to Yenala *et al.* (2018) as a starting point of recent work in this space.

Along the same lines, fairness is an essential aspect for information seeking and recommendation tasks, yet largely beyond our scope. We note that this includes both fairness in terms of biases that may exist in recommendation to different groups (Ge *et al.*, 2021) as well as fairness when considering both consumers of recommendations as well as producers of items being recommended (Abdollahpouri *et al.*, 2020).

1.4 Applications

An alternative way to characterize the scope of this work could be in terms of the relevant *applications* that are addressed. Section 2 will focus on this formulation, starting with a brief introduction on conversational information seeking (Section 2.3). This includes a discussion of different modalities' (that is, text, speech, or multi-modal) impact on the seeking process, as for instance studied by Deldjoo *et al.* (2021). We then continue with the topic of conversational search and its various proposed definitions (Section 2.5), culminating with one that relates CIS to many other related settings (Anand *et al.*, 2020). Following this, Section 2.6 introduces conversational recommendation (Jannach *et al.*, 2021) followed by conversational question answering in Section 2.7, where for instance Qu *et al.* (2019a) provide a powerful characterization of the relationships between these areas of study. We continue Section 2 by explaining how CIS applications can be used in different domains, and focus on e-commerce, enterprise, and health in Section 2.8. The section concludes with details on intelligent assistants with relation to CIS.

1.5 A High-Level Architecture for CIS Systems

To create a structure for the remainder of this work, we follow the general structure of most CIS systems. This choice guides the main body of this monograph: Each section in this part focuses on one of the core technological competencies that are essential to a modern CIS system.

1.5.1 Conversational Interfaces and Result Presentation

Section 3 provides an overview of conversational interfaces. We begin with a historical perspective, where we explain differences between existing conversational interfaces such as spoken dialogue systems, voice user interfaces, live chat support, and chatbots. This overview illustrates the use of conversations within closely related CIS applications (McTear *et al.*, 2016). Next, research on result presentation through different mediums (desktop or small device) and modalities (text, voice, multimodal) are discussed in Section 3.2, such as recent work by Kaushik *et al.* (2020). This overview emphasizes the difficulties with highly interactive result presentation and highlights research opportunities. Following this, Section 3.3 introduces different kinds of initiative in conversational systems, including system-initiative, mixed-initative, and user-initiative, for instance well-characterized by Zue and Glass (2000) and Wadhwa and Zamani (2021). This section aims to explain the different kinds of initiative, and the consequences on human-machine interactions. We finish the section with a discussion of conversational interfaces limitations including, for instance, limitations as experienced by visually impaired searchers (Gooda Sahib *et al.*, 2015).

1.5.2 Tracking and Understanding Flow

The focus of Section 4 is on the varying approaches that make it possible to follow conversational structure. We begin with an overview of how to represent a single turn, such as is done with Transformer models (Raffel *et al.*, 2020), and how turns are often classified into dialogue acts (Reddy *et al.*, 2019). Section 4.2 then looks at how the different turns of a conversation are usually tied together through state tracking

and text resolution across turns. In particular, the structure of longer conversations is looked at in-depth in Section 4.3, although noting that existing models are limited in their ability to capture long-distance conversational structure (Chiang *et al.*, 2020). We cover work that operates over long-term representation of CIS exchanges in Section 4.4, followed by recent work that attempts to model longer conversations in the final section, epitomized by work on selecting the right context for understanding each turn (Dinan *et al.*, 2019a).

1.5.3 Determining Next Utterances

The next step for a canonical conversational system is selecting or generating a relevant response in the conversational context. This is the focus of Section 5. We begin with an overview of the different types of responses, including short answers, long answers, and structured entities or attributes. The short answer section presents early Conversational QA (ConvQA) systems then discusses the transition to more recent Transformer architectures based on pre-trained language models. Section 5.1.5 then examines how ConvQA is performed over structured knowledge graphs including systems that use key-value networks (Saha *et al.*, 2018), generative approaches, and logical query representations (Plepi *et al.*, 2021). Following this, we discuss open retrieval from large text corpora as part of the QA process. In particular, Section 5.2 goes beyond short answer QA to approaches performing conversational passage retrieval from open text collections including multi-stage neural ranking, for instance recently considered by Lin *et al.* (2021). We briefly discuss long answer generation approaches in Section 5.3 including both extractive and abstractive summarization methods. We conclude the section with conversational ranking of items in a recommendation context, including models that use multi-armed bandit approaches to trade-off between elicitation and item recommendation.

1.5.4 Initiative

Section 6 provides a detailed look at mixed-initiative interactions in CIS systems. We start with reviewing the main principles of developing mixed-initiative interactive systems, and describing different levels of

mixed-initiative interactions in dialogue systems (Allen *et al.*, 1999; Horvitz, 1999). We briefly review system-initiative interactions with a focus on information seeking conversations, such as the work of Wadhwa and Zamani (2021), in Section 6.1. We then delve deeply into intent clarification as an example of important mixed-initiative interactions for CIS in Section 6.2. We introduce taxonomy of clarification and review models for generating and selecting clarifying questions, such as those by Aliannejadi *et al.* (2019) and Zamani *et al.* (2020a). In presenting the work, we include models that generate clarifying questions trained using maximum likelihood as well as clarification maximization through reinforcement learning. Additionally, Section 6.3 discusses preference elicitation and its relation with clarification, followed by mixed-initiative feedback (*i.e.*, getting feedback from or giving feedback to users via sub-dialogue initiation) in Section 6.4.

1.6 Evaluation

Beyond the details of how a CIS system functions, fair evaluation is key to assessing the strengths and weaknesses of the solutions developed. Section 7 looks at evaluation in CIS holistically. After considering possible ways of studying this broad space, this section breaks down evaluation by the setting that is evaluated. Specifically, offline evaluation is treated first, in Section 7.2. A variety of frequently used offline datasets are presented (such as Multi-WOZ (Budzianowski *et al.*, 2018)), and strengths and limitations are discussed including the use of simulators to produce more privacy-aware evaluations as well as the use of non-text datasets. Online evaluation is considered next, with Section 7.2 contrasting lab studies, crowdsourcing, and real-world evaluations. An example of these is where commercial systems may ask evaluation questions of their users (Park *et al.*, 2020). Finally, the metrics applied in these settings are covered in Section 7.4. While readers are referred to Liu *et al.* (2021) for a full treatment, we present an overview of typical turn-level as well as end-to-end evaluation metrics.

1.7 Open Research Directions

Section 8 provides a brief overview of this monograph and discusses different open research directions. We collate the major themes discussed throughout this manuscript instead of presenting a detailed account of all possible future research problems. We highlight four key areas for future exploration. First, Section 8.1.1 covers challenges related to modeling and producing conversational interactions as a way to transfer information between user and system. Second, we highlight the importance of result presentation and its role in CIS research in Section 8.1.2. Third, we emphasise the importance of different CIS tasks in Section 8.1.3 and Section 8.1.4 covers measures of success during the highly interactive CIS process and ultimate evaluation of CIS systems.

1.8 Further Resources

Beyond the main body of this work, Appendix A briefly presents a more holistic historical context for this monograph. This appendix mainly includes information about early research on interactive information retrieval, as well as on dialogue-based information retrieval, such as the I^3R (Croft and Thompson, 1987) and THOMAS (Oddy, 1977) systems (see Section A.1). We discuss approaches for theoretical modelling of interactive information retrieval systems, such as game theory-based models (Zhai, 2016) and economic models (Azzopardi, 2011) in Section A.2. We also include introductory information about existing literature on session search, such as the TREC Session Track, and evaluation methodologies for session search tasks (Carterette *et al.*, 2016) in Section A.3. Finally, we briefly cover exploratory search (White and Roth, 2009) and discuss its relationship to conversational information seeking in Section A.4, followed by a very brief overview of chit-chat and task-oriented dialogue systems in Section A.5. Newcomers to the field of information retrieval are highly encouraged to review this appendix to develop an understanding of where the core ideas behind CIS originated.

2

Definitions and Applications

In this section, we provide relevant concepts from previous work in conversational information seeking (CIS) and its tasks, contexts, and applications illustrating the multi-dimensional nature of CIS. This introductory section aims to guide the reader with background knowledge on definitions and basic concepts related to CIS. We cover three CIS subdomains, namely conversational search, conversational recommendation, and conversational question answering. These topics are closely related and their boundaries are often blurred. We also introduce some domain-specific applications of CIS, including e-commerce, enterprise, and health, and illustrate their use-cases. Lastly, we cover how CIS can be embedded within the subdomain of intelligent assistants.

2.1 Conversation

The term "conversation" carries different definitions in different contexts. The Merriam-Webster Dictionary defines conversation as "oral exchange of sentiments, observations, opinions, or ideas".[1] This refers to the everyday use of conversation by humans. Brennan (2012) defined

[1] https://www.merriam-webster.com/dictionary/conversation

conversation as "a joint activity in which two or more participants use linguistic forms and nonverbal signals to communicate interactively", highlighting the possible use of nonverbal signals in conversations. In contrast, researchers in dialogue systems consider a more pragmatic definition by identifying a few attributes in human conversations. These attributes include turn, speech acts, grounding, dialogue structure, initiative, inference, and implicature (Jurafsky and Martin, 2021, Ch. 24). In this monograph, we provide a new definition of conversation which we believe is best suited for conversational information seeking research.

A conversation is a sequence of interactions between two or more participants, including humans and machines, as a form of interactive communication with the goal of information exchange. Unlike most definitions of conversation in linguistics and dialogue systems that only focus on natural language interactions, we argue that a conversation can also exhibit other types of interactions with different characteristics and modalities, such as click, touch, body gestures, and sensory signals. The reason behind including these interactions is the rich history of using them in search technologies that shape the fundamentals of CIS research. That being said, long form natural language is still the dominant interaction type in conversations. Therefore, a conversation can be defined as follows.

Definition 1. ***Conversation*** is interactive communication for exchanging information between two or more participants (*i.e.*, humans or machines) that involves a sequence of interactions. While natural language is considered as prerequisite for conversational interactions, conversations can also exhibit other types of interaction with different characteristics and modalities (*e.g.*, click, touch, and gestures).

An important characteristic of conversation is its style: *synchronous* versus *asynchronous*. Synchronous conversations happen in real time, where at least two participants (or agents) exchange information. Most human-machine conversations are expected to be synchronous. Asynchronous conversations, on the other hand, happen when information

can be exchanged independent of time. Therefore, asynchronous conversations do not require the participants' immediate attention, allowing them to respond to the message at their convenience. Conversations between humans in forums and email threads are asynchronous. A conversation can also be a mixture of synchronous and asynchronous interactions. For instance, a user can have synchronous interactions with a conversational system and later a human representative can reach out to the user to follow up the conversation and better address the user's needs, if the conversational systems failed.

Researchers in the area of CIS are interested in *information seeking conversations*: conversations in which at least one participant is seeking information and at least another participant is providing information. Information seeking conversations are mostly either among humans (*e.g.*, the interactions between users and librarians for finding information in a library) or between humans and machines (*e.g.*, the interactions between a user and a CIS system). They can be either synchronous, asynchronous, or a mixture of both.

Definition 2. ***Information seeking conversation*** is a conversation (cf. Def. 1) in which the goal of information exchange is satisfying the information needs of one or more participants.

2.2 Interaction Modality and Language in Conversation

According to the mentioned definition of conversation, a conversational system's input channel from the users may involve many different input types such as touch, speech, or body gestures. These signals can be translated through traditional input devices such as a mouse or keyboard. For more modern input devices, users can also input gestures, motion, or touch. The output channels from the conversational system can vary from 2D screens to audio output to holograms.

Users can interact with a conversational system through a range of input devices, including keyboards for typing, microphones for speech,

smartphones for touch, or through a mixture of these and other input devices (Deldjoo *et al.*, 2021). Using a mixture of modalities offer numerous benefits. The key is accessibility; for example, systems with spoken interfaces may be more accessible to users for whom traditional search interfaces are difficult to use (Weeratunga *et al.*, 2015). Even though research in CIS primarily refers to conversation as textual or spoken input, other modalities and the mixture of modalities are receiving increased research attention (Liao *et al.*, 2021; Hauptmann *et al.*, 2020; Deldjoo *et al.*, 2021).

The system output or presentation, similarly to the input from the user, can consist of different output channels. Given the user's device, context, and task complexity, conversational systems need to decide which output modality to use for result presentation (Deldjoo *et al.*, 2021).

2.3 Conversational Information Seeking

CIS, the process of acquiring information through conversations, can be seen as a subset of information seeking (Wilson, 1999). In the case of information seeking, *any* interaction that aids the finding of information is considered. Hence, searching for information in a book is considered part of information seeking. In contrast, CIS specifies the interaction type as conversational in which thoughts, feelings, and ideas are expressed, questions are asked and answered, or information is exchanged. CIS is often partitioned into three subdomains: conversational search, conversational recommendation, and conversational question answering. However, we do not make a strong distinction between these subdomains. The reason is that the boundaries between these subdomains are blurred. For instance, a system that helps a user to find and purchase shoes through a conversational interface can be seen as either a conversational search or conversational recommendation. Or a system that answers a sequence of non-factoid questions by retrieving passages can be seen as either conversational search or conversational question answering. Therefore, this monograph focuses on conversational information seeking in general and describes models, theories, and techniques that can be used across all CIS subdomains. We define CIS systems as follows:

Definition 3. A ***Conversational Information Seeking (CIS) system*** is a system that satisfies the information needs of one or more users by engaging in information seeking conversations (cf. Def. 2). CIS responses are expected to be concise, fluent, stateful, mixed-initiative, context-aware, and personalized.

In this definition, we provide a number of properties that are expected from CIS systems. They are explained in the next subsection. Even though we believe that there is no clear distinction between CIS subdomains (depicted in Figure 2.1), we describe prior work that focused on each of these subdomains in Sections 2.5 – 2.7.

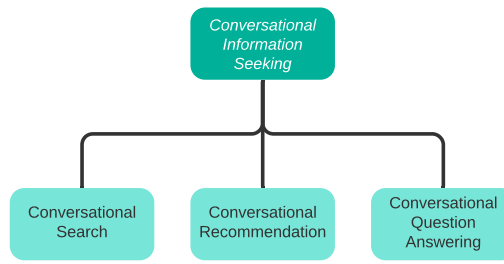

Figure 2.1: Conversational Information Seeking and example subdomains including conversational search, conversational recommendation, and conversational question answering.

2.4 System Requirements of CIS Systems

To create a truly conversational system, it has been argued that the system should pro-actively participate in the conversation (Andolina *et al.*, 2018; Avula and Arguello, 2020; Tabassum *et al.*, 2019; Trippas *et al.*, 2018; Vuong *et al.*, 2018; Wadhwa and Zamani, 2021). This requires *mixed-initiative*, which implies that the system both responds to utterances, but also at times drives the conversation. Furthermore, the user-system interactions should create a *multi-turn* dialogue where each participant takes multiple turns to state their information need, clarify this need, or maintain communication functions such as discourse

management (Aliannejadi *et al.*, 2019; Deits *et al.*, 2013; Trippas *et al.*, 2020; Zamani *et al.*, 2020a). Indeed, systems can utilize interactive feedback signals such as clarifying questions to optimize the advantages of the conversational technique (Aliannejadi *et al.*, 2019; Zamani *et al.*, 2020a). Mixed-initiative interactions, and in particular clarifying questions, are thoroughly reviewed in Section 6.

The requirements of a system to support the users' interactions are multiple. For example, the interaction history (*e.g.*, queries, relevance feedback, type of interaction device) has to be saved and, where necessary, retrieved by the system (Reichman, 1985; Zamani and Craswell, 2020). The interaction history as well as user-specific and contextual information can be adopted to provide *personalized* and *context-aware* access to information. A system should also be able to adapt the results presentation strategies depending on the users' needs. This could be that a user is cognitively engaged, in which case the system can present the results *concisely and fluently with a high comprehensibility*. We note that conciseness and fluency are not specific to natural language and it should be extended to multi-modal conversations. For instance, in speech-only setting, the CIS outputs are expected to be "listenable" (Trippas, 2019).

Due to the proposed interactive, adaptive, and conversational nature of these user-system interactions, both user and system turn-time can be less predictable. For example, if the users' input is natural language-based, it can increase the time needed to convey their information need versus a query-based information need. Simultaneously, a system can ask clarifying questions to overcome errors and thus engage with the user through multiple interactions (Skantze, 2007).

One system requirement which is particularly relevant to a speech-only setting is the system's ability to assist the user when speech recognition errors have occurred (Trippas *et al.*, 2018). These errors may occur due to background noise, speaker accents and spoken language ability, or out-of-vocabulary words among other reasons. Speakers often compensate with hyper-articulation or restarting voice inputs (Jiang *et al.*, 2013; Myers *et al.*, 2018). It has been suggested that systems should design for ways to handle the myriad of possible errors and use meta-communication to overcome them (Trippas, 2019).

Overall, a CIS system is concerned with dialogue-like information seeking exchanges between users and system. Furthermore, the system is pro-actively involved with eliciting, displaying, and supporting the user to satisfy their information need through multi-turn transactions, which can be over one or more sessions. We note that given the complexity of the system and properties listed in Definition 3, most research articles make several simplifying assumptions. For instance, TREC Conversational Assistance Tracks 2019 - 2021 (Dalton *et al.*, 2019; Dalton *et al.*, 2020b; Dalton *et al.*, 2020c) ignore some of these properties including personalization and mixed-initiative interactions.

2.5 Conversational Search

Conversational search, or the process of interacting with a conversational system through natural conversations to search for information, is an increasingly popular research area and has been recognized as an important new frontier within IR (Anand *et al.*, 2020; Culpepper *et al.*, 2018). Furthermore, mobile devices and commercial intelligent assistants such as Amazon Alexa, Apple's Siri, and Google Assistant, in which users interact with a system to search for information, are becoming accepted. We can use these systems to receive weather updates, directions, calendar items, and information on any topic covered on the Internet by stating our needs in natural language.

Information seeking, or the process by which people locate information, has traditionally been viewed as a highly interactive process (Oddy, 1977; Croft and Thompson, 1987). More specifically, searching has been approached as an interactive user-system activity for many years. Furthermore, with the rise in machine learning (ML), natural language processing (NLP), and spoken language comprehension, understanding many users' natural language statements has become more feasible. Simultaneously, with ever-growing computing power, it has been easier to comprehend, categorise, or analyze major datasets, which helped to develop genuinely interactive systems that go beyond the conventional "query box" action-reaction search model (Trippas *et al.*, 2018). For example, instead of posing a query word in which the user needs to filter through a search engine results page (SERP), the user can describe

their information need. In addition, the system could inform the user in a more conversational style which documents might be relevant to the query and thus have the system actively involved in the search process. As described by Radlinski and Craswell (2017), the system could reason about the retrieved documents and actively help the user sift through the information. Intuitively, as a new interaction paradigm, conversational search opens up many different possibilities. For example, we may learn how to optimize traditional browser-based "query box" searching, improve information accessibility, and decrease information access barriers by incorporating search into everyday dialogues (Balasuriya *et al.*, 2018; Trippas *et al.*, 2021).

For example, when someone is inquiring about whether superannuation is compulsory in Australia within a traditional browser-based search engine, they may input the following:

USER: Compulsory superannuation

SYSTEM: In Australia, superannuation, or "super", is compulsory for all people working and residing in Australia. The balance of a person's superannuation is used to provide an income stream when retiring. Federal law dictates minimum amounts that employers must contribute to the employees' super accounts on top of standard wages or salaries. Employees are also encouraged to supplement compulsory superannuation contributions with voluntary contributions, including diverting their wages or salary income into superannuation contributions under so-called salary sacrifice arrangements.

As seen in the above example, the statement from the user is short and resembles a keyword style query and the system response is a long and information-dense passage making it hard for the user to consume. In addition, the presentation of the results is not interactive, but instead, all the information is presented in one turn, thus inhibiting the strength of interactivity as an interaction paradigm. Furthermore, the user cannot interact with the content through query refinements or clarifications. This also reinforces the perceived importance of the initial

user query, requiring them to formulate "excellent" queries from the beginning (Gooda Sahib *et al.*, 2015). In contrast to the above example, the following instance is an example of a conversational search dialogue and enables the user to provide their query in a more natural style:

USER: Is it compulsory to contribute to superannuation?

SYSTEM: Do you mean in Australia?

USER: Yes.

SYSTEM: Yes, in Australia, the law states that employers must contribute to the employees' super accounts called the "superannuation guarantee". Do you want to know more about the set percentages of contributions?

USER: No, not really but I can add some more money from my own?

SYSTEM: Yes, you can voluntary supplement the compulsory superannuation with is often referred to as salary sacrifice.

The dialogue in this example is more natural and involves greater natural language exchanges. The dialogue is intuitively divided into pieces to minimise information overload. Furthermore, the system recognizes the user's context, assisting them in refining their enquiry, and maintains an account of prior encounters, eliminating the need for repetition. In addition, the system creates a model of the user and their information needs through problem elicitation. All of these user-system interactions are made possible by both sides conversing in a human-like manner.

As part of CIS, several definitions of conversational search have been proposed (Anand *et al.*, 2020; Radlinski and Craswell, 2017; Azzopardi *et al.*, 2018; Trippas *et al.*, 2019), which are all inline with the CIS definition provided earlier in this section. For example, researchers who attended the Dagstuhl seminar on Conversational Search created a typology based on existing systems as a definition (Anand *et al.*, 2020). Radlinski and Craswell (2017) and Azzopardi *et al.* (2018) viewed the

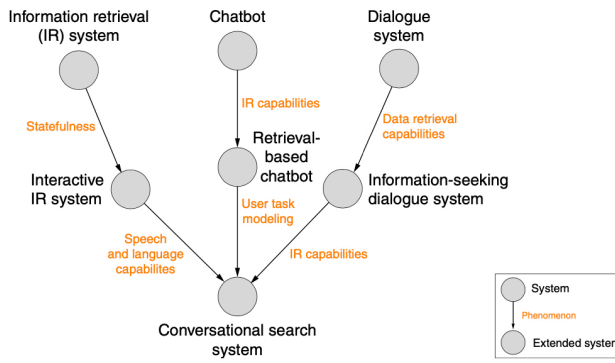

Figure 2.2: The Dagstuhl Conversational Search Typology defines the systems via functional extensions of IR systems, chatbots, and dialogue systems (Anand *et al.*, 2020).

process mainly from a theoretical and system perspective, while Trippas (2019) viewed it from a cognitive, user-system, and empirical perspective.

As seen in Figure 2.2, the Dagstuhl typology aimed to position conversational search with respect to other disciplines and research areas. For instance, they drew the lines from IR systems and added properties such as statefulness to derive IIR systems and thus specify conversational search as:

> *"A conversational search system is either an interactive information retrieval system with speech and language processing capabilities, a retrieval-based chatbot with user task modeling, or an information seeking dialogue system with information retrieval capabilities."* (Anand *et al.*, 2020, p. 52)

While Radlinski and Craswell (2017) define conversational search systems with a more focused and applied view on which properties need to be met.

> *"A conversational search system is a system for retrieving information that permits a mixed-initiative back and forth between a user and agent, where the agent's actions are chosen in response to a model of current user needs within*

> *the current conversation, using both short- and long-term knowledge of the user."* (Radlinski and Craswell, 2017, p. 120)

Lastly, Trippas (2019) expanded on Radlinski and Craswell's definition and stated that for spoken conversational search:

> *"A spoken conversational system supports the users' input which can include multiple actions in one utterance and is more semantically complex. Moreover, the conversational system helps users navigate an information space and can overcome standstill-conversations due to communication breakdown by including meta-communication as part of the interactions. Ultimately, the conversational system multi-turn exchanges are mixed-initiative, meaning that systems also can take action or drive the conversation. The system also keeps track of the context of particular questions, ensuring a natural flow to the conversation (i.e., no need to repeat previous statements). Thus the user's information need can be expressed, formalised, or elicited through natural language conversational interactions."* (Trippas, 2019, p. 142)

All of these definitions look at the provided CIS definition from a search perspective by focusing on retrieving/selecting information items.

2.6 Conversational Recommendation

Recommender systems can be seen as information seeking systems that provide users with potentially relevant items based on historical interactions. Unlike a conventional search engine that takes a query as input, a recommender system uses past user-item interactions as a way to produce relevant recommendations (Konstan and Riedl, 2012). As such, traditional recommender systems aim to help users filter and select items for their information need, often in a closed domain such as books, restaurants, or movies. These systems select possible items from an extensive database and filter through facets to present the user with

the best suitable option (Resnick and Varian, 1997; Thompson *et al.*, 2004).

Recently, two survey papers on conversational recommender systems have proposed definitions of this research area as:

> *"A conversational recommender system is a software system that supports its users in achieving recommendation-related goals through a multi-turn dialogue."* (Jannach *et al.*, 2021, p. 105)

and

> *"A recommendation system [ed. conversational recommender system] that can elicit the dynamic preferences of users and take actions based on their current needs through real-time multi-turn interactions."* (Gao *et al.*, 2021a, p. 101)

Based on the above definitions and similar to conversational search, conversational recommender systems ultimately should be *multi-turn*, meaning that there is more than one interaction or two utterances (*i.e.*, one utterance from the user and one from the system). Current conversational recommender systems can answer recommendation requests reasonably well, but have difficulties maintaining multi-turn conversations (Jannach *et al.*, 2021).

Even though the usage of multi-turn interactions could imply some kind of memory that can keep track of the communication and current state, previous definitions fail to mention this fundamental requirement for conversational recommendation. Indeed, some form of user-system interaction history with conversational recommender systems is necessary for a system to be able to provide recommendations based on those previous interactions. Thus, storing past interactions to refer to is a key component, similarly to conversational search. At the same time, it is important to simultaneously consider privacy implications of such an interaction history: What exactly is being retained, how it may be used in future, and how people can control what is stored. This is currently an open area of research.

Conversational recommender systems are sometimes referred to as a "systems ask, users answer" paradigm (Sun and Zhang, 2018; Zhang *et al.*, 2018). This means that only the recommender system could ask questions to elicit users' preferences. Furthermore, this approach involving one-way elicitation can have difficulties capturing the users' needs thoroughly. However, more recent work in conversational recommender systems has been investigating this rigid paradigm, introducing the *mixed-initiative* approach (Ren *et al.*, 2020b). Indeed, a conversational recommender system should be able to elicit, acquire, store, and utilize user preferences through implicit (*e.g.*, clicking) or explicit (*e.g.*, rating) user feedback (Pommeranz *et al.*, 2012; Christakopoulou *et al.*, 2016). This implies that conversational recommender systems should be capable of taking the initiative and thus support mixed-initiative interactions. An example of acquiring the user's preference can be seen in Figure 2.3.

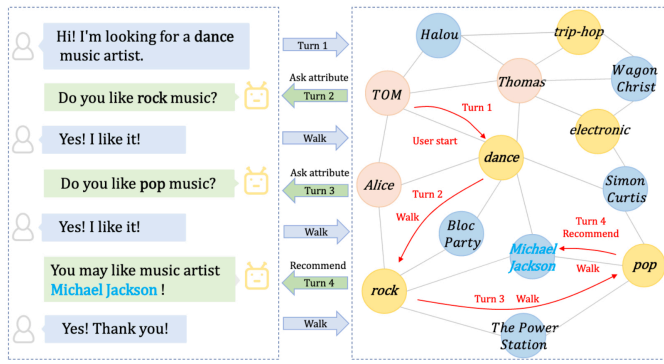

Figure 2.3: An example of user interactions with conversational recommender systems from Lei *et al.* (2020b) with each interaction demonstrating the reasoning.

A fundamental characteristic of conversational recommender systems is that they support specific tasks and goals. The system should suggest recommendations while the user interacts with that system to help them find relevant information and thus support the user's decision making process.

Another way to elicit user preferences is through product reviews. However, one drawback of this method is that the user must have

reviewed for the system to create a user profile (Chen *et al.*, 2015). Conversational interactions may overcome this issue by simply engaging the user in a conversation about products they liked or disliked in the past or the most important features of products for them (Iovine, 2020). Another advantage of the conversational format for recommendations is to explain why (and/or how) particular items are retrieved (Laban and Araujo, 2020). In conversational search, users submit a query and explain their information need, which means there can be some transparency on why the system retrieves the given results. However, the decision making process in recommender systems is much less visible for the users since it is based on prior interactions. Further research on systems that reason and explain through natural language and conversational actions why particular results are retrieved, how they yield ethically sourced recommendations that are culturally relevant, and respect laws and societal norms is warranted. By providing explanations, conversational systems will enhance human decision-making and will also be improved from ethical standpoints.

Conversational search and conversational recommender systems share many commonalities. In essence, both tasks aim to provide users with relevant items based on a ranking, either through a query (search) or user preference (recommendation). This point has been raised in the 1990s by Belkin and Croft (1992) and has recently been reconsidered in (Zamani and Croft, 2020a; Zamani and Croft, 2020b). Furthermore, both systems will interact through *conversations* with the system and share the same characteristics to interaction modality (see Section 2.2). In conclusion, as been repeatedly mentioned, the boundaries between these CIS applications are often blurred, mainly because many comparable technological and computing methods are applied. Using the strengths and advances from each CIS subdomain will move the area of conversational systems forward.

2.7 Conversational Question Answering

Question answering (QA), the task of providing one of more answer(s) to a given question, has been a longstanding information seeking task within the IR and NLP communities (Dwivedi and Singh, 2013;

Kolomiyets and Moens, 2011; Warren and Pereira, 1982; Winograd, 1974). One of the first QA systems were created in the 1960s and 70s, such as BASEBALL (Green *et al.*, 1961) and LUNAR (Woods *et al.*, 1972). Both systems were an interface to a structured database that could be accessed through very restricted natural language questions. The subject domain was also very restricted so that the user query could be processed and parsed through a manually created domain-specific vocabulary.

Other early systems, usually created for a specific domain, include SHRDLU by Winograd (1974) and CHAT-80 QA by Warren and Pereira (1982). The SHRDLU system was designed as an interactive dialogue interface to give commands, ask questions, or make statements while the system could react by carrying out the commands, answering questions, and taking in new information. However, this early system had some missing system capabilities. For example, as Winograd (1974) explained, the system was narrow and only accepted a limited range of information, specifically in understanding human language and the reasoning behind these interactions.

QA is a specific form of information seeking where the users' needs are expressed in a form of (natural language) question. For example, a QA query could be "Which country has the longest period without a government?". QA questions are also frequently classified by common properties and can often be classified as factoid, list, definition, relationship, procedural, and conformation questions (Kolomiyets and Moens, 2011). These particular question types have specific characteristics, such as a factoid question often starts with WH-interrogated words (what, when, where, who) and list questions often start with *List/Name [me] [all/at least NUMBER/some]* according to Kolomiyets and Moens (2011).

In contrast to classical IR, in which full documents are considered relevant to the user's need, QA is concerned about finding and presenting relatively short pieces of information to answer the queries. Therefore, QA uses NLP and IR techniques to retrieve small text snippets containing the exact answer to a query instead of the document lists traditionally returned by text retrieval systems (Voorhees *et al.*, 1999; Gao *et al.*, 2019). The short answers are often retrieved and pre-

sented as short text passages, phrases, sentences, or knowledge graph entities (Lu *et al.*, 2019).

With the developments around conversational systems, the QA work received increased attention in the context of CIS (Christmann *et al.*, 2019; Qu *et al.*, 2019b; Kaiser *et al.*, 2020). Conversational QA (ConvQA) can be seen as a subsection of CIS but with a narrower focus than conversational search. Even though ConvQA is a popular research topic, no comprehensive definition of ConvQA has been proposed so far. The main reason is that it is difficult to distinguish it from many conversational search tasks.

Traditionally, QA has been focused on a single question, which means that no historical interaction data is kept. However, it could be argued that conversations should be composed of more than one interaction. Thus, in conversational QA, more than one question may be posed by the user. Furthermore, as explained in earlier sections, conversational interactions imply that the history of previous dialogues is kept and used to answer the user's questions enabling follow-up questions or references to earlier concepts. Using the advantage of the conversational aspect, users can query the system interactively without having to compose complicated queries (Gao *et al.*, 2019). However, to correctly answer the user's question, ConvQA systems need to handle more complex linguistic characteristics of conversations, such as anaphoras (words that explicitly refer to previous conversational turns) or ellipsis (words that are redundant in the conversation) (Vakulenko *et al.*, 2020).

An example of a series of ConvQA interactions is seen in Figure 2.4. Furthermore, ConvQA is often seen in relation to machine comprehension (Yang *et al.*, 2018b), which is often based on questions about a given passage of text. The main difference is that machine comprehension organizes the questions into conversations (Qu *et al.*, 2019a). This means that leveraging the history is crucial to creating robust and effective ConvQA systems. For example, history can help map the state and changes of the information need to inform current or future responses. Recent work from Kaiser *et al.* (2020) also mentions the importance of dialogue context to improve ConvQA. That is, the user in later interactions can refer to the implicit context of previous utterances.

q_1: *When was Avengers: Endgame released in Germany?*
ans_1: *24 April 2019*
q_2 : *What was the next from Marvel?*
ans_2: *Spider-Man: Far from Home*
q_3 : *Released on?*
ans_3: *04 July 2019*
q_4: *So who was Spidey?*
ans_4: *Tom Holland*
q_5 : *And his girlfriend was played by?*
ans_5: *Zendaya Coleman*

Figure 2.4: An ideal conversational QA interaction example with five turns from Kaiser *et al.* (2021) where q_i and ans_i are questions and answers at turn i, respectively.

2.8 Conversational Information Seeking in Different Domains

As a paradigm to interact with information, CIS can find items on the web, databases, or knowledge graphs. Conversational information access can also be applied to specific domains such as the financial industry, hospitality, or cooking recipes. This section expands different domains where CIS can be applied in addition to their unique properties. These domains include e-commerce, enterprise, and health.

2.8.1 Conversational Information Seeking in E-Commerce

Finding and buying products through conversational interactions is becoming popular (Papenmeier *et al.*, 2021). E-commerce transactions, the process of buying and selling of goods and services online, are steadily increasing.[2] Simultaneously, with the uptake of CIS systems with consumers (*e.g.*, Amazon Alexa or Google Assistant), it becomes increasingly easier to identify consumers context, resulting in more accurate context-aware responses.

It has been suggested that conversational e-commerce search and task-oriented dialogues share commonalities. For example, the dialogue for flight reservation and e-commerce will elicit user preferences such as flight destinations akin to an e-commerce product (Yang *et al.*, 2018b).

[2] https://www.forbes.com/sites/joanverdon/2021/04/27/global-ecommerce-sales-to-hit-42-trillion-as-online-surge-continues-adobe-reports/

However, differences between task-oriented dialogue systems and e-commerce queries have been observed, making e-commerce information need expression much more complex (Yang *et al.*, 2018b). For instance, e-commerce products often have different facets, such as brand, color, size, or style, resulting in different preference slot combinations or shopping schema. Thus, e-commerce schemas can be complex. It is even suggested that they can be incomplete due to the extended range of product facets. Zhang *et al.* (2018) suggested that user-system interactions in e-commerce CIS systems can be classified into three stages which are initiation, conversation, and display. In their proposed paradigm, the system will loop through questions to understand all user preferences of the product's facets before presenting the user's query results.

Some advantages of using CIS in e-commerce include accessing products through conversational-enabled devices such as mobile phones or smart devices. Furthermore, instead of going to a shop for support, customers can access help instantly through these devices. In addition, when users are logged in to their shopping profile, personalization and shopping history can optimize shopping experiences. Conversely, CIS systems embedded in an intelligent assistant have the potential to be a virtual shopping assistant.

2.8.2 Conversational Information Seeking in Enterprise

An application of CIS which has not received much attention is searching through conversational interactions in an enterprise setting (Teevan, 2020). CIS enterprise systems aim to help people in a work environment such as meeting rooms and at desks, with predictions that by 2025, 50% of knowledge workers will use a virtual assistant daily. This prediction is up from 2% in 2019.[3] Even though there has been an increased interest in workplace-oriented digital assistants in general (*e.g.*, Alexa for Business[4] or Cortana Skills Kit for Enterprise[5]), the uptake has been limited.

[3] https://blogs.gartner.com/anthony_bradley/2020/08/10/brace-yourself-for-an-explosion-of-virtual-assistants/

[4] https://aws.amazon.com/alexaforbusiness/

[5] https://blogs.microsoft.com/ai/cortana-for-enterprise/

It is well known that enterprise search applications have different needs than a traditional web search engine, including challenges such as searching over enterprise Intranets or multiple internal sources (Hawking, 2004). Furthermore, besides the use of CIS systems in a traditional office environment, many different applications of more varied and complex environments such as airplane pilots create an extra layer of complexity (Arnold *et al.*, 2020; Gosper *et al.*, 2021). Many open problems in the intersection of CIS applications and enterprise need further investigation. In particular, issues such as defining appropriate test collections, effective conversational search over distributed information sources, identifying tasks that lend themselves to use a CIS application, and understanding the way employees interact with these systems need to be investigated.

2.8.3 Conversational Information Seeking in Health

Searching for health information is a logical application for CIS. Many people already search for health advice online. For example, people will go to symptom checkers to understand if they have an underlying health condition or try to identify whether they need to seek professional advice (Cross *et al.*, 2021). Furthermore, a recent study of a CIS application to enable patients to search for cancer-related clinical trials suggest that CIS could help to make health information more accessible for people with low health or computer literacy skills (Bickmore *et al.*, 2016).

A recent survey suggests that the main areas of CIS applications literature are located in areas for patients such as treatment and monitoring, health care service support, and education (Car *et al.*, 2020). However, besides patients, user groups such as carers and other health professionals can benefit from these systems. For example, in a study where physicians used an information seeking chatbot, they reported that the advantages of CIS include diagnosis helping and that it could be used as a research tool (Koman *et al.*, 2020).

Even though CIS has major potential, some concerns about implementing these systems in the health domain need to be addressed. For example, these systems may not have sufficient expertise to answer all

questions and may even misinterpret or misunderstand these questions, potentially providing a wrong answer (Su *et al.*, 2021). Although a common challenge to all search systems, this may be exacerbated in a CIS setting if a system were to naively present health *mis*information in a way that reinforces it. Furthermore, these systems can deal with sensitive patient data and thus need to be safeguarded. Voice-only CIS systems may also encounter issues with speech recognition, especially when people are distressed or are in noisy environments (Spina *et al.*, 2021).

2.9 Intelligent Assistants

Intelligent assistants are often associated with CIS and are rising in popularity. The number of intelligent voice assistants worldwide is predicted to double between 2020 and 2024, from 4.2 billion to 8.4 billion.[6] Intelligent assistants are frequently embedded in existing phones, laptops, mobile devices or smart speakers. For instance, assistants such as Google Assistant, Amazon's Alexa, AliMe, or Apple's Siri enable users to receive assistance on everyday tasks with a specific goal (*e.g.*, turning on or off appliances) or conduct simple question answering tasks such as asking the weather forecast or the news. With the increase in mobile devices and mobile internet connections, users instantly have access to powerful computational and digital intelligent assistants. These intelligent assistants may be designed to access the user's situation or context through GPS locations, the people who are around them through Bluetooth scans, and previous interactions users had with their electronic devices (Liono *et al.*, 2020; Trippas *et al.*, 2019) when enabled on the mobile device. However, more research is needed to use all the contextual signals to optimize CIS responsibly and with user privacy in mind.

Different CIS tasks may require access to different knowledge sources and databases. Intelligent assistants need to disambiguate which knowledge source they need to retrieve the information. For instance, Aliannejadi *et al.* (2018b) introduced the problem of unified mobile search,

[6] https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use

in which intelligent assistants identify the target mobile apps for each search query, route the query to the selected apps, and aggregate the search results. In a follow up work, the authors demonstrated the impact of user context and app usage patterns on unified mobile search (Aliannejadi *et al.*, 2018a; Aliannejadi *et al.*, 2021b). Identifying knowledge sources was also used in the Ninth Dialog System Technology Challenge (DSTC9) with a track called "Beyond domain APIs - Tasks-oriented conversational modeling with unstructured knowledge access". This track aimed to expand different task-oriented dialog systems by incorporating external unstructured knowledge sources (Gunasekara *et al.*, 2020). The track's purpose was to investigate how to support *frictionless* task-oriented situations so that the flow of the conversation does not break when users have questions that are out of the scope of APIs/DB but possibly are available in external knowledge sources.

Other applications that can incorporate CIS systems are embodied robots, *e.g.*, the Multi-Modal Mall Entertainment Robot (MuMMER) (Foster *et al.*, 2016). MuMMER was a collaborative challenge in which a robot was made to behave appropriately to human social norms and engage through speech-based interactions. Similarly, social bots enable users to search and engage in information dialogues. This has been thoroughly studied in the context of Alexa Prize Socialbot Challenge (Ram *et al.*, 2018). Although these interactions involving search for information may differ from a focused CIS system, embedding CIS enables a wider variety of use-cases.

2.10 Summary

This section provided a high-level overview of CIS and its applications. We first started with providing definitions for conversation, information seeking conversation, and CIS systems. Under these definitions, conversational search, conversational question answering, and conversational recommendation are seen as the subdomains of conversational information seeking tasks. This section also included a number of system requirements that are expected from CIS systems.

We later reviewed previous work that characterizes the three subdomains of CIS and discussed their connections. We lastly provided an

overview of how CIS can be used in particular domains and compared CIS to intelligent assistants. CIS is still being developed and is rapidly expanding as a multi-dimensional and multi-disciplinary research area. Overall, this section summarized prior work in conversational information seeking applications to provide an overview.

3

Conversational Interfaces and Result Presentation

The emergence of conversational systems has empowered the development of a new kind of human–computer interface supporting users to converse with the interface through spoken interactions. In this section, we introduce different kinds of conversational interfaces, set out the limitations, how they support the entire interaction from the users' speech input to the system's output, and investigate the latest research in the presentation of results.

A conversational interface, also identified as conversational user interface (CUI), presents the front-end to a chatbot or virtual personal assistant, enabling the user to interact with the application through various input and output modalities such as speech, text, or touch (McTear *et al.*, 2016; McTear, 2017). Besides being the system's front-end, the conversational interface integrates or glues together all the underlying system components, represented in a usable application (Zue and Glass, 2000). Even though all the recent developments of the separate components have made conversational interfaces more functional, they act as the orchestrator of all the information with their challenges.

Overall, this section introduces the different conversational interfaces and illustrates the limitation of transferring information in a

conversational style for different interfaces. We discuss *initiative* as a critical element in conversational interactions, including the interface limitations with regards to CIS.

3.1 Conversational Interfaces

Interfaces that provide users with the ability to interact *conversationally* with systems through different modalities such as speech, gesture, text, or touch are commonly referred to as CUIs. Many additional terms refer to these systems that enable conversational interactions, including chatbots, intelligent assistants, or conversational agents.

An interface is often referred to be *conversational* when it covers two basic attributes (1) natural language and (2) conversational interaction style (McTear, 2017). The *natural language* attribute means that the system and user can use language as in naturally occurring conversations between two or more participants; this contrasts to restricted commands, mouse clicks, or phrases in a graphical user interface (GUI). Furthermore, natural language is more flexible, permitting input to be expressed in many different ways versus one fixed expression. Intuitively, allowing the user to input natural language contributes to a more complex system. In addition, *conversational interaction style* is often referred to as basic turn-taking behavior in which the user and system converse one after another. This contrasts with clicking or swiping on GUI elements such as buttons or drop-down menus. Furthermore, to make an interface even more conversational, the usage of *mixed-initiative* is introduced. Mixed-initiative is more human-like and flexible because both actors can independently contribute to the conversation. Lastly, a more advanced system could include *context* tracking enabling follow-up questions and persistent tracking of the topic. Even though many dialogue systems are seen as conversational, they may not be tracking the context and therefore never refer back to a previous question or answer. Instead, they attend to every input individually.

Basic conversational interfaces often consist of two primary attributes and sub-attributes: natural language which does not consist of fixed expressions, and conversational interaction style which could support turn-taking, mixed-initiative, and context tracing.

Even though various forms of conversational interfaces have been around for a long time, we have recently seen a revival of the topic, mostly due to the advances in automatic speech recognition (ASR), natural language processing (NLP), and machine learning in general. Nevertheless, much fundamental research dates back to the 1960s with the first well-known chatbot, called ELIZA, which simulated a Rogerian psychologist (Weizenbaum, 1966). In the following, we provide some historical context for four distinctive groups of conversational interfaces, (1) spoken dialogue systems (SDSs), (2) voice user interfaces (VUIs), (3) live chat support, and (4) chatbots.

3.1.1 Spoken Dialogue Systems

Spoken dialogue systems (SDSs) enable users to interact with a system in spoken natural language on a turn-by-turn basis and are an instance of a conversational interface. Many of these systems are used for task-oriented issues with clear task boundaries, such as travel planning. In the 1960s and 70s, the earliest SDSs were mainly text-based. However, once technologies improved in the 80s, more complex components were added, such as more advanced ASR or components that helped recover from conversational breakdowns. Much government funding from Europe and the U.S. supported research in SDS, which resulted in the European SUNDIAL (Speech Understanding and DIALog) project (Peckham, 1991) and the DARPA spoken language system in the U.S. (Clark, 1988). The SUNDIAL project aimed to design systems that could be used by the public, while the DARPA program focused on the technical aspects. Many of the early research outcomes are still applicable today, such as the Information State Update Theory (Traum and Larsson, 2003), information presentation techniques (Gibbon *et al.*, 1997), or the

CSLU toolkit (Sutton and Cole, 1997).

Traditionally, the task for SDSs has often been time-tabling for travel services which is a good example for providing the interface between the user and an existing database (Fraser, 1998). In the following example, the user has the need of finding a reasonable travel plan.

USER: I want to travel from Paris to Brussels.

SYSTEM: What time would you like to arrive in Brussels?

USER: At 8pm.

SYSTEM: There are no arrivals at 8pm. The closest other arrival is either 7pm or 10pm. Do you want one of these arrivals?

As seen in the first utterance from the system, it is narrowing down the information need by adding a refinement or clarification question to help the user. These back and forth interactions are part of the elicitation process for the system to understand and specify the information need.

3.1.2 Voice User Interfaces

Companies have traditionally developed VUIs for commercial benefits, which is in contrast with SDS, created mainly by academic and research labs. For example, AT&T created an early VUI called *How May I Help You?* which supported call routing (Gorin *et al.*, 1997). The automated customer self-service systems are task-oriented and engage in conversation to help the client, thus being classified as a conversational interface. Instead of helping the customer with their problem, VUIs typically aim to understand the customer's initial problem sufficiently, after which the user can be routed to the appropriate (human) call taker to help the user with their problem further. Thus, these call routing services only need to elicit the general problem to refer the call to someone or a specific system module. The system responses are pre-recorded, which is possible for highly structured domain-specific settings. For example, a scenario where a user wants to pay for a service might follow a very scripted interaction:

SYSTEM: How may I help you?

USER: I'd like to pay my bill.

SYSTEM: Do you want to charge the credit card we have on file which ends in 0987?

USER: No, an other one.

SYSTEM: Please speak your card number.

USER: 123-456-789

In these systems, when none of the options are relevant to the user, the system will narrow down the problem to re-route the call to an appropriate human agent. The connection with CIS is the human-like interactions, eliciting information needs, and narrowing down the relevant answers or services.

The VUI community has involved with the development of W3C standards for scripting spoken dialogues such as VoiceXML,[1] VoiceXML-based toolkits,[2] and the development for speech analtyics.

3.1.3 Live Chat Support

The above interfaces (*i.e.*, SDS and VUI) are mainly used with an underlying automated system. However, many support systems are powered by humans in which the interface is the connection between a user and service provider. Live chat support is real-time communication between a customer and a support person via instant messaging, often through a pop-up dialogue box. The service providers can include librarians on a library website (see Figure 3.1) (Matteson *et al.*, 2011), technical or sales support on e-commerce websites (Goes *et al.*, 2012), or health assistance (Stephen *et al.*, 2014). Such chat support interfaces are often embedded as a web widget in websites or an extra feature within an application. The main advantage of live chat support interfaces is that the chat history is persistent and can be referred to by the people using the interface. Furthermore, these chats can support asynchronous and synchronous interactions (Fono and Baecker, 2006).

[1] https://www.w3.org/TR/voicexml21/
[2] http://evolution.voxeo.com/

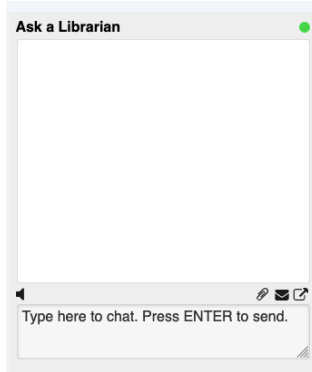

Figure 3.1: Example of a live chat support interface to interact with librarians.

Some recent work by Vakulenko *et al.* (2021) investigated virtual reference interviews of professional librarians. They suggest major differences between librarian interviews and existing datasets used to investigate, analyze, and train CIS topics. For example, they suggested that professional intermediaries are more proactive, write more extended responses, ask follow-up questions, and actively steer the topic of conversation. Further research efforts are needed to understand the impact of different conversational styles of CIS systems (Thomas *et al.*, 2018).

A "live chat support" provider (*e.g.*, the call taker or customer provider) is often synchronous, meaning that the support person answers questions from the user in real-time. However, many support providers are required to answer multiple customers simultaneously, creating a one-to-many relationship. The importance of the support provider's interface, which could support decision making by ranking response suggestions on the information-seeking process or incorporating machine reading to track the conversation, has been understudied (Xu and Lockwood, 2021; Yang *et al.*, 2018b). Furthermore, research on how the support providers deal with task-switching and interruptions could suggest future conversational interface optimisations (Pajukoski, 2018).

3.1.4 Chatbots

The interactions with chatbots are often based on social engagement through chit-chat (*i.e.*, small talk), which is in contrast to the very

task-oriented interactions with SDSs and VUIs. Traditionally, chatbots are mainly text-based. However, more recent chatbots have started incorporating spoken interactions, images, and even avatars for creating a more human-like persona.

Note that a chatbot is different from a bot (McTear, 2017). A chatbot is a software application that can perform automated tasks while engaging in conversations with the user. This contrasts with bots, which complete repetitive and mundane automated tasks such as crawling the web or harvesting email addresses from social networks.

All the above systems aim to support users to interact with datasets or databases. Due to the conversational aspect of the interaction, no technical expertise is required to interact with these databases, making them more accessible. As illustrated with the different CUIs (*i.e.*, SDS, VUI, live chat support, and chatbots), these systems cover a large range of applications and tasks (*e.g.*, from travel booking to chit-chat). Although all these CUIs may be considered conversational, they still differ in the degree that people are searching for information, the system maintains control, and flexibility allowed by the user to ask for what they want to find or how they want to have the information presented. In contrast, searching for information on the web over documents is much less predictable and cannot be implemented by pre-set refinement options. Due to the vast amount of information, more advanced techniques are needed to support users' information needs. Other questions such as the ambiguity in people knowing when they are talking to a human or machine (*e.g.*, chatbot),[3] the trust of people have in these systems, appropriateness of these systems, or transparency around the usage of artificial intelligence in general[4] are relevant (Mori *et al.*, 2012; Zamora, 2017).

3.2 Result Presentation: From Search Boxes to Speech Bubbles

Result presentation in CIS is tightly coupled with decades of research on interface development for search engines and other information retrieval

[3] https://botor.no/

[4] https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

systems. In this section, we draw the connection between conversational user interfaces required in CIS and past research on result presentation in search engines.

Results presentation, the way search results are communicated, has been a major research area for many years. The general approach to present search results is as a vertical list of information summarising the retrieved documents. These results should not only return relevant results but also display them so that users can recognize them as relevant to their information need.

Even though many people have become accustomed to searching through these search boxes, finding information can still be a demanding task with much information to filter through. Traditionally, a user would submit an information need through keywords in a search engine search box. In return, search engines present a ranked list with potential relevant documents for that query, also referred to the search engine result page (SERP). This SERP consists of the traditional "ten blue links" in which each item or result consists of a document title, a short summary (*i.e.*, snippet), URL, and often other meta-data such as date or author (see Figure 3.2) (Hearst, 2009; Paek *et al.*, 2004).

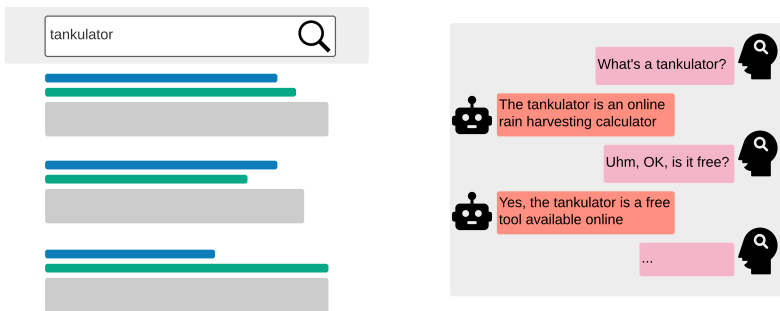

Figure 3.2: Traditional SERP example versus a conversational style interaction

The user would then review this returned ranked list and select an item they think would satisfy their information need. However, the first clicked item will often not satisfy the users' information need. Instead, the user will go back and forth between inspecting SERPs, looking at the contents of documents and submitting new queries. These interactions

mimic a limited or one-sided conversation driven by the user. In this instance, the user has "control" over the actions taken and the system has limited capabilities to interact with the user. These systems are sometimes referred to as *passive* (Avula, 2020; Trippas *et al.*, 2018).

The alternative interaction paradigm of CIS aims to overcome the limitations of the results presentation strategies of existing search engines by becoming more active. That is, instead of presenting a ranked list, these CIS systems can be more flexible with their information presentation strategies by adapting to the user's needs.

Even though research has shown that different presentation techniques and answer organization are needed for different modalities, limited research has been conducted in *how* (the content expression) and *what* (the content response) to present in conversational search (Chuklin *et al.*, 2018; Trippas *et al.*, 2015b; Vtyurina *et al.*, 2020). Furthermore, not only the retrieved information needs to be presented but depending on the modality of the results presentation, other interactions such as meta-conversations, need to be presented (Kiesel *et al.*, 2021a; Trippas *et al.*, 2018).

Overall, it is well accepted that people search differently depending on the device (*e.g.*, desktop versus mobile) and modality (*e.g.*, text versus audio). Differences are highlighted in the following subsections.

3.2.1 Text-Only Result Presentation on Desktops

Much research has been conducted on the appearance of SERPs in browsers (Hearst, 2009). In a visual setting, researchers have investigated features such as snippet length (Cutrell and Guan, 2007; Kaisser *et al.*, 2008; Maxwell *et al.*, 2017; Rose *et al.*, 2007), snippet attractiveness (Clarke *et al.*, 2007; He *et al.*, 2012), or the use of thumbnails (Teevan *et al.*, 2009; Woodruff *et al.*, 2002).

Research on results presentation has suggested that the presentation has an impact on the usability of the system. For instance, Clarke *et al.* (2007) investigated the influence of SERP features, such as the title, snippets, and URLs on user behavior. They suggested that missing or short snippets, missing query terms in the snippets, and complex URLs negatively impacted click-through behavior. In addition, Cutrell and

Guan (2007) used an eye-tracking study to explore the effects of changes in the presented search results. They manipulated the snippet length with three different lengths (short [1 text line], medium [2-3 lines], and long snippets [6-7 lines]) as shown in Figure 3.3. Their results suggested that depending on the search task (*i.e.*, navigational or informational), the performance improved with changing the length of the snippet. That is, the performance improved for informational queries but degraded for navigational queries.

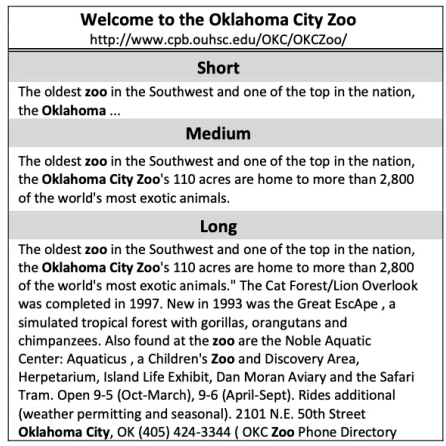

Figure 3.3: Snippet length differences (Cutrell and Guan, 2007).

Further research into snippet summary length confirmed the findings that different snippet lengths were preferred depending on the task (Kaisser *et al.*, 2008). A more recent study by Maxwell *et al.* (2017), re-investigated the varying snippet length and the information content within the snippets. Their results suggest that users preferred more informative and extended summaries, which they perceived as more informative. However, even though participants felt that longer snippets were more informative, they did not always help the users to identify relevant documents.

Techniques in which visual changes to text are made, such as clustering, highlighting, or "bolding" query words in their context, sentence fragments, or query-biased summaries have been extensively investigated for traditional results presentation. Furthermore, besides only

showing text in the SERP, search engine companies have added more techniques to display results through feature snippets, knowledge cards, query suggestions, or knowledge panels. More research for these presentation styles in CIS is needed to understand the impact of these techniques in a conversational setting.

Limited research has been conducted into conversational results presentation for desktop. A recent prototype for text-only chat-based search by Kaushik *et al.* (2020) combined a conversational search assistant with a more traditional search interface, see Figure 3.4. The user can either interact with the agent itself on the left side of the application through the Adapt Search Bot or with the retrieved information on the right panel, which is called the Information Box. The authors described this design as flexible for users to interact with the agent and the search engine itself. Furthermore, their design supported users interacting with the search engine with the agent initiating dialogues to support the search process. However, further research should be done to understand the impact of different presentation techniques, chat-based search, and distributed results presentation (*e.g.*, results on both left and right panel).

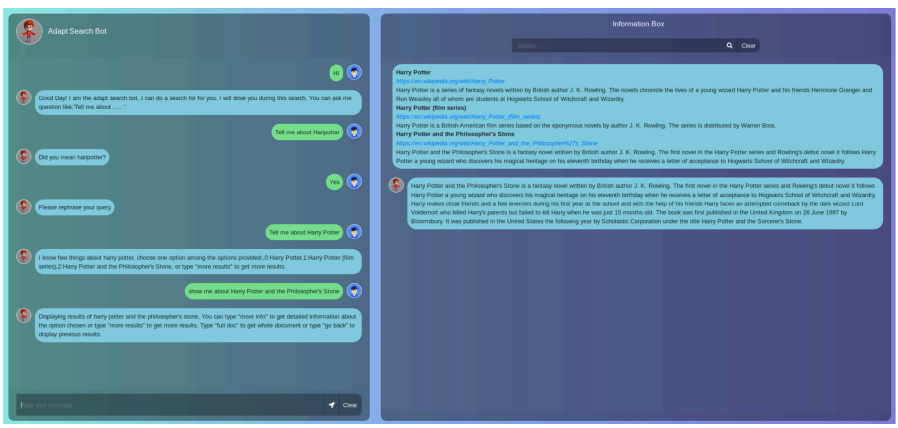

Figure 3.4: A visual example of the Conversational Agent by Kaushik *et al.* (2020). The agent exist of a conversational search assistant (left) with a more traditional search interface (right).

Another alternative for searching through conversational interactions

on a desktop was presented by embedding a *searchbot* directly into an existing messaging platform (*i.e.*, Slack) by Avula *et al.* (2018). The searchbot interfered in a collaborative setting by injecting information relevant to the conversation between the two users. An example of a searchbot results page within Slack is presented in Figure 3.5. As seen in the figure, the results were always followed by a "click here for more" option, redirecting the users to a different SERP. The results of this study suggest that dynamically injected information can enhance users' collaborative experience. Further research into the presentation of the results in such a collaborative CIS setting is needed to enhance our understanding of optimizing this search experience.

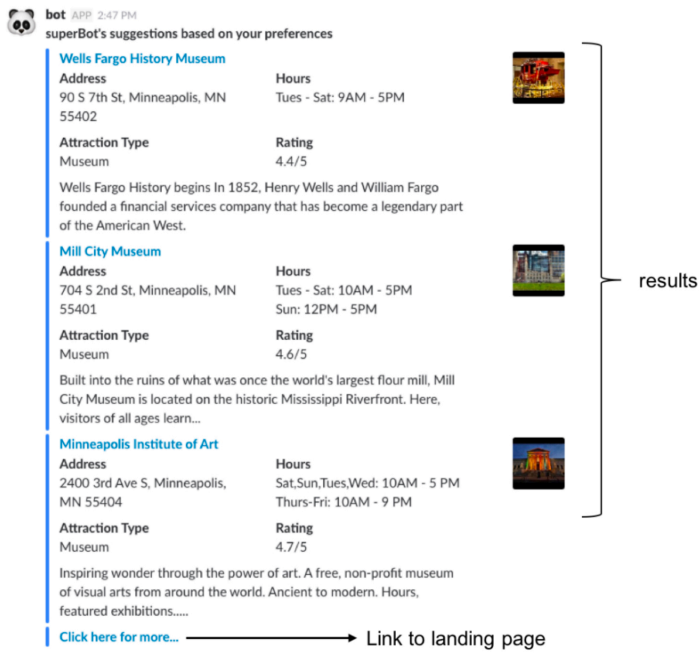

Figure 3.5: A *searchbot* results presentation example inside Slack on a desktop (Avula *et al.*, 2018).

3.2.2 Text-Only Result Presentation on Small Screens

People interact differently when searching for information on a mobile or desktop device (Jones *et al.*, 1999; Church and Oliver, 2011; Ong *et al.*, 2017). Researchers have suggested that with the shift to mobile search, there has also been a paradigm shift in web search (Ong *et al.*, 2017). Differences in screen size and being able to access search engines in different contexts or "on-the-go" have impacted how we search.

With the increasing use of mobile devices such as smartphones, researchers have also investigated the results presentation on different screen sizes (Ong *et al.*, 2017; Kim *et al.*, 2015). Because of the smaller screen sizes on mobile devices, it is important to investigate the result presentation and optimize for the screen real-estate. For example, an average-sized snippet for a desktop site may not be acceptable for a smaller screen since it may involve more scrolling and swiping.

Kim *et al.* (2017) studied different snippet lengths on mobile devices. An example of varying snippet length on a small screen is presented in Figure 3.6. They demonstrated that participants who were using more extended snippets took longer to search because it took them longer to read the snippets. They suggested that unlike previous work on the effect of snippet length, the extended snippets did not seem that useful for mobile devices and that snippets of two to three lines were most appropriate. Furthermore, it has been suggested that short snippets may provide too little information about the underlying document, which can have an adverse effect on the search performance (Sachse, 2019). In general, depending on the information need, different snippet lengths could be used to optimize the user experience.

Even though results presentation has not been fully explored in a CIS context, they can be developed and deployed on already installed mobile messaging applications such as Telegram (see Figure 3.7) (Zamani and Craswell, 2020). This means that people are already familiar with the application and can often be deployed and accessed over multiple devices and platforms. Furthermore, embedding these CIS systems within existing messaging applications means that the user does not need to download and install new apps for every service.

However, limited research on how users could interact with in-

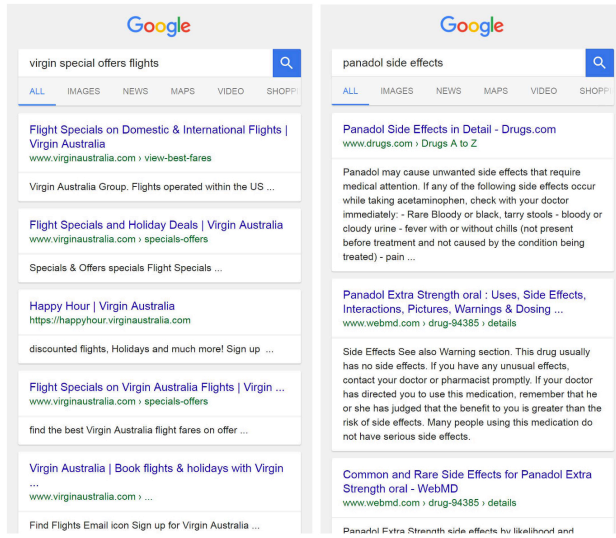

Figure 3.6: Examples of SERPs with short (left) and long (right) snippets by Kim *et al.* (2017).

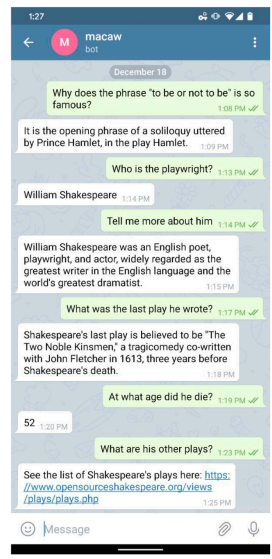

Figure 3.7: Example screenshot of results presentations with Macaw (Zamani and Craswell, 2020) using Telegram.

formation through these messaging applications has been conducted. Questions such as what medium to use, enabling users to click on messages to respond or read more information are still open.

3.2.3 Speech-Only Result Presentation

Result presentation research has traditionally been focused on visual representation. However, with the ongoing trend of CIS and improvement of speech recognition, researchers have started investigating how to present results in a speech-only setting.[5] It has been suggested that using speech to search is a natural extension of the visual search engines, potentially changing the way we access information (Trippas, 2019). However, several researchers have also suggested that simply translating a SERP from a visual to a speech setting is not desirable (Lai *et al.*, 2009; Trippas, 2019; Vtyurina *et al.*, 2020). Just translating text results into audio impacts the user experience negatively and requires higher cognition (Vtyurina *et al.*, 2020). Thus, it has been suggested to steer away from the "ten blue link" paradigm and instead re-think the interactions with search systems. Furthermore, due to the temporal nature of speech, results can be adapted on the fly to change the presentation, thus supporting the user in their information need more actively.

Similarly, as in traditional web search versus searching on smaller screens, it has been suggested that snippet length should be altered depending on the information need. In a study by Trippas *et al.* (2015b), the preference for summary length was investigated with a crowdsourcing setup. More specifically, they studied the summary length by comparing user preference between text and speech-only results. They observed that users preferred longer, more informative summaries in a text setting, however, this was not observed for audio summaries. Furthermore, different results were observed depending on the query style (single- or multi-faceted). More concretely, users preferred shorter audios for single-faceted queries, however, for more ambiguous queries, this preference

[5] We use speech-only which is the structural act or mechanism to speak. However, some of the studies described use *audio* as a sound or *voice* which is a sound produced by a person's voice-box.

was not clear.

More recent work by Vtyurina *et al.* (2020) also compared results presented over text versus speech. They used a mixed-methods study with a crowdsourcing and laboratory component, suggesting that user preferences are different depending on the presentation mode (text or speech). However, they also suggest that users can still identify relevant results even if presented in a more cognitively demanding speech format. The authors suggested that further improvements to the snippets can help optimize and guide the use of speech-based search interfaces. As part of this study, the authors provided the following presentation guidelines for speech-only results presentation:

- Use prosody to avoid monotone voice
- Avoid abbreviations in the spoken results
- Avoid truncation of sentences
- Avoid repetitive terms in spoken results

Research on using prosody for results presentation was conducted by Chuklin *et al.* (2018) and Chuklin *et al.* (2019). They investigated audio manipulation as an alternative to "highlighting" or "bolding", which is frequently done in a visual interface. They used a crowdsourcing study by modifying speech prosodies such as pitch, pauses, and speech rate in readout snippets. Some emphasis features help users identify relevant documents more often. They also increase the snippet informativeness.

Many open problems related to the support and guiding of searchers through results presentation exist. For example, presentation order bias (Azzopardi, 2021; Kiesel *et al.*, 2021b), interaction with tabular data (Zhang *et al.*, 2020a), personas of the conversational system (Nass and Brave, 2005), persuasiveness of synthetic speech (Dubiel *et al.*, 2020b), meta-communication to support communication breakdowns (Trippas *et al.*, 2018), or using non-speech sounds to increase user engagement with search results (Winters *et al.*, 2019; Arons, 1997). For example, order bias has been suggested to affect which results summaries receive the most attention from users in a visual setting (Joachims *et al.*,

2005). Work has suggested a possible bias towards first and last readout search results depending on the kinds of information need, single- versus multi-faceted (Trippas *et al.*, 2015a). This example of a *serial-position effect* (*i.e.*, the tendency to recall the first and last items best and the middle items worst) are open problems.

3.2.4 Multi-Modal Results Presentation

Research on CIS primarily focuses on uni-modal interactions and information items. That is, all information is either exchanged in text or speech-only format. However, more recently, researchers have started investigating in more detail the advantages of multi-modal CIS (MMCIS), in which multiple input and output devices are used (Deldjoo *et al.*, 2021; Liao *et al.*, 2021). Presenting search engine results over a multi-modal channel aims to increase the knowledge transfer of different modalities, enhancing the search experience (Schaffer and Reithinger, 2019).

A multi-modal interface can process two or more user input modes, for instance, speech, images, gestures, or touch (Furht, 2008). Multimodal systems try to recognize human language, expressions, or behaviors which then can be translated with a recognition-based system. These multi-modal interfaces are often seen as a paradigm shift away from the conventional graphical interface (Oviatt and Cohen, 2015). Similar to a multi-modal dialogue system, a MMCIS system aims to provide completeness to the unimodal counterpart by providing information through multiple modalities (Firdaus *et al.*, 2021). Furthermore, the theoretical advantage of these different inputs is that they are very close to human expression and thus are an efficient way of human-computer interaction. Thus, multi-modal interfaces enable humans to input signals to machines naturally through a mixture of interactions to convey the intended meaning (Rudnicky, 2005) and it is often suggested that multi-modality increases the intuitiveness of an interface.

By coupling the intuitiveness of conversations with human conversation, which is inherently multi-modal, the strengths of human communication can be combined, enabling a natural form of information seeking. In addition to the system trying to elicit information from

the user to satisfy the information need and perform queries in the background, the system also needs to decide *which*, *what*, *how*, and *when* to present information.

Rousseau *et al.* (2006) created a conceptual model, called WWHT, describing four main concepts of multi-modal information presentation, based on four concepts "What", "Which", "How", and "Then":

- **What** is the information to present?
- **Which** modality(ies) should we use to present this information?
- **How** to present the information using this(ese) modality(ies)?
- and **Then**, how to handle the evolution of the resulting presentation?

When designing multi-modal CIS interactions, a fundamental problem is the option, combination, or sequence of different outputs of "displaying" results. For example, it is logical that relying only on a speech-only result presentation in a loud environment will be undesirable. Instead, using a combination of modalities to present the results in such an environment may be advantageous. Furthermore, as identified and demonstrated by Deldjoo *et al.* (2021), MMCIS, and therefore the information presentation problem, is suitable in the following conditions:

- the person who is searching has **device(s)** available which allow for more than one interaction mode (multi-device and multi-modal),
- when the task's **context** is important and can be captured with a device in a suitable modality enhancing personalization,
- when task **complexity** can be supported by the mode of device interaction,
- when the results can be returned in an appropriate **output modality** given the device, context, and complexity.

Many open challenges for CIS results presentation in a multi-modal domain exist. Problems include selecting the optimal output modality

depending on the context or the user's ability, adapting or changing the output modality to be different from the retrieved modality, or fusing the response to present the results in multiple modalities (Deldjoo *et al.*, 2021).

3.3 Initiative in Conversational Systems

The demand to access information rapidly in a natural way has substantially increased due to the proliferation of reliable mobile internet, mobile devices, and conversational systems. Humans create and collect more information than ever before[6] through blog posts, social media, emails, news articles, or videos while using them for education, entertainment, finance decisions, or other decision making (Zue and Glass, 2000). In addition, querying this information has become omnipresent with an estimated 75,000 Google searches per second in 2019.[7] DuckDuckGo, a privacy-focused search engine with an estimated market share of 0.18% of global searches, received 23.65 billion search queries in 2020,[8] illustrating the scale of search in our daily life.

Furthermore, with the rise of smartphones and mobile internet, we have been accustomed to access this information on the go and while multitasking.[9] However, accessing information through a small screen and on-screen keyboard while travelling can be cumbersome. Therefore, conversational interfaces in which natural language can be used to interact with information have great promise. Indeed, spoken human language is attractive since it is the most intuitive way of conversation. Furthermore, it is often seen as a very efficient, flexible, and inexpensive means of communication (Zue and Glass, 2000; Trippas, 2019). In addition to human language, additional support input can be given through gestures as part of the multi-modal input (see Section 3.2.4).

Independent of the kind of conversational interface, these interfaces

[6] https://www.forbes.com/sites/nicolemartin1/2019/08/07/how-much-data-is-collected-every-minute-of-the-day/

[7] https://www.domo.com/learn/infographic/data-never-sleeps-7

[8] https://www.theverge.com/2018/10/12/17967224/duckduckgo-daily-searches-privacy-30-million-2018

[9] https://www.thinkwithgoogle.com/intl/en-aunz/marketing-strategies/app-and-mobile/device-use-marketer-tips/

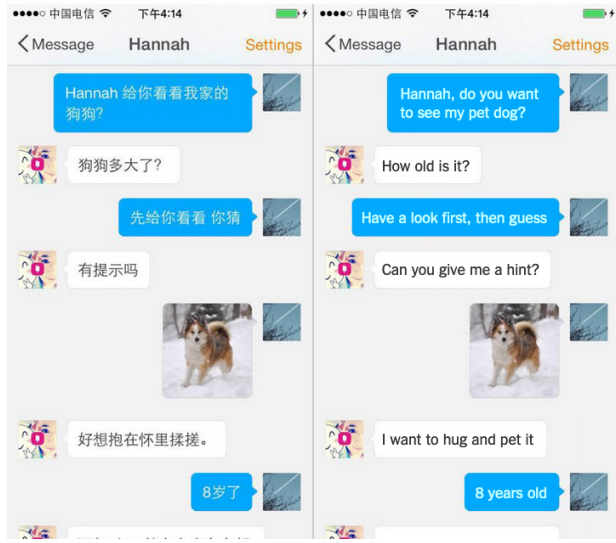

Figure 3.8: Example chat with XiaoIce (Zhou *et al.*, 2020).

are often considered from the perspective of *initiative*. That is, to which degree does the system maintain an active role in the conversation (Zue and Glass, 2000; McTear *et al.*, 2016; McTear, 2017). Three different levels are often used to distinguish these, *i.e.*, system-initiative, mixed-initiative, and user-initiative and are often used interchangeably with levels of control *system*, *user*, *interchangeable*. With *system-initiative* applications, or system directed dialogue, the computer takes control over the sequences in the conversation and which information needs to be exchanged. The aim of the system is to elicit information from the user in order to provide relevant details back to the user. This can be done through asking open-ended questions such as in utterance U1 in the below example, in which the system invites the user to provide information and then elicits further details (utterance U3):

U1 - System: OK, tell me what you are looking for.

U2 - User: I want to find information about the health benefits of honey.

U3 - System: Would you like to know more about the medical health

benefits of honey or skin benefits?

U4 - User: Medical benefits.

U5 - System: I have some information from WebMD: "Claims for honey's healthfulness are unproven – except in the area of wound care and, to a lesser extent, cough suppression."

U6 - User: Right, so does that mean there are benefits for the skin?

As seen in the example, in *system-initiative* dialogues, the system takes the initiative to drive the conversation and the user only answers the system's queries. This strategy aims to constrain the user input or request variety, thus making the dialogues more efficient. However, this comes at a cost, with rigid and restricted conversations making the interactions less natural.

In utterance U6, the user takes control over the dialogue by asking a question, turning the conversational interaction into a *mixed-initiative* dialogue. Hence, both user and system now actively participate in addressing the information need through the interactive conversational paradigm. Thus, *mixed-initiative* dialogues are known for a more natural exchange, however, more advanced ASR and language understanding is needed.

Lastly, *user-initiated*, or user-directed dialogues, are conversations in which the user has complete control and can say anything to the system and the user always has the initiative. This means that the system will only respond to the user's requests. The disadvantage of this approach is that the user may find it challenging to understand the system's capabilities because the system will never suggest anything. Furthermore, dialogues with *user-initiative* may lead to frustration from the user because the system is not a conversational *partner* but rather only replies to requests.

3.4 Interface Limitations in Conversational Systems

Even though conversational systems can have many advantages, such as enabling users or supporting natural language input, expression of

multiple information needs in one turn, cross-platform compatibility and integration, and increasing engagement through personalization, many limitations need to be addressed.

For example, natural language input components, such as ASR and NLU, need to be optimized to handle the huge number of unknown and unexpected user inputs. Furthermore, conversational systems need to be optimized to handle *non-explicit* information needs. For example, a users' tone of voice may imply that they want the conversational partner to do something, even though that need was not explicitly stated. Current CIS systems work reasonably well with narrow or factoid queries, however, they still have issues when the information need is more complex (*e.g.*, multi-faceted) or has multiple information needs in one turn.

Besides the limitation of results presentation or output from the system discussed in Section 3.2, such as highlighting or bolding keywords, other more general limitations need to be considered. For example, GUIs should be carefully investigated before being directly translated into conversational or voice user interfaces. Even though many chatbots support menu-based interactions within the application, using buttons or menus will limit the benefits of natural language input. Furthermore, issues that already exist in GUIs are now passed on to conversational systems. Instead, conversational systems now inherit the GUI experience devaluing the natural language advantage.

In addition to these existing output difficulties, speech-only conversational systems come with their distinct challenges. For example, simply reading out textual components or reading out lists has shown to be ineffective (Trippas *et al.*, 2015b; Vtyurina *et al.*, 2020; Gooda Sahib *et al.*, 2015). Furthermore, the serial and transient nature of audio can challenge the users' ability to recall all information presented (Dubiel *et al.*, 2020a). This exacerbates the difficulty of skimming audio and makes it challenging to present results while not overwhelming the user with information nor leaving them uncertain as to whether they have covered the information space (Trippas *et al.*, 2015b; Trippas, 2019). These CIS systems cannot maintain a lengthy information exchange or keep sufficient track of the context. In addition, images and graphs are more challenging to be displayed and may need to be interpreted by the system to inform the user what is displayed (Trippas *et al.*, 2019).

Other limitations, such as the tone of voice or persona of the system interacting with existing stereotypes or biases of humans speaking in particular ways may plausibly both reinforce existing biases as well as cause systems to be perceived in particular ways Nag and Yalçın, 2020

Considerations must also be made for limitations automatic speech recognition (ASR). For example, users' speech-input may include disfluencies or errors. Users may mispronounce words, use filler words such as "uhm" or "ah", or add extra pauses. Furthermore, different speech variabilities such as patterns, dialects, age, gender, or speech impairments can impact ASR performance. For example, speaking faster or slower can have an impact the acoustic models used for transcriptions (Benzeghiba *et al.*, 2007). Indeed, an apparent challenge for conversational systems is the barrier to recognize speech from a diverse population (Zue and Glass, 2000). To make information more accessible and enable wide adaptation of conversational systems, including by people with cognitive or physical impairments is needed (Baldauf *et al.*, 2018; Derboven *et al.*, 2014). Beyond this, there has been very limited published work on design of speech-only systems to consider users who either hard of hearing or vision impaired.

General issues related to conversational interfaces include the lack of transparency on which capabilities a system has (*i.e.*, discoverability), personalization in the context of conversations, and privacy issues related to talking out-loud in public spaces.

3.5 Summary

This section covered conversational interfaces, results presentation, different kinds of initiative in conversational systems, and interface limitations. We explained some of the essential building blocks for conversational interfaces and presented the central features such as natural language, conversational style, mixed-initiative, and context tracking. This section also provided a historical context on existing conversational interfaces (*i.e.*, SDS, VUIs, live chat support, and chatbots) and their application. We highlighted the differences between these various types of conversational interfaces with respect to CIS. Since information presentation is key to transferring the found information to the users

for CIS, we provided an overview of information presentation research. We described text-only presentations on desktops and small screens, speech-only presentations, and multi-modal presentations. In each part, we highlighted the limited research on *how* and *what* to present with interactions in CIS systems. Lastly, we explained key elements regarding initiative in conversations and how different initiatives create different conversational styles, followed by a brief summary of limitations in conversational interfaces.

To conclude, different interfaces can support people in their information seeking journey differently. Thus, CIS systems can benefit from different interfaces, and models and evaluation methodologies for handling multiple interfaces will be needed. Understanding the impact of interactivity in CIS is crucial to make user-friendly systems. Different ways of understanding these conversational interactions are discussed in the next section.

4

Understanding Conversational Interactions

A distinctive property of CIS systems is the ability of the systems to understand multi-turn interactions expressed in the form of natural language. This section discusses the varying approaches to conversational understanding required to follow the conversational structure, with a focus on information seeking conversations.

We given an example of a possible CIS conversation taken from the TREC Conversational Assistance Track (CAsT) Year 3 (Dalton *et al.*, 2020c) – topic number 113. This illustrates coreference issues, dependence on previous result state, and explicit feedback with topic switching behavior.

USER: How do genes work?

SYSTEM: A gene is a short piece of DNA. Genes tell the body how to build specific proteins. There are about 20,000 genes in each cell of the human body. Together, they make up the blueprint for the human body and how it works. A person's genetic makeup is called a genotype... Many personal characteristics, such as height, are determined by more than 1 gene. However, some diseases, such as

sickle cell anemia, can be caused by a change in a single gene.

USER: What others are caused by a single change?

SYSTEM: The four types of (human) Genetic diseases are: (1) Single-gene/monogenic Genetic Diseases: In this category the starting point is a mutation/change in one gene... Almost 6000 single gene disorders are known and it is estimated that 1 of 200 newborns face a single gene genetic disorder...

USER: What are the other types of diseases?

SYSTEM: (3) Chromosomal Genetic Diseases: Chromosomes are big DNA molecules composed from genes... Abnormalities in the structure, number (and not only) of the chromosomes can cause some of the most dangerous genetic disorders...

USER: You missed the second type. What was that?

SYSTEM: A large genetic mistake typically occurs in the woman's egg, which may partially explain why older women are more likely to have babies with Down syndrome...

USER: That's not what I wanted. How about recent developments in gene therapy to treat those defects?

SYSTEM: ...

As illustrated above, CIS systems must "understand" a natural language utterance and represent the state of an information dialogue as it evolves. Second, they need to track the evolution of that state across turns and relate them to one another including detecting feedback and sub-topic switches. The fundamental mechanisms and representations for conversation understanding vary widely. The architecture may have an *implicit* internal representation based upon the model's input (*e.g.*, attention-based models that encode the conversation history). The

model may also have an *explicit* representation of the salient state that could include representations that are structured (entity or schema), lexical (word-based), or vector-based.

This section begins by discussing the representation used within a single turn. It then discusses conversation modeling and history tracking for information and how this differs from previous work in dialogue state tracking. Next, it covers models of conversational discourse and discusses approaches that identify turn salience, extract contextual keywords, and vector representations. It provides an overview of core conversational tracking sub-tasks including: (1) entity recognition, (2) query expansion, (3) salient term selection, and (4) conversational query rewriting (CQR). It concludes with a discussion of how these approaches continue to evolve beyond short conversations towards longer and multi-session conversations.

4.1 Modeling within Turn State

This subsection introduces the building block for multi-turn conversations — the representation of the state for a single turn. Because CIS systems usually operate in an open-domain environment, they often do not use predefined domain state (frame) ontologies. At its most basic level, the state representation includes the utterance *text*; whether it is typed or from automatic voice transcription. It may also include *explicit* structured annotations of the utterances from the output of an NLP system: part of speech information, dependency parse, semantic frame parsing (*e.g.*, FrameNet), entity recognition and linking, semantic parsing to a logical representation, and others. The state representation for a single turn in CIS is related to work on natural language query understanding and question answering (Huang *et al.*, 2018; Raffel *et al.*, 2020). These sub-tasks evaluate understanding at query/turn level using standard text evaluation measures.

A widely adopted approach to conversational representation uses pre-trained language models with contextualized embeddings, particularly Transformer models (Vaswani *et al.*, 2017; Raffel *et al.*, 2020). These exhibit transfer learning capability that allow them to be fine-tuned for one or more conversational ranking or question answering (QA) tasks.

For conversations, utterances may be encoded separately, compared, and possibly combined in a dense embedding space (Khattab *et al.*, 2020; Xiong *et al.*, 2021; Prakash *et al.*, 2021).

Pre-trained language models demonstrate general language understanding behavior commonly found in NLP pipelines including coreference resolution, entity recognition, and relations (Tenney *et al.*, 2019). As a result, systems in many leading CIS benchmarks (*e.g.*, CAsT (Dalton *et al.*, 2019) and QuAC (Choi *et al.*, 2018)) do not explicitly use the output from an NLP system, but instead rely on the models to handle these tasks implicitly. The result is that for CIS, the system's ability to track the cumulative explicit state representation is usually not measured because the state is not an explicit structure. Most CIS datasets do not contain traditionally labeled NLP annotations with ground-truth salient state. Instead, the focus is on CIS-specific subtasks such as query rewriting or classifying turn salience. The primary evaluation of this effectiveness is often extrinsic in the overall end-to-end retrieval or QA system.

The key differentiating element for CIS compared with single-turn information seeking is the type of interaction and discourse structure. There are various proposed models of a conversation in the literature. Structure in a conversation builds on actions of the participants, namely the speech or dialogue acts. A common task is 'dialogue act recognition' to label the utterances with the type of interaction (Bunt *et al.*, 2017) (*e.g.*, INFORM, REQUEST, GREETING) that encodes how the current turn relates to previous ones explicitly. The definition of these act types and their usage varies widely.

One model developed more specifically for CIS systems by Azzopardi *et al.* (2018) presents a model of conversational search evolution and includes a taxonomy of the user and system action space. A conversational information need evolves with a single turn being the Current Information Need (CIN), past turns with results as Past Information Needs (PINs), and an agent's model of the information space including a model of varying trajectories with Alternative Information Needs (AINs). The action space includes rich types of both user and system revealment of varying forms. The work of Ren *et al.* (2021) refine this with a focus on conversational interaction with existing search engines,

including explicit user intents (such as reveal, revise, chit-chat) and system actions (suggest clarifications, show results, chit-chat, *etc*).

Many of the current benchmark datasets have simplistic discourse with the user asking questions and the system returning answers of varying types. For example, the widely used QuAC (Reddy *et al.*, 2019) conversational QA dataset contains three categories of dialogue act annotations for each turn, (1) continuation (follow up, maybe follow up, or don't follow up), (2) affirmation (yes, no, or neither), and (3) answerability (answerable or no answer). Later developments to tackle challenges in this area include richer types of revealment, feedback, and others (Dalton *et al.*, 2020c).

The action space and intents vary widely according to the task and interface constraints. What makes CIS distinctive is the unique focus on satisfying a user's information need that may encompass short answer, long answer, and other rich types of interactions.

4.2 Modeling Conversation History and Tracking State

Understanding a conversation is primarily concerned with organizing how a series of turns relate to one another. The relationships in CIS differ from previous work in search systems in that they often exhibit natural language phenomena that span turns - *coreference* (two or more expressions referring to the same thing) and *ellipsis* (omitting words or topics implied by the context). It also requires handling informal language use and implicature. Dalton *et al.* (2020a) looked at the use of coreference of varying kinds - anaphora, zero-anaphora (omission), and others. Compared with traditional NLP corpora (such as OntoNotes and CoNLL coreference) conversation information seeking has a higher rate of ellipsis and zero-anaphora, which are extremely rare in narrative text.

Informal conversational phenomena also include interpreting indirect answers in context (Louis *et al.*, 2020). An example is: "Would you like to get some dinner together?" with a reply, "I'd like to try the new

Sushi place.", which is an implicit affirmative that indirectly implies an answer. For voice-based applications, they must also handle noise because of disfluency removal and the noisy channel from speech-to-text transcription (Hassan Awadallah *et al.*, 2015).

The use of an explicit structured state is widely adopted by task-oriented dialogue systems. Frame-based approaches model the dialogue state with structured domain-specific schemas that have intents (actions) and typed slots with values. Keeping track of this evolving state is a standard task, Dialogue State Tracking (DST), with long-running benchmarks in the Dialogue State Technology Challenge (DSTC) (Williams *et al.*, 2016). These systems usually support a fixed number of pre-defined domains with schemas; the widely used MultiWoz dataset (Budzianowski *et al.*, 2018) has an ontology with twenty-five slots spanning seven domains. The largest, the Schema Guide Dialogue dataset (Rastogi *et al.*, 2020) contains sixteen domains with an average of five intents per domain and 214 slots (2.5 per intent on average). In contrast, CIS systems typically do not have pre-defined domains, intents, or slot representations.

In contrast to task-oriented dialogue systems, CIS systems typically do not have pre-defined domains, intents, or slot representations.

4.3 Modeling Conversation Discourse

There are varying approaches to modeling the evolution of a conversational topic across turns. These leverage the natural language phenomena used in conversational dialogue. Automatic approaches look for topic shifts based on changes in coreferent mentions, shared noun phrases, and common patterns (Mele *et al.*, 2020). The realism of the conversational discourse varies widely among the conversational corpora based upon their creation methodology. The TREC CAsT topics are inspired by informational sessions in web search (Dalton *et al.*, 2020a), but are also engineered to be challenging for trivial reformula-

tion systems. Other datasets such as SQuAD and CoQA are derived from artificially created information needs. The widely used QuAC dataset is limited to discussing an information need about a single person because the crowdsourcing setup does not assume pre-existing background knowledge, resulting in a bias towards entertainment (Choi *et al.*, 2018). The result is that the discourse of conversations varies based upon the type of information, the topic being discussed, broader user task, and the modalities supported for interaction. Note that most of the aforementioned datasets, including TREC CAsT (2019), assume that the sequence of questions are fixed and are independent of the system's response, which is different from real interactions. Furthermore, they assume that the only action the system can take is answering the questions and thus do not support mixed-initiative interactions. This is changing, with increased result dependence in CAsT 2021 and mixed-initiative sub-tasks in 2022 (Dalton *et al.*, 2020c). It represents part of the larger trend towards greater dependence on previous system responses as well as richer types of system responses.

In the following subsections, we discuss the evolution of approaches to modeling conversational history including how it is represented. We then break down history understanding sub-tasks and discuss each. Finally, we conclude by looking towards the evolution of conversations to longer, more complex tasks and information needs.

4.3.1 History Models

Modeling the conversation history requires determining relevant parts of the history, how the history is encoded, and how the encoding is leveraged by the model. Approaches to modeling state across turns vary.

A simple and widely used approach to modeling history is the simple heuristic to concatenate the *last-k* turns. The approaches vary in length and type of context appended. One example of this approach to modeling history uses approximately the previous two turns (Ohsugi *et al.*, 2019) – the previous user utterances, system responses, or one of those two. For conversational QA, a popular approach is to only append the previous answer as context (Choi *et al.*, 2018). Similar heuristics that append the first turn and previous turn(s) of a conversation were

also used in the first year of TREC CAsT (Dalton *et al.*, 2019).

The most important feature in modeling history is the positional relationship between turns to capture common patterns of conversational discourse. In particular, in current datasets, most references refer to immediate or short-term contexts (Chiang *et al.*, 2020). Qu *et al.* (2019a) append encodings of the history but do not model position explicitly. Multiple threads of work improved on this by adding the relative position of previous answers (Qu *et al.*, 2019b; Chen *et al.*, 2020a). Beyond position, Qu *et al.* (2019b) adds a History Attention Module that takes the encoded representations of sequences or tokens and learns the importance of the representations to the current answer. Recent analysis shows that models appear to be relying heavily on positional understanding more than on textual semantic relationships (Chiang *et al.*, 2020).

A challenge for history models is that many of the existing benchmarks only exhibit simple discourse with strong local positional bias. Many also lack explicit discourse annotation. As shown in CAsT, most dependencies are local, on directly preceding turns (Dalton *et al.*, 2020b). This is evolving as CIS systems become more capable, with non-local dependencies increasing from 12% of the turns in CAsT 2020 to 22% in CAsT 2021.

Improved models of conversation history is an area for future work, particularly for long and complex conversations where appending short-term history does not adequately model the discourse. Behavioral analyses of existing models show that they rely heavily on short-term distance cues rather than deeper historical understanding.

4.3.2 History Representation

As mentioned above, a simple approach to model representation is to concatenate relevant turns to history in the order they appear. This creates an *explicit* text-based representation for downstream tasks including query expansion and rewriting.

This may also be performed implicitly through the latent state from a sequence model. Recurrent networks (such as LSTMs (Hochreiter and Schmidhuber, 1997)) encode long-term conversation history dependencies via latent hidden states (Yang *et al.*, 2017). More recent neural language models based on Transformer architectures (Vaswani *et al.*, 2017), *e.g.*, BERT (Devlin *et al.*, 2019), use attention to latently encode relationships. A key consideration is how to encode turn structure (for example using separators) to indicate boundaries between previous user and system responses (Reddy *et al.*, 2019; Qu *et al.*, 2020). This may also be done in the model as in Qu *et al.* (2019a) using a separate embedding indicator to determine if an utterance is a part of a question (user) or answer (system). Chiang *et al.* (2020) use a special word embedding indicator if a token is used in a previous answer.

A separate vein of research creates an *explicit* history model with a mechanism to integrate the representation. The FlowQA approach (Huang *et al.*, 2019) introduces the concept of *Flow* which generates an explicit latent representation of previous context. Modeling the conversation was subsequently evolved in the context of graph networks to model the flow of information as a graph using recurrent neural networks (Chen *et al.*, 2020a). Yeh and Chen (2019) extended this to a Transformer architecture and makes the context flow dependence explicit.

Following recent trends in retrieval approaches, the adoption of approximate nearest neighbor search applied to learned dense representations, also known as dense retrieval, and/or sparse representations resulted in a significant shift. In these representations the query and history is combined into one or more vectors issued as queries to a dense retrieval system. Yu *et al.* (2021) encoded the history representation with a dense vector that is *learned* with a teacher-student model to mimic a dense representation of the manually rewritten query. The model for multiple turns uses composition with dense retrieval approaches similar to those in multi-hop QA (Khattab *et al.*, 2021), but applied to a conversational context. The results from Dalton *et al.* (2020c) include these as a baseline, and they are widely adopted by many of the top-performing teams in TREC CAsT. Most of the CIS systems, although they use a dense vector representation, still adopt simplistic history heuristics

to create one or more representations that may also leverage words combined via fusion. Further, for effectiveness most systems require an explicit text representation for further reranking (as discussed in the next section). This represents the start of a shift towards compressing history into a compact vector representation.

4.4 History Understanding Tasks

Given a representation of history and the parts of the conversation relevant to the current turn, the key consideration is how to use them in retrieval. This includes identifying important concepts or entities within them, performing query expansion, generating extractive summaries of key terms, and performing query rewriting.

4.4.1 Entity Recognition and Resolution

Tracking the evolution of topics discussed in a conversation is important to overall system effectiveness, particularly for mixed-initiative systems. Work on tracking entities across turns first appears in multi-turn factoid QA at TREC 2004 (Voorhees, 2004). This evolved with the introduction of 'related' questions that included anaphora and TREC 2005 with dependence on previous factoid responses (Voorhees, 2005). A related line of research uses search session history to improve named entity recognition effectiveness in queries (Du *et al.*, 2010). Approaches grounded in concepts and entities were widely used by Alexa Prize socialbot systems (Ram *et al.*, 2018) that allowed them to track topics and shifts across turns in similar ways as CIS systems.

A key consideration for these systems is that they need to be able to identify general concepts, commonly referred to as Wikification (Cucerzan, 2007). Joko *et al.* (2021) studied the effect of entity linking on conversational queries and found that existing linking systems perform significantly worse on conversations than on other types of text. As the current benchmarks only study non-personalized CIS systems, it would be expected that they will be even less effective for personal entities (Li *et al.*, 2014). They may also exhibit more localized issues of personal bias (Gerritse *et al.*, 2020).

4.4.2 Query Expansion

It is common practice to augment the representation of the current turn with additional representational information from previous turns in the conversation as well as from pseudo-relevance feedback (Yang *et al.*, 2019; Mele *et al.*, 2020; Hashemi *et al.*, 2020).

Similar to the previously discussed history modeling approaches, a commonly used approaches include heuristics as well as supervised approaches. In TREC CAsT a simple heuristic baseline expansion approach is to expand with the first and previous turn in the conversation (Clarke, 2019) because these turn often represent an overall topic and the most recent previous information. A mixture of feedback models (Diaz and Metzler, 2006) can be used to combine feedback models across turns. A common approach is to use previous questions and top-ranked answers to either expand the initial query, rerank candidate results, or both. The HQExp proposed by Yang *et al.* (2019) does both and leverages the combination of scores from a BERT model across past turns.

Some models perform expansion across turns, such as passage ranking by conversational reasoning over word networks (CROWN) (Kaiser *et al.*, 2020), an unsupervised method for propagating relationships across turns based upon a network of words related by mutual information.

The overall utility of these approaches is mixed, with many of the expansion approaches being outperformed by straightforward rewriting and reranking pipelines (Dalton *et al.*, 2019; Dalton *et al.*, 2020b). The best performing model applies diverse techniques: rewriting, expansion, and reranking in a multi-stage pipeline (Lin *et al.*, 2020c). An approach based upon both early and late fusion of multiple expansions and rewrites across both retrieval and reranking is currently required for effectiveness (Lin *et al.*, 2020c; Lin *et al.*, 2021). This indicates there is opportunity for more unified approaches combining the different sub-components.

4.4.2.1 Turn Salience

A key property of dialogue is the relation of the turns in the conversation to one another. Recent efforts focus on explicitly modeling the relationship of turns in a dialogue to determine their importance. The CAsTUR dataset created by Aliannejadi *et al.* (2020) adds turn salience data to the TREC CAsT 2019 dataset (Dalton *et al.*, 2019). The authors performed a detailed analysis of the dependencies. The resulting relation labels were used to train classifiers of turn salience (Kumar and Callan, 2020). We note that subsequent iterations of CAsT in 2020 and 2021 (Dalton *et al.*, 2020b) included explicit dependence annotation labels by the topic creators with labels on dependence on previous user utterances as well as previous results.

4.4.2.2 Term Selection and Summarization

Additional recent work is beginning to frame the task of expansion as a summarization task – extractive or abstractive. A recent direction is to automatically select previous turns and terms from them to use in query expansion. The Query Resolution by Term Classification (QuReTeC) model proposed by Voskarides *et al.* (2020) models the task as a binary term classification, effectively performing term-level extractive summarization. Parallel and similar work, the Conversational Term Selection (CVT) method by Kumar and Callan (2020) frames the problem as a term extraction task but further applies the same extraction to pseudo-relevant results. These methods extend previous methods that extract key concepts from long verbose queries for web search (Bendersky and Croft, 2008).

4.4.3 Conversational Query Rewriting

Given a query and a dialogue history, the goal of conversational query rewriting is to generate a new query that contains the relevant context needed to rank relevant content in a single unambiguous representation. In a pipeline system with multiple passes of retrieval, this step is critical because it determines the effectiveness of both candidate passage retrieval as well as subsequent re-ranking.

A widely adopted approach is to model the task as a sequence-to-sequence task (Sutskever *et al.*, 2014). The task-oriented dialogue systems community used pointer-generator networks and multi-task learning to rewrite turns but they are limited to a handful of task domains (Rastogi *et al.*, 2019). This approach rapidly evolved with pre-trained language models based on Transformer architectures (Vaswani *et al.*, 2017) and with evaluations on a Chinese dialogue dataset (Su *et al.*, 2019). They showed that these architectures implicitly solve coreference resolution more effectively for the target task than previous state-of-the-art coreference models.

Subsequent work by Vakulenko *et al.* (2020) in the TREC CAsT 2019 benchmark (Dalton *et al.*, 2019) demonstrated the effectiveness of pre-trained models based on GPT-2, resulting one of the top three best performing automatic runs in that year. Subsequent work showed the model can generalize with relatively few examples, particularly when combined with weak supervision based on rules to handle omission and coreference (Yu *et al.*, 2020). Improvements continued to evolve by training the models with additional data spanning both CAsT and conversational QA datasets (Elgohary *et al.*, 2019; Vakulenko *et al.*, 2020).

Subsequent work on newer generations of sequence-to-sequence models (*e.g.*, T5 (Raffel *et al.*, 2020)) based on larger corpora, increased model size, and refined objectives continued to show improvements (Lin *et al.*, 2020b). Additionally, recent work from the dialogue community (Henderson *et al.*, 2020; Mehri *et al.*, 2020) demonstrated that pre-training and fine-tuning on conversational data provides significant gains for both task-oriented and chit-chat models over models pre-trained on general corpora only. It is common to train on public large-scale social media data, such as Reddit (heavily filtered), because of its size and diverse informal language (Henderson *et al.*, 2020; Roller *et al.*, 2020). The use of informal chat data does raise, however, issues of safety and bias in the models.

4.5 Modeling Longer Conversations

Many of the existing conversational information seeking and question answering models are limited to conversations with a relatively small number of turns. Years one and two of TREC CAsT averaged between 8-10 turns (Dalton *et al.*, 2019; Dalton *et al.*, 2020b), QuAC has fewer, with approximately 7 turns (Choi *et al.*, 2018).

As conversations become longer, simple methods for modeling conversations break down. One reason for this is that most current neural models are only capable of encoding a limited (short) context of a few hundred tokens. To address this, approaches select conversational context to use, a task referred to as sentence selection in a dialogue (Dinan *et al.*, 2019a). For this task, previous turns and responses are ranked for relevance to the current turn using proven neural ranking models (Humeau *et al.*, 2020). Another common heuristic based on the observation that many dependencies are local is to use only the three previous turns (Mehri *et al.*, 2020). New methods for tackling the challenge of length are emerging, including the use of independently encoded contexts that are concatenated and then used in the decoder of an encoder-decoder model (Izacard and Grave, 2020) as well as methods that fuse independently encoded representations across turns (Xiong *et al.*, 2020).

Related work on summarization, text compression (Rae *et al.*, 2020), cascading (Hofstätter *et al.*, 2021), and efficient Transformer variants, *e.g.*, Conformer (Mitra *et al.*, 2021) and Longformer (Beltagy *et al.*, 2020), could also be applied to conversation histories. This is an area we discuss in future work.

4.5.1 Long Answer Result Dependence

Another dimension of modeling conversations is understanding long responses. Much of the previous related work focused on tracking and reformulation mostly based on previous utterances (queries) with only limited result interaction (Dalton *et al.*, 2019). The structure of previous conversational QA tasks had limited reliance on immediate previous results (Choi *et al.*, 2018; Voorhees, 2005). This is because the responses

given by the system are short factoid responses.

In contrast, the broader scope of CIS systems has richer questions that require varied length responses. These may be one or more passages, a document, or even an entire search engine results page. These long answers make the overall conversational history much longer than typical chit-chat dialogue models.

Interactions with long results in later turns make the language understanding significantly more challenging for CIS systems. They need to be able to understand references across a longer distance in more complex discourse. This remains a challenging area, where current neural approaches for conversation understanding struggle (Dalton *et al.*, 2020c).

4.6 Summary

This section reviewed conversational state tracking approaches and models. We examined the fundamentals of modeling intra-turn states including vector representations, entity representations, and discourse classification. We discussed approaches to model conversational history and differentiating features of conversational search as contrasted with voice search or traditional text narratives, with a key differentiating feature being wider use of implied context including indirect answers, zero-anaphora, and ellipsis. We discussed long-range history models with many current approaches using a static window of context (last few turns) as well as dynamic turn salience, or attention-based models. Within this history, we examined key sub-tasks: entity recognition and linking, query expansion, query summarization, and query rewriting. Overall, this section suggested that notably challenging areas in understanding conversational interactions include result dependence on long responses as well as modeling long conversations, possibly spanning multiple sessions.

5

Response Ranking and Generation

In this section, we discuss response ranking and generation used in conversational information seeking. The task of response ranking is selecting the relevant information item(s) for a turn in the conversation from the knowledge available to a conversational system. The types of methods are often categorized based upon the type of conversational response provided: short answer (QA), longer single passage or document, automatically generated responses from extractive or abstractive summarization, and structured entities (products, restaurants, locations, movies, books, *etc*).

5.1 Short Answer Selection and Generation

This section covers an overview of Conversational QA (ConvQA) approaches, also referred to as Conversational Machine Comprehension in the NLP community. ConvQA often assumes that the question in each turn is answerable by a span of text within a particular passage (from a conversational retrieval model) and selects one or more spans of text from the passages. We begin by discussing traditional models, then more recent neural approaches, and end with recent work combining elements of retrieval and selection with end-to-end approaches.

The evolution of ConvQA follows advances in QA and machine comprehension. The adoption of deep neural models brought new interest in the task and approaches. They are the building blocks for later ConvQA models. Early models are extractive and select one or more spans of text as the answer(s). These models have evolved to use generative sequence-to-sequence models. This evolution has been heavily influenced by the publicly available resources. To support comparison between methods this section discusses most methods in the context of these datasets.

5.1.1 Early Conversational QA Models

Early ConvQA models started in the TREC 2004 QA Track (Voorhees, 2004; Voorhees, 2005) with questions grouped into different series related to a single target entity (or event). Each question asks for more information about the target. This required models to use previous questions in the sequence, mainly the first with the target. Unlike a dialogue or conversation, the questions did not mention answers (responses) from previous questions in the series, resulting in a limited discourse structure. Because the dependence could be resolved using rule-based and straightforward models, the response ranking methods did not leverage the multi-turn nature of the series.

A building block for later ConvQA models is extractive neural models for single turn QA. Notable models include DrQA (Chen *et al.*, 2017) and BiDAF (Seo *et al.*, 2017) that use Recurrent Neural Networks – specifically bidirectional long short-term memory networks (Bi-LSTMs). The BiDAF++ QA model (Peters *et al.*, 2018) includes self-attention and the use of pre-trained contextualized word vectors (ELMo). Later Pointer Generator Networks (See *et al.*, 2017) extended these by supporting copying spans from an input context in the decoder. These models and related datasets are extractive QA and do not focus significantly on ranking the input text. They are also not conversational, although as we discuss next they were later adapted to encode conversational context.

The shift from QA to ConvQA for these models required the development of new benchmark datasets. The Question Answering in Context

(QuAC) dataset (Choi *et al.*, 2018) is one of the early ones. The baseline model on the dataset was BiDAF++, the state-of-the-art QA model at the time. To adapt it for ConvQA, the conversational history was appended (as described in Section 4) and was referred to as 'BiDAF++ w/ k-Context'. This model appends previous k (1-3) answers (their contextual embeddings) as context, along with the question turn number. We note that the QuAC dataset is limited to people entities, with a particular emphasis on entertainment.[1] Concurrent with QuAC, the CoQA benchmark (Reddy *et al.*, 2019) was released with similar goals. Because of its crowdsourcing task setup, the extracts are shorter (2.7 words vs over 15 for QuAC). It includes conversational questions from seven diverse domains: children's stories, literature, middle and high school English exams, news, Wikipedia, Reddit, and science. The CoQA baseline models were also similarly single-turn QA models adapted for conversation. They used BiDAF++ w/ k-Context. They also extended the DrQA model (Chen *et al.*, 2017) by including context history markers to separate turns, which outperforms the BiDAF model variants. These datasets and models are important because they represent the first steps towards a large-scale evaluation of ConvQA systems with models simply adapted from previous QA systems.

One of the first steps towards new models explicitly designed for conversation are models that incorporate Flow (FlowQA) (Huang *et al.*, 2019) to model the conversational dialogue. Instead of appending history with a marker, they introduce a method that provides the model access to the full latent state used to answer the previous questions. This is a stack of two recurrent networks - one for each turn and one across turns. Their first model uses Bi-LSTMs to encode each turn and then processes each full turn representation linked with GRUs (for efficiency reasons). This represents a significant advancement over previous models that were extended to other types of networks including Transformers, discussed next.

[1] See the datasheet description of QuAC for details including details of bias, https://quac.ai/datasheet.pdf

5.1.2 Conversational QA with Transformers

The introduction of pre-trained language models based on Transformer architecture that support transfer learning represents a significant shift for ConvQA systems. This subsection describes this evolution in approaches and challenges with these models.

Following the early success of Transformer-based models, such as BERT (Devlin *et al.*, 2019) in QA tasks, these models were applied to ConvQA and yielded similarly impressive improvements. In many cases, the early work using Transformer approaches simply appended previous turn context with separators similar to previous extensions of BiDAF and DrQA. However, results show this has significant limitations. Naive approaches appending answer context degrade faster because of the limitations of the sequence input length (Qu *et al.*, 2019a). To overcome these issues, Qu *et al.* (2019a) proposed the History Answer Embedding (HAE) model that uses BERT for ConvQA while modifying the representation to explicitly encode whether parts of the input are present in the previous history. On QuAC they found that this model outperforms BERT models that naively append the question or answer history, and is also more robust to appending longer conversations. In a different thread, Yeh and Chen (2019) introduced the FlowDelta model that extends the previously discussed Flow model to use BERT for encoding, as well as changing the Flow loss to focus on the difference in Flow (Delta) across turns. They found that the proposed FlowDelta outperforms the previous Flow and BERT-based models.

A long-standing top performing system on the CoQA leaderboard is RoBERTa+AT+KD (Ju *et al.*, 2019), an extractive model using a RoBERTa language model in combination with Adversarial Training (AT) that performs perturbation of the contextual embedding layer and Knowledge Distillation (KD) using a student-teacher setup. It ensembles nine models and has a post-processing step for the multiple-choice questions to match extracted spans to the target answer. Beyond the leaderboard, Staliunaite and Iacobacci (2020) studied the behavior of BERT- and RoBERTa-based models on CoQA. They found that the key gain between the base models is that RoBERTa provides a better lexical representation. However, it does not capture more of the

fundamental linguistic properties in ConvQA. To address these issues, they tested incorporating varying types of linguistic relationships in a multi-task model and combined the models in an ensemble. They found that incorporating the linguistic structure outperforms the base models. This indicates that the base representation of the language model is important for effectiveness and that there is an opportunity for models that incorporate more linguistic and conversational discourse structure.

The behavior of the current models is constrained by issues with current datasets and task formulation. An issue highlighted by Mandya *et al.* (2020) is exposure bias: CoQA systems use gold answer labels for previous turns in both training and test time. In contrast, using the models' predictions may result in compounding error, with a significant reduction in effectiveness compared with ones that use system predictions rather than previous gold answers. They found this is particularly problematic for longer conversations and longer (more complex) questions. As discussed later, there is a similar phenomenon for retrieval systems using conversational query rewriting. Systems that use manual query rewrites instead of predicted ones for earlier turns overestimate their effectiveness (Gemmell and Dalton, 2020).

The ConvQA models and datasets (QuAC and CoQA) use a short pre-defined narrative of 200-400 tokens with the conversation focusing on one passage. As a result, the previously discussed ConvQA systems work well for extracting information from short passages with conversations grounded to a single paragraph. Further because of the way they were constructed, the headroom for generative models is very limited, approximately 5% on CoQA (Mandya *et al.*, 2020). The next subsection covers more realistic models that include the retrieval of passages in the QA process.

5.1.3 Open Retrieval Conversational QA

This subsection discusses approaches that incorporate retrieval into the ConvQA task. This is referred to as open retrieval ConvQA (OR-ConvQA) or end-to-end ConvQA. The distinguishing feature of these models is that they operate on a large corpus of passage content and rank the passages used in the conversation. A common architecture

for these systems is that they consist of two components - a *Retriever* and a *Reader*. The Retriever takes the conversational context and uses it to identify candidate passages. The Reader takes the context and candidates (text) and produces an answer. The base retrieval systems are effectively the ConvPR long answer systems discussed below in Section 5.2 combined with a QA reader model to extract or generate the answer. A key challenge is that the existing ConvQA benchmarks are not designed for open retrieval QA and that current conversational passage retrieval benchmarks do not have short answer annotations. As a result, recent efforts (Qu *et al.*, 2020; Gao *et al.*, 2021b; Ren *et al.*, 2020a) adapted and extended the datasets to bridge this gap. The first of these by Qu *et al.* (2020) extended QuAC to incorporate passage retrieval over Wikipedia, creating the OR-QuAC dataset. To do this a synthetic query representing the information need is created by providing the Wikipedia title and first paragraph with the initial question that is rewritten to be unambiguous.

Recent developments in dense retrieval are also being applied to OR-ConvQA. Qu *et al.* (2020) performed retrieval using a dot-product of a query history representation (previous k queries) and a passage that is based upon a learned query and passage encodings using ALBERT (Lan *et al.*, 2020), a lite BERT representation. One of the novel contributions is the multi-task training objective where the retriever, a BERT-based cross-attention reranker, and a BERT-based reader are trained concurrently to avoid issues of error propagation. Another contribution is that it uses a distribution over candidate answers. One potential issue is that for training the model, golden query rewrites are used rather than employing a noisy query rewriter. This approach was further extended and improved upon by leveraging distant supervision to handle the free-form responses more effectively (Qu *et al.*, 2021).

One of the large-scale efforts in OR-ConvQA is the development of the Question Rewriting in Conversational Context dataset (Anantha *et al.*, 2021). For a baseline, they used a BERTserini passage retriever combined with a BERT-large reader model. They found that a key factor in the success of reader models that leverage retrieval is incorporating the passage relevance score into the reader model (Anantha *et al.*, 2021). Recent results by Tredici *et al.* (2021) demonstrate that different

representations should be used for retrieving and reading models.

5.1.4 Response Generation for Conversational QA

Recent trends in QA increasingly have a focus on generative sequence-to-sequence models. These models are used to (1) perform generation and put the answer in the conversational context, and (2) to make the model more effective by generating responses from retrieved passages and past sessions.

The first type focuses on the conversational natural language generation of answers, putting the answer into the natural conversational context and focusing on fluency. They follow the pattern of *retrieve* and *refine* (Weston *et al.*, 2018). The initial results, retrieved from previous document collections or previous conversation responses, are used as the context that is refined during generation. The refining processing connects the answer to the dialogue and puts it into a natural language response form. An example of this is AnswerBART (Peshterliev *et al.*, 2021), which provides an end-to-end model that performs answer ranking, generation, and includes abstaining from answering when there is none. A novelty of this model is that it jointly learns passage reranking with the extraction task. A variation of this is treating generation as a ranking task. Baheti *et al.* (2020) used syntactic patterns and templates to generate multiple candidate responses. This was combined with a GPT-2 based model that was pre-trained on Reddit conversations. These models focus primarily on fluency and putting the answer in context.

The ability to incorporate long-term memory from past sessions is important for CIS systems. The work from Shuster *et al.* (2021) extended (Lewis *et al.*, 2020) by incorporating the turn structure for knowledge-grounded conversations and they found this reduces model hallucination (i.e., producing factually invalid information) and results in a model that generalizes more effectively. Going beyond this, the work from Xu *et al.* (2021) extended the retrieval aspect to incorporate retrieval from past conversational sessions. The model retrieves sessions as a document and uses these as the context in the generation process.

One limitation of many of these ConvQA approaches is that because the answers are short (even if they are put into a natural language

utterance), they are usually simple factoid responses. As a result, the level of discussion in the conversation does not discuss aspects of the response and the the ability to reference previous results in follow-up parts of the conversation is limited. The Question Rewriting in Conversational Context (QReCC) dataset from Anantha *et al.* (2021) is noteworthy because approximately 25% of the answers are not simple extractions, but are human-generated paraphrases, possibly of multiple passages. Systems and behavior with these types of richer responses continue to evolve and represents an area for further work.

This section covered multiple threads of the evolution of these systems to use Transformer and attention-based architectures for ConvQA. They focus on improving the contextualized encoding (BERT vs RoBERTa), multi-task learning of discourse or token importance, stacking networks to capture cross-turn relationships, and approaches to make the models more robust using adversarial training and data augmentation. Recent work by Kim *et al.* (2021) brought together generative conversational query rewriting using T5 in the QA process and showed that it outperforms more complex models that attempt to model both simultaneously. The models largely target factoid QA with most being extractive, possibly with minor adaptions for yes/no questions or multiple choice. None of the existing ConvQA benchmarks are based upon real user information needs (queries) with multiple results from retrieval. This represents an opportunity for new systems and methods to evolve towards more realistic tasks based upon real information needs.

5.1.5 Conversational QA on Knowledge Graphs

Similar to parallel threads in question answering over unstructured text, ConvQA can also be performed on structured knowledge graphs (KGs) containing entities. This sub-area of conversational QA over a knowledge graphs is called KG-ConvQA. These approaches allow conversational information seeking over structured data. Therefore, the nature of the questions they can answer are also structured and may involve logical operations including joins, aggregations, quantitative comparisons, and temporal references. KG-ConvQA systems may be partitioned into two

distinct types. The first performs QA directly using the KG internally or traversing it using actions to produce an answer. The second type performs conversational semantic parsing and produces an executable logical structured query for producing the answer.

For the first type of KG-ConvQA systems, a neural sequence-to-sequence model is combined with a memory network to generate an answer. One of the first attempts to do this was done by Saha *et al.* (2018), who introduced the Complex Sequential QA (CSQA) dataset and baseline model. A baseline for KG-ConvQA is HRED+KVmem, which combines a base conversational recurrent neural network (RNN) model, HRED, with a key-value memory network for modeling the KG, and finally a RNN decoder to generate the answer. This baseline model works well for many categories of questions, but struggles with quantitative and comparative reasoning.

Another approach, CONVEX, proposed by Christmann *et al.* (2019) starts from a seed entity and performs actions to traverse the graph to identify an answer. To handle the conversational evolution, CONVEX maintains a dynamic sub-graph that changes with each conversational turn using look-ahead, weighting, and pruning techniques to limit the graph size. This is effective because traversing the graph on the evaluation benchmark CONVQUESTIONS finds answers that are relatively close (no more than five edges away from the seed entity that starts the conversion).

The dynamic sub-graph approach was extended by Kaiser *et al.* (2021) with their model, CONQUER. It uses reinforcement learning to select graph traversal actions. CONQUER maintains a set of context entities from which the agents traverse the graph. Their model uses a policy network that uses weak supervision from a fine-tuned BERT model. One of the key differences from previous work is that the model also predicts if the query is a reformulation. This is built on an extension of the CONVQUESTIONS dataset that adds manual reformulations when the baseline system produces incorrect answers. The results show that CONQUER outperforms CONVEX, demonstrating that reformulation and the policy network outperforms the previous sub-graph tracking approach, particularly when there is implicit feedback from reformulation for wrong answers.

The second direction taken to address this task is based upon conversational semantic parsing. Instead of generating an answer, these approaches generate a structured grammar. Guo *et al.* (2018) propose the Dialog-to-Action (D2A) model that builds on a GRU sequence-to-sequence model with a question and context from interaction history and outputs an action sequence from a predefined grammar. The dialogue history is managed in the action space. In contrast to the earlier HRED+KVmem model the D2A model is much more effective, particularly for queries requiring reasoning.

Subsequent approaches improve upon the semantic parsing quality by incorporating entity recognition and disambiguation in the semantic parsing process with multi-task learning. For instance, Shen *et al.* (2019) presented the Multi-task Semantic Parsing (MaSP) model, performing both entity typing and coreference resolution together with semantic parsing. A subsequent multi-task model is CARTON (Context trAnsformeR sTacked pOinter Networks) (Plepi *et al.*, 2021), with an encoder and decoder model to model the conversational representations. A series of three stacked pointer networks focus on the logical form needed for execution (types, predicates, and entities).

A later approach using Transformers with multi-task learning and graph attention (LASAGNE) by Kacupaj *et al.* (2021) built on this semantic parsing approach leveraging a graph attention network. It has a grammar guided Transformer model to generate logical forms as well as a sub-model that learns correlations between predicates and entity types to avoid spurious logical forms. LASAGNE appears to outperform CARTON across most categories. However, CARTON performs better on coreference and quantitative reasoning. They perform ranking on the KG by selecting the relevant entity and are implicit in the semantic parses produced by the model.

The work from Marion *et al.* (2021) generates a hierarchical JSON-like logical form that is KG executable. They used an Object-Aware Transformer that includes entity linking. They highlighted that the CSQA approaches often use a base gold seed entity and only require coreference to the previous turn. The results demonstrate strong effectiveness across multiple datasets using pre-trained encoder-decoder models.

The focus of most of KG-ConvQA models is traversing the graph for structured comparison. The conversational structure, such as ellipsis and dependency support are limited in current models.

5.2 Conversational Long Answer Ranking

This subsection discusses open-domain conversational long answer retrieval, sometimes referred to as ConvPR (for Passage Ranking). Analogous to the previous distinction between ConvQA and OR-ConvQA (see Section 5.1.3), this subsection distinguishes between ConvPR and OR-ConvPR. ConvPR focuses on conversational passage reranking from a closed set of responses. In contrast, OR-ConvPR includes full retrieval over a corpus of passages in the ranking step. The questions may require one or more long answers to sufficiently answer the questions. This class of responses covers work on Ubuntu/Quora, MSDialog, AliMe, TREC CAsT, and similar corpora.

The task of ConvPR has a rich history that builds on response retrieval and selection from discussion forums. These models have a long history in retrieval-based chatbots, see (Tao *et al.*, 2021) for details. For the ConvPR task, the Deep Attention Matching Network (Zhou *et al.*, 2018) encodes each turn with a transformer model and combines them with a matching network and a final 3D convolutional network that incorporates the history. The intent-aware ranking model from Yang *et al.* (2020) extends this model by adding explicit conversation intents. The encoding is similar to DAM, but it also produces a vector representing the user intent. This represents dialogue discourse types specific to CIS and includes: asking a question, clarification, elaboration on details, and both positive and negative feedback. The encoded turns are incorporated combined with the intent classification with a weighted attention model and aggregated into a matching tensor. Similar to DAM, the result is used in a final two-layer 3D-CNN model to rerank the candidate responses.

One of the fundamental aspects of the effectiveness of any ConvPR model is the language model used in encoding. Many of the encodings

used are off-the-shelf language models, but an alternative is to perform a step of model fine-tuning with the language modeling objective on conversational corpora. Current leading approaches in chatbots and similar use models are trained on heavily filtered and curated conversations from web forums like Reddit. For example, the ConveRT model (Henderson *et al.*, 2020) fine-tunes a BERT-based model on Reddit discussions and applies the resulting model to the task of response selection. This pre-training objective results in significant gains on Ubuntu DSTC7 and the AmazonQA response selection tasks. It is also widely used as a pre-training objective for dialogue system models. In contrast to the intent-aware model, these do not use pre-defined intents and instead learn common discourse patterns directly from the text.

In contrast to the previous ConvPR models, the OR-ConvPR models must retrieve and optionally rerank from large passage corpora. As a result, a current pattern exemplified by many CAsT systems is a pipeline with specialized modules. This includes modules that focus on understanding the context, as described in Section 4, that include conversational question rewriting and expansion across turns. These are then used with neural ranking models for passage retrieval. For more information on neural ranking models, we refer the readers to the recent survey articles (Mitra and Craswell, 2018; Guo *et al.*, 2020; Lin *et al.*, 2020a). This architecture allows existing components trained on large existing datasets for query expansion, rewriting, and ConvPR to be used in the open retrieval context.

It is common to use a multi-stage cascaded architecture for OR-ConvPR tasks. One of the prototypical multi-stage systems that perform this is developed by Lin *et al.* (2021). A core building block of their approach is Historical Query Expansion (HQE) that generates expanded queries based upon the dialogue history using a sequence-to-sequence model. The conversational query rewriting is a standard T5 model trained on QuAC/Canard. One aspect is that the system additionally performs rank fusion to combine multiple query interpretations and formulations. This fusion can be performed early (in initial retrieval) or late (in reranking) and they find that fusion in early retrieval is critical for getting sufficient candidate passages in the pipeline for later reranking.

Instead of the multi-stage cascade architecture, an alternative is end-to-end approaches based upon dense retrieval, sometimes referred to as Conversational Dense Retrieval (ConvDR) (Yu *et al.*, 2021). The distinguishing feature is that retrieval and conversation are encoded with a dense vector rather than an explicit word-based query. The representation of the turn is a combination of the vectors from previous turns. This approach can also be applied to OR-ConvQA. Although not (yet) as effective as the best complex pipeline systems, it is an active research area.

5.3 Long-Form Response Generation for CIS

The previous subsection discussed retrieval of (mostly) passage-level responses. In contrast, a recent development is extractive or generative summarization of retrieved results appropriate to a conversational interface and in a conversational context, referred to as ConvSum.

One approach to handling multiple retrieved documents or passages for CIS is to combine them with extractive summarization approaches. This is particularly important for summarizing long documents for CIS interfaces and interaction. A text editing approach is to keep, delete, or make small insertions. This approach is used by the LaserTagger (Malmi *et al.*, 2019) and Felix (Mallinson *et al.*, 2020) models. They leverage pre-trained Transformers trained with supervised data. They produce responses for CIS applications that are true to the document and add elements of fluency by putting them in a conversational form.

Beyond extractive approaches, generative chatbots are evolving towards longer and more complex information responses. Recently, these developments include pre-training of language models for generation on dialogue, such as Language Model for Dialogue Applications (LaMDA) that builds upon the previous Meena (Adiwardana *et al.*, 2020) architecture based on Evolved Transformers and trained on social media data. A key consideration for generative models is their factual consistency and fidelity to the input passages, with previous work showing that the degree to which the model uses the input varies (Krishna *et al.*, 2021). To address this for short answers, an early benchmark by Dziri *et al.* (2021), Benchmark for Evaluation of Grounded INteraction (BE-

GIN), uses generated responses from Wizard-of-Wikipedia (Dinan *et al.*, 2019b).

Generative approaches for long answers remain a significant open area of research for CIS. This area is particularly important as generated answers become longer and more complex. Recent work on long-form generation by Krishna *et al.* (2021) uses the Routing Transformer with dynamic attention routing to support longer sequences. Similar models that take into account long-term conversational context will need similar approaches. New models will evolve to handle longer conversations, more varied types of summarization for complex questions, and ones that focus on factuality and knowledge-grounding to overcome issues of correctness and hallucination.

5.4 Procedural and Task-Oriented Ranking

The previous subsections describe formulations of CIS response ranking that largely extend previous research from QA, retrieval, and recommendation to a conversational context. However, because CIS systems are often embedded or used in combination with task assistants, the types of information needs and tasks performed are more likely to be grounded in procedures and real-world tasks. Information seeking is interleaved with task-oriented processes and structured dialogue actions, such as task navigation (Ren *et al.*, 2021; Azzopardi *et al.*, 2018). This subsection discusses multiple veins of work in these areas and their connection to CIS.

5.4.1 Procedural Question Answering

In Procedural QA, the task is to interact conversationally to determine outcomes based on complex processes represented in text documents. To address this task, Saeidi *et al.* (2018) introduced the Shaping Answers with Rules through Conversation (ShARC) benchmark. It contains varied types of discourse and natural language inference required within it. The procedures come from conversations on complex regulatory decisions. Because they are vague, the model must generate clarifying questions and understand the complex rule structures encoded in docu-

ments. Instead of providing excerpts like a typical QA task, the goal is to use rules in the text and the conversational responses to infer a yes/no answer. Similar to the evolution of other QA systems, a baseline model for this task includes a conversational BiDAF model for encoding history which is then combined with a natural language inference model, such as the Decomposed Attention Model (DAM) (Parikh *et al.*, 2016) for interpreting rules.

Subsequent work (Gao *et al.*, 2020) focused on segmenting documents into elementary discourse units (EDUs) which are tracked through the conversation. Going further, recent work built on this by explicitly modeling the conversational structure using Graph Convolutional Networks (GCNs) (Ouyang *et al.*, 2021). The results show that using both explicit and implicit graph representations allows the model to more effectively address conversations with complex types of discourse structure. Mirroring the evolution of QA towards open retrieval, Gao *et al.* (2021b) extended the ShARC conversational entailment task by adding rule retrieval, creating OR-ShARC. In this task setup systems must first search a knowledge base of rule texts with context from the user and scenario (although it is limited to rule texts used in the original ShARC benchmark). It uses a basic TF-IDF retriever achieving over 95% recall in the top 20 rules; approximately the top five rules are used with a recall of over 90%. These are used in a RoBERTa machine comprehension system that also leverages inter-sentence Transformer layers to combine evidence. It is noteworthy that systems capable of reading multiple passages in the top-k retrieved results, *e.g.*, (Dehghani *et al.*, 2019), can be more effective than systems that only use the top (often golden) rule.

5.4.2 Task-Oriented Information Seeking

Task-based virtual assistants perform tasks in the world. They are largely separate from CIS systems. Recently, there is a trend towards systems and models capable of both. A critical aspect for CIS is that information seeking is occurring within an explicit task context with domains and intents. It may start with conversational search to find an appropriate agent or task to execute (for example, finding a recipe to cook) and then

unfold as the task is performed. This may involve answering procedural questions grounded in the task execution, questions requiring external knowledge for QA, and other types of information needs. The CIS should also respond to changes in the task execution environment. From the dialogue community, this task was proposed and evaluated as part of the DSTC challenge, (Kim *et al.*, 2020).

The recent Amazon Alexa Taskbot challenge[2] introduced the challenge of using multi-modal conversation to solve real-world tasks. This challenge includes conversational task retrieval and refinement, task-oriented QA, and conversational procedural instruction responses. Further, because interactions are multi-modal (including voice and screen), the responses may include images and videos in response to the information need. In practice, this means that elements of a dialogue system to navigate the task are interleaved with task-specific question answering and open-domain question answering. Additionally, the goal is also implicitly to select responses that are natural and engaging for the user with elements of social chat related to the task.

5.5 Conversational Recommendation

Traditionally, recommender systems mainly exploit historical user-item interactions for predicting user preferences. This has led to the development of collaborative filtering methods which are at the core of effective real-world recommendation engines. Other recommendation models, such as content-based and demographic filtering, have also been studied and showed promising results mostly for cold-start users and items. All of these approaches provide users with little control over the recommended list. For instance, users often cannot ask for a revised recommendation list based on their preferences. Conversational recommender systems address this limitation. During a human-machine dialogue, the system can elicit the current user preferences, provide explanations for the recommended items, and/or take feedback from users for recommendation refinement.

[2] https://www.amazon.science/academic-engagements/amazon-launches-new-alexa-prize-taskbot-challenge

Interactive and conversational recommender systems have been studied for several years (Thompson *et al.*, 2004; Mirzadeh *et al.*, 2005; Mahmood and Ricci, 2009; Blanco and Ricci, 2013). Due to the potential real-world applications, conversational recommender systems have recently attracted considerable attention. Most efforts in this domain focus on preference elicitation by asking questions from users. Christakopoulou *et al.* (2016) studied this task and proposed a conversational model based on probabilistic matrix factorization for a restaurant recommendation. They proposed to initialize the conversational recommendation model's parameters by training the model on offline historical data and updating the parameters while the users interact with the system through online learning. They focused on question selection from a question bank during the online interactions for preference elicitation. This approach was later revisited by Zhang *et al.* (2018) who used multi-memory neural networks for template-based question generation. They unified conversational search and recommendation and trained their model based on item reviews in the e-commerce domain. In more detail, they extracted attribute-value pairs mentioned by users about items in their reviews and train a model that generates attribute-based questions based on the attributes mentioned by the user.

Another line of research focuses on modeling conversational recommendation using reinforcement learning (RL). Sun and Zhang (2018) developed an early interactive RL-based recommendation model that can take two actions: (1) selecting an attribute (or facet) for preference elicitation, or (2) making a personalized recommendation. They simply used a two-layer fully-connected neural network as the policy network and defined the reward function based on the recommendation quality at every timestep during the dialogue. They demonstrated the benefits of conversational recommendation via both offline and online experimentation.

More recently, Lei *et al.* (2020a) introduced the *Estimation-Action-Reflection* (EAR) framework for conversational recommendation. This framework unifies the following three fundamental problems in conversational recommendation: (1) what questions to ask, (2) when to recommend items, and (3) how to adapt to the users' online preferences. Another approach to conversational recommendation is to exploit

multi-armed bandit solutions which have shown promising results in the sequential and interactive recommendation. Zhang *et al.* (2020c) followed this path and proposed conversational contextual bandit. Later on, Li *et al.* (2021) improves this model by introducing the Conversational Thompson Sampling (ConTS) model. ConTS builds upon multi-armed bandit and models items and attributes jointly. This enables the model to compute the exploration-exploitation trade-off between preference elicitation and item recommendation automatically.

5.6 Summary

This section focused on core conversational response ranking. The models started with ConvQA, with basic extractive factoid QA with context naively appended that operated in a closed environment. These evolved towards variants that are more realistic by incorporating retrieval from a corpus (OR-ConvQA), including incorporating multiple results and their retrieval scores as well as other forms of external memory, including past turns or conversations.

As the retrieval task evolved towards longer and exploratory responses (OR-ConvPR and OR-ConvDR), the systems evolved to be complex pipelines that required query rewriting, query expansion, dense retrieval, multi-pass re-ranking, and result fusion. However, the ranking components are still largely separated and trained on external datasets specific to those tasks.

Later, models evolved to include conversational models of richer types of responses, including entities (KG-ConvQA), as well as ones that are longer and more natural. Longer and more complex responses support richer types of result dependency and more natural conversations. This includes generating responses from one or more retrieved sources. Most of the effective ranking and generation models build upon pre-trained language models. The effectiveness of the models varies depending on their lexical representations and training to support linguistic and conversational structure. The most effective ones include additional fine-tuning with a language modeling objective on the target conversational data before a final task-specific training. For ranking models there is a common pattern of having a model for a single turn

and then incorporating evidence across turns by stacking models to capture conversational context (*e.g.*, Flow or 3D-CNNs).

Finally, the section covered response ranking for structured prediction in the form of task-oriented dialogues and recommendation. Beyond these, the ranking tasks and models will continue to evolve to include richer types of responses and to support more realistic and complex information seeking tasks.

6

Mixed-Initiative Interactions

Most approaches to human-computer interactions with intelligent systems are either controlled by a person or the system (i.e., user- or system-initiative). For example, in current search engines, users always initiate the interaction by submitting a query and the search engine responds with a result page. Therefore, search engines are user-initiative systems. That being said, developing intelligent systems that support *mixed-initiative interactions* has always been desired. Allen *et al.* (1999) believed that development of mixed-initiative intelligent systems will ultimately revolutionize the world of computing. Mixed-initiative interactions in dialogue systems have been explored since the 1980s (Kitano and Van Ess-Dykema, 1991; Novick and Douglas, 1988; Walker and Whittaker, 1990). Early attempts to build systems that support mixed-initiative interactions include the LookOut system (Horvitz, 1999) for scheduling and meeting management in Microsoft Outlook, Clippit[1] for assisting users in Microsoft Office, and TRIPS (Ferguson and Allen, 1998) for assisting users in problem solving and planning.

Horvitz (1999) identified 12 principles that systems with mixed-initiative user interfaces must follow. They are listed in Table 6.1. In

[1] https://en.wikipedia.org/wiki/Office_Assistant

Table 6.1: Principles of mixed-initiative user interfaces by Horvitz (1999).

Principle
1. Providing genuine value
2. Considering uncertainty about user intents
3. Considering the user status in the timing of services
4. Inferring ideal action in light of cost, benefit, and uncertainties
5. Employing dialogue with users to resolve key uncertainties
6. Allowing efficient direct invocation and termination
7. Minimizing the cost of poor guesses about action and timing
8. Scoping precision of service to match uncertainty in goals
9. Providing mechanisms for efficient result refinement
10. Employing socially appropriate behaviors
11. Maintaining working memory of past interactions
12. Continuing to learn by observing

summary, mixed-initiative interactions should be taken at the right time in the light of cost, benefit, and uncertainties. Many factors mentioned in these principles can impact cost and benefit of interactions. In addition, systems with mixed-initiative interactions should put the user at the center and allow efficient invocation and termination. Systems with mixed-initiative interactions are expected to memorize past interactions and continuously learn by observation. Based on these principles, conversational systems by nature raise the opportunity of mixed-initiative interactions.

Allen *et al.* (1999) defined four levels of mixed-initiative interactions in the context of dialogue systems, as follows:

1. **Unsolicited reporting:** An agent notifies others of critical information as it arises. For example, an agent may constantly monitor the progress for the plan under development. In this case, the agent can notify the other agents (e.g., user) if the plan changes.

2. **Subdialogue initiation:** An agent initiates subdialogues to clarify, correct, and so on. For example, in a dialogue between a user and a system, the system may ask a question to clarify the user's

intent. Since the system asks the question and the user answers the question, and this may be repeated for multiple turns, the system has temporarily taken the initiative until the issue is resolved. This is why it is called subdialogue initiation.

3. **Fixed subtask initiation:** An agent takes initiative to solve predefined subtasks. In this case, the agent can take initiative to ask questions and complete the subtask. Once the subtask is completed, initiative reverts to the user.

4. **Negotiated mixed-initiative:** Agents coordinate and negotiate with other agents to determine initiative. This is mainly defined for multi-agent systems in which agents decide whether they are qualified to complete a task or it should be left for other agents.

When it comes to (pro-active) open-domain conversational information seeking, some of these mixed-initiative levels remain valid. Mixed-initiative interactions in the context of CIS have been relatively less explored, but are nevertheless identified as critical components of a CIS system (Radlinski and Craswell, 2017; Trippas *et al.*, 2018; Aliannejadi *et al.*, 2019; Wadhwa and Zamani, 2021). Vakulenko *et al.* (2021) conducted a large-scale analysis of 16 publicly available dialogue datasets and established close relations between conversational information seeking and other dialogue systems. Clarification and preference elicitation are the two areas related to mixed-initiative interactions that have attracted much attention in recent years. Therefore, in the rest of this section, we first review the role of agents in initiating a conversation (Section 6.1), and continue with discussing methods for generating, analyzing, and evaluating clarification in conversational search (Section 6.2). We further summarize preference elicitation in conversational recommendation (Section 6.3). We finally discuss how the user and system can be involved in mixed-initiative interactions with the goal of providing feedback (Section 6.4).

6.1 System-Initiative Information Seeking Conversations

Typically, users initiate the interaction with a conversational system, for example by clicking or touching a link or button, by using pre-defined

voice commands such as "Alexa" or "OK Google", or by asking a question or submitting an action request. In mixed-initiative conversational systems, the agent is also able to initiate the conversation. This is also called a system-initiative (or agent-initiative) conversation. Making a recommendation is perhaps the most common scenario for initiating an interaction by the system. For example, a CIS system can initiate a conversation by recommending an item based on the situational context of the user (e.g., location and time) and their preferences. Note that this is different from many conversational recommendation settings, where users first submit a request about the item they are looking for, *e.g.*, (Sun and Zhang, 2018; Zhang *et al.*, 2018). Joint modeling of search and recommendation (Zamani and Croft, 2020a; Zamani and Croft, 2020b) is a step towards developing mixed-initiative search and recommendation systems. However, initiating a conversation by the system is not limited to recommendation. For instance, Avula and Arguello (2020) developed a system for conducting wizard-of-oz experiments to study system-initiative interactions during conversational collaborative search. This system can be integrated into Slack.[2] In this system, while a group of users are performing a collaborative search task, another user (who plays the role of wizard) can intervene and provide additional information. Although little progress has been made in this area, there is a great potential for systems to initiate conversations based on context and engage with users or even get feedback. For instance, assume a user drives to a restaurant using a map application, e.g., Google Maps. When it has access to the context, a CIS system can initiate a conversation when the user is driving back, by asking about their experience at the restaurant. This can potentially lead to improving the user experience with the conversational system, collecting feedback on the restaurant, and also collecting information on the user's preferences for improving the user profile. As another example, if a user is struggling with completing a task, a CIS system can be automatically triggered to start the conversation with the user, hear their complaints, and help them complete the task. Related to this line of research, Rosset *et al.* (2020) studied how a system can lead a conversation while users are

[2] https://slack.com/

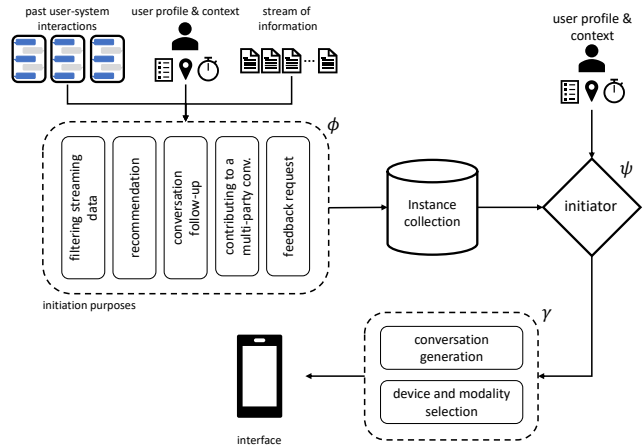

Figure 6.1: A generic pipeline for conversation initiation in CIS systems by Wadhwa and Zamani (2021).

searching for or exploring a topic. They formulated the problem as a conversational question suggestion task and demonstrated its impact by presenting the question suggestions in search engine result pages.

Initiating a conversation by the system can be risky and it may annoy users and hurt user satisfaction and trust. For instance, in some situations, a user may not be interested in engaging in a conversation, and thus predicting opportune moments for conversation initiation is an important part of developing system-initiative CIS systems. Therefore, whether and when to initiate a conversation are the key decisions a mixed-initiative CIS system should make.

Wadhwa and Zamani (2021) studied system-initiative CIS systems and discussed its challenges and opportunities. They introduced a taxonomy of system-initiative CIS systems by defining three orthogonal dimensions: (1) initiation moment (*when* to initiate a conversation), (2) initiation purpose (*why* to initiate a conversation), and (3) initiation means (*how* to initiative a conversation). They further identified five different purposes for initiating conversations for CIS systems, some of

which have been mentioned above: (1) filtering streaming information, (2) context-aware recommendation, (3) following up a past user-system conversation, (4) contributing to a multi-party human conversation, and (5) requesting feedback from users. Based on this taxonomy and conversation initiation purposes, they introduced a generic pipeline that is depicted in Figure 6.1. According to this pipeline, several algorithms are constantly monitoring the user's situation (user context) and the stream of generated information to produce conversation initiation instances. These instances are stored in a database which is constantly monitored by a conversation initiator component. Based on the situation, the initiator may select one of the initiation instances. Then, a fluent conversation will be initiated. For more information on this architecture, we refer the reader to Wadhwa and Zamani (2021).

6.2 Clarification in Information Seeking Conversations

Clarification is defined as "an explanation or more details that makes something clear or easier to understand."[3] In information seeking systems, it is often used to clarify the user's information need, and it can be in any form. For instance, relevance feedback is one form of clarification that is provided by the user. In mixed-initiative interactions, systems can take initiative to ask for clarification. This is why asking for clarification has been identified as a necessary component in developing ideal CIS systems (Radlinski and Craswell, 2017; Aliannejadi *et al.*, 2019; Anand *et al.*, 2020; Zamani *et al.*, 2020a; Trippas *et al.*, 2020). As pointed out earlier, subdialogue initiation is one of the four levels of mixed-initiative interactions in conversational systems, which involves *asking a clarification*. In a study of mixed-initiative collaborative planning in human conversations, clarification accounts for 27% of interactions, more than any other mixed-initiative interaction (Allen *et al.*, 1999). A conversational agent can ask a clarifying question to resolve ambiguity, to prevent potential errors, and in general to clarify user's requests and responses. Clarification may happen in multiple levels for various purposes. Stoyanchev *et al.* (2014) used clarification

[3] https://dictionary.cambridge.org/us/dictionary/english/clarification

for resolving ambiguity and uncertainty in speech recognition, while Aliannejadi *et al.* (2019) used clarification to identify query intent in a conversational search setting. Besides CIS systems, asking clarifying questions has been explored in various tasks. For instance, Rao and Daumé III (2018) used clarification for identifying missing information in a passage, such as community question answering posts. Trienes and Balog (2019) identified the community question answering posts that require clarification. Subsequent work by Tavakoli *et al.* (2021) studied properties of clarification in community question answering websites based on user responses. Asking clarifying questions has also been studied in the context of task-oriented dialogue systems which are mostly closed-domain (Krum *et al.*, 2005; Rieser and Moore, 2005). In the following subsections, we mostly focus on query intent clarification which is the most relevant type of clarification for information seeking systems.

6.2.1 A Taxonomy of Clarification Types

In the context of information seeking systems, clarification has been mostly studied in both synchronous and asynchronous information seeking scenarios.

For instance, Braslavski *et al.* (2017) studied clarifications asked in community question answering (CQA) websites as an example of asynchronous human-human information seeking conversations. They derived a taxonomy of clarification types for questions asked in CQA websites. The clarification types and their examples are reported in Table 6.2.

Later on, Zamani *et al.* (2020a) studied clarification in open-domain search systems by analyzing a large-scale query reformulation data collected from a commercial web search engine. This resulted in a clarification taxonomy for open-domain information seeking queries. Their taxonomy consists of four main categories (with no particular order) and a number of subcategories as follows:

- **Disambiguation:** some queries (or part of the queries) are ambiguous and could refer to different concepts or entities. Clarifying questions can be used to disambiguate the query intent.

Table 6.2: A taxonomy of clarification types for questions asked in CQA websites by Braslavski *et al.* (2017).

Clarification Type	Example
More Information	What OS are you using?
Check	Are you on a 64-bit system?
Reason	What is the reason you want a drip pan?
General	Can you add more details to this question?
Selection	Are you using latex or oil based Kilz?
Experience	Have you tried to update video card drivers?

- **Preference:** Besides disambiguation, a clarifying question can help identify a more precise information need. Four major subcategories of preference clarifications are:
  - Personal information ("for whom"): personal information, such as gender, age, language, and expertise, can limit the search space.
  - Spatial information ("where"): spatial information is also reflected in reformulations in many cases.
  - Temporal information ("when"): some queries have a temporal aspect which can be clarified by the system.
  - Purpose ("for what purpose"): if the answer to a query depends on the purpose of user, a clarifying question can seek to identify the purpose. For example, a user searching for "screwdrivers" may be interested in screwdrivers for different kinds of screws in different sizes, depending on the user's purpose.
- **Topic:** In case of broad topics, the system can ask for more information about the exact need of the user. This would narrow down the search space and would potentially lead to more accurate results. Topic clarification includes:
  - Sub-topic information: The user might be interested in a specific sub-topic of the query.

  - Event or news: based on an event or breaking news, many users often search for a topic related to the news, while the query may have different meanings out of the context of that event or news.

- **Comparison:** Comparing a topic or entity with another one may help the user find the information they need.

Note that clarifying the information need of a user may lie in multiple categories in this taxonomy. As mentioned earlier, this taxonomy was obtained based on web search query logs. Therefore, it can be considered as a taxonomy for open-domain information seeking queries. However, there may be other domain-specific types of clarification that are not easily recognizable in the web search query logs. For more information on this taxonomy, we refer the reader to Zamani *et al.* (2020a).

For all clarifying questions, we note that it is also essential to consider a system's need for specific information, with particular attention to personal or private information. As an example, while personal information such as gender or age may help a CIS system better answer a particular information need, is it clear to the user why this is being asked? Is it clear how this information will be processed and/or recorded? What would be the effect should the user decline to answer this question? While there are commonly accepted UI affordances for visual search systems (such as an asterix for required fields and hover-over information tags to provide background on questions), such affordances rarely exist in verbal modalities.

6.2.2 Generating Clarifying Questions

There exist three categories of solutions for generating clarifying questions: (1) selecting and filling out pre-defined question templates, (2) generating clarifying questions based on sequence-to-sequence modeling by maximizing the likelihood of generating the questions in a training set, and (3) generating clarifying questions by maximizing a clarification utility. In the following subsections, we briefly discuss solutions from each of these categories.

6.2.2.1 Template-based Slot Filling Models

Template-based slot filling is the simplest approach for asking a clarification. In this approach, a small set of question templates is first defined. The templates are task- and domain-dependent. For instance, Coden *et al.* (2015) simply used the question template "Did you mean _____ or _____?" for entity disambiguation. The question template "Did you mean _____?" has been widely used by various commercial search engines, such as Bing and Google, to clarify misspelling. Zamani *et al.* (2020a) listed a handful of question templates for search clarification. The question templates can be as generic as "What would you like to know about _____?". However, more specific questions, such as "What _____ are you using?" or "Who are you shopping for?" would be desired in most scenarios.

Once the question templates are defined, the task is to select one of the templates and fill it out. The template selection can be as simple as a rule-based algorithm or modeled as a machine learning problem, either as a multi-class classification or a learning to rank task. Similarly, rule-based solutions can be used to fill out the templates. For example, a substring of the user request or its entity type obtained from a knowledge base can be used to fill out some templates. Machine learning solutions are often preferred due to their superior performance for filling out the templates. Slot filling is not specific to clarification. A number of slot filling models used in task-oriented dialogue systems can be employed in clarification as well (Wu *et al.*, 2019; Budzianowski and Vulić, 2019; Zhao *et al.*, 2019).

6.2.2.2 Sequence-to-Sequence Models

As discussed in Rao and Daumé III (2019) and Zamani *et al.* (2020a), generating clarifying questions can be seen as a sequence generation task, in which the inputs are the query q and the context c and the output is a clarifying question q^*. The context here may refer to the query context, e.g., short- and long-term search or conversation history (Bennett *et al.*, 2012) and situational context (Zamani *et al.*, 2017), or some additional knowledge about the query, such as query aspects. Sequence-to-sequence models, including seq2seq (Sutskever *et al.*, 2014) and the Transformer

encoder-decoder architecture (Vaswani *et al.*, 2017), can be adopted and extended to address this task.

Sequence-to-sequence models consist of at least one encoder and one decoder neural network. The encoder model E takes the query q and the corresponding context c and learns a representation v for the input tokens. The decoder model D uses the encoder's outputs and generates a sequence of tokens, i.e., a clarifying question. The training objective is to maximize the likelihood of generating the clarification q^* by the decoder. This likelihood probability is computed as:

$$p(q^*|q,c) = \prod_{i=1}^{|q^*|} p(q_i^*|v, q_1^*, q_2^*, \cdots, q_{i-1}^*) \tag{6.1}$$

where q_i^* denotes the i^{th} token in the clarifying question q^*. In fact, this objective maximizes the likelihood of generating tokens in the clarification provided by the training data one by one. This maximum likelihood objective is equivalent with minimizing the cross-entropy loss.

Once the model is trained, it is used to generate the clarification as follows:

$$\arg\max_{q^*} p(q^*|q,c) \tag{6.2}$$

This decoding step can be achieved using beam search, its variants, or in the most simplest case, generating the clarification token by token until observing an `end` token. For more detail on sequence-to-sequence modeling, we refer the reader to Sutskever *et al.* (2014) and Vaswani *et al.* (2017).

It is widely known that training text generation models by maximizing likelihood of generating a ground truth output will result in frequent generation of the most common outputs. Thus, the models often suffer from generating diverse outputs. This has been addressed using different techniques, such as unlikelihood training (Welleck *et al.*, 2020) and F^2-Softmax (Choi *et al.*, 2020). Clarification utility maximization (next subsection) also implicitly addresses this issue.

6.2.2.3 Clarification Utility Maximization Models

An alternative to the presented sequence-to-sequence models that maximize the likelihood of generating clarification observed in the training set

is clarification utility maximization models. The intuition is to generate a question that best clarifies the user information need, while there is no notion of clarification in the training objective of sequence-to-sequence models.

In this approach, the goal is to maximize a clarification utility function U that measures the likelihood of clarifying the user information need or a similar objective. For instance, Rao and Daumé III (2019) estimated the information value of the possible answer that a user may give to the generated clarification as a utility function. Zamani *et al.* (2020a) estimated the likelihood of covering all information needs observed in the query logs based on the past interactions.

The clarification utility functions are often non-differentiable, which prevents us from using gradient descent based optimization. Therefore, clarification generation can be modeled as a reinforcement learning task whose reward function is computed based upon the clarification utility function U. The REINFORCE algorithm (Williams, 1992) can then be used for learning the clarification generation model. It has been shown that using the models that are pre-trained using maximum likelihood training for the REINFORCE algorithm can lead to more effective and more robust outcomes. This approach is called Mixed Incremental Cross-Entropy Reinforce (MIXER) (Ranzato *et al.*, 2016). For more information, we refer the reader to Rao and Daumé III (2019) and Zamani *et al.* (2020a).

6.2.3 Selecting Clarifying Questions

Clarifying question generation models can be evaluated using human annotations or online experimentation. However, both of these approaches are time consuming and are not always available. On the other hand, offline evaluation based on text matching metrics, such as BLEU (Papineni *et al.*, 2002) and ROUGE (Lin, 2004), are not reliable for clarification generation models. Therefore, due to the challenges in offline evaluation of clarifying question generation models, Aliannejadi *et al.* (2019) introduced the task of *selecting* clarifying questions from a set of candidate human- or machine-generated clarifying questions. The authors created and released the Qulac dataset, consisting of over

10K human-generated (through crowdsourcing) question-answer pairs for 198 topics associated with the TREC Web Track 2009-2012. An alternative dataset is MIMICS (Zamani *et al.*, 2020b) that contains over 450K unique real queries and machine-generated clarifying questions along with user engagement signals (*i.e.*, clickthrough rate).

Baseline models that use a combination of contextual representations of the query and clarifying questions (*e.g.*, BERT) and query performance prediction indicators (*e.g.*, standard deviation of retrieval scores) demonstrate the best performance on clarification selection tasks on Qulac (Aliannejadi *et al.*, 2019). Zamani *et al.* (2020c) showed that the clarifying question selecting model can benefit from query reformulation data sampled from search engine query logs. Subsequent work by Hashemi *et al.* (2020) proposed Guided Transformer, an extension to the Transformer architecture that uses external information sources (*e.g.*, pseudo-relevant documents) for learning better representations for clarifying questions. This model significantly improves upon the baseline models for clarification selection tasks. Specifically, they showed that the model performs well for clarifications with short negative responses. Subsequently, Bi *et al.* (2021) focused on a BERT-based model for clarification selection based on negative feedback. This model works well for document retrieval when clarifying questions are asked. Both clarifying question generation and selection tasks are still active areas of research in both the IR and NLP communities.

6.2.4 User Interactions with Clarification

The way users interact with clarification can reveal information on the clarification quality. For example, user engagement with clarifying questions can be studied as a proxy to measure clarification quality. Zamani *et al.* (2020c) studied how users interact with clarifying questions in a web search engine. They found out that more specific questions have a higher chance to engage users. They showed that the majority of engagement comes for one of two reasons: (1) high ambiguity in the search queries with many resolutions, and (2) ambiguity but where there is a dominant "assumed" intent by users where they only realize the ambiguity after issuing the query. Interestingly, users are more likely to

interact with clarification in case of faceted queries in comparison with ambiguous queries. Note that the user interface may affect these findings. For instance, in the web search interface with ten blue links, users can simply skip a clarification and directly interact with the retrieved web pages. However, this may not be possible in a conversational search system with a speech-only interface. Therefore, besides generating high-quality clarifying questions, (spoken) CIS systems should make a (binary) decision at every step on whether to ask a clarifying question or to show the result list or answer. Wang and Ai (2021) addressed this issue by developing a risk-aware model that learns this decision-making policy via reinforcement learning. Their model considers the common answers to each clarification in order to minimize the risk of asking low-quality or out-of-scope clarifications. The model enables the CIS system to decide about asking a clarification with different levels of user tolerance.

In a separate line of research, Tavakoli *et al.* (2021) studied user interactions with clarifying questions in asynchronous conversations. They focused on user interactions in community question answering websites, e.g., StackExchange.[4] To study user interactions, they categorized clarifying questions to three categories: (1) clarifications that have been answered by the Asker (the person who submitted the questions/post), (2) clarifications that have been answered but not by the Asker, and (3) clarifications that are left unanswered. They found that clarifications with the goal of disambiguation account for the majority of clarifying questions and they are very likely to be answered by the Asker. On the other hand, clarifications with the goal of confirmation are more likely to be left unanswered. For more analysis on user interactions with clarification in asynchronous information seeking conversations, refer to Tavakoli *et al.* (2021).

6.3 Preference Elicitation in Conversational Recommendation

Preference elicitation in conversational recommender systems forms another type of mixed-initiative interactions. Typically, recommender

[4] https://stackexchange.com/

systems create a user profile or user representation based on the user's past interactions (e.g., click) (Jannach *et al.*, 2018; Oard and Kim, 1998) and/or her explicit feedback on items using ratings and reviews (Resnick *et al.*, 1994; Ricci *et al.*, 2010). Conversational systems enable recommendation engines to ask for user preferences in a natural language dialogue. This creates a significant opportunity for the system to learn more about the current context of the user, and how their preferences at this point in time may differ from their preferences in general. Christakopoulou *et al.* (2016) studied the task of conversational recommender systems by focusing on preference elicitation in a closed-domain scenario, like restaurant recommendation. They observed 25% improvements over a static model by asking only two questions. Following their work, Sun and Zhang (2018) proposed a reinforcement learning model for preference elicitation by asking questions about item facets in a closed-domain setting, i.e., restaurant recommendation. Zhang *et al.* (2018) focused on a broader domain by automatically extracting user preferences about item facets from user reviews on an online e-commerce website. They showed that multi-memory networks can be successfully used for asking questions about item facets in their setting. Sepliarskaia *et al.* (2018) used a static questionnaire to ask questions from users in the context of movie and book recommendation. They studied different optimization strategies for the task with a focus on cold-start users. In this work, user responses to the system questions are automatically generated and may be different from real-world settings. To mitigate this issue, Radlinski *et al.* (2019) conducted a crowdsourcing experiment with a wizard-of-oz setting, where a crowdworker plays the role of user and another person (i.e., assistant) plays the role of the system. They introduced a "coached" preference elicitation scenario, where the assistant avoids prompting the user with specific terminology. Preference elicitation in recommendation is tightly coupled with the design of conversational recommender systems. Refer to Section 5.5 for further information.

6.4 Mixed-Initiative Feedback

The system can take advantage of mixed-initiative interaction to get feedback from users and even give feedback to them. For instance, in the

middle (or at the end) of a dialogue in a conversational recommender system, the system can ask for explicit feedback from the users. Existing systems often have a static pre-defined questionnaire that will automatically be triggered after a conversation ends. For instance, the Alexa Prize Challenge (Ram *et al.*, 2018) has sought explicit rating feedback from users upon the completion of the conversation and used average ratings for evaluating the participant teams. This simple approach can be further improved by asking context-aware questions for feedback and making natural language interactions within the conversation.

Mixed-initiative feedback can be also relevant to the concept of "grounding as relevance feedback" introduced by Trippas *et al.* (2020). Grounding is defined as discourse for the creation of mutual knowledge and beliefs. The authors demonstrated grounding actions in a spoken conversational search data, such as providing indirect feedback by reciting their interpretation of the results. This grounding process can potentially enable CIS systems to better understand a user's awareness of the results or information space.

As mentioned earlier, mixed-initiative interaction can be used to give feedback to the users. As an emerging application, users may not directly know how to effectively use the system. Hence, the system can take advantage of this opportunity to educate users on the system capabilities. Educating users on interacting with CIS systems has been relatively unexplored.

6.5 Summary

In this section, we discussed the opportunities and challenges that mixed-initiative interactions bring to CIS systems. We drew connections with mixed-initiative user interfaces and mixed-initiative interactions in dialogue systems. We discussed system-initiative CIS and reviewed different purposes for conversation initiation. We also provided an overview of clarification in CIS systems and how a clarifying question can be generated or selected to identify the user's intent. We briefly reviewed preference elicitation and demonstrated its connections with intent clarification. We finished by showing how systems can get feedback from and give feedback to the users through mixed-initiative interactions.

Overall, understanding mixed-initiative interactions and initiating conversations have been identified as a key part of CIS research. Clarification, as a form of mixed-initiative interaction, has been studied quite extensively. However, other forms of mixed-initiative interactions require further significant efforts. Evaluating mixed-initiative CIS systems is another under-explored yet important research area.

7

Evaluating CIS Systems

Evaluation of conversational information seeking systems continues to be a rapidly evolving research area due to unique challenges of assessing the quality of *conversations*, and the parallel difficulty in creating benchmark datasets.

In contrast to non-conversational information seeking settings, the multi-turn nature of conversations requires evaluations to model long-term state, and consider *what* information is conveyed, *when* the information is conveyed, as well as *how* this communication happens. All these are dependent on *why* a user is engaging in a conversational interaction in the first place (as opposed to non-conversational alternatives). The same conversation may be considered of high or of low quality depending on context: For example, if a user is in a rush or not, or if the user requires high confidence in the conclusion or not.

7.1 Categorizing Evaluation Approaches

There are a number of ways that CIS evaluation may be presented. We structure Section 7 by evaluation modality: Offline or online evaluation, and sub-types of these modalities.

However, evaluation approaches can be broken down in other ways

(see Chapter 4.2 of Anand *et al.*, 2020). For example, individual components of conversations can be evaluated at a micro-level, leading to a catalogue of micro-evaluation techniques including *How well does the system predict the dialogue act of a given utterance? How well does the system predict the user's goals and sub-goals? Can the system identify terms in statements to fill slots in a structured search query? How well does the system select responses from a set of candidates? How well does the system answer individual questions?* As we shall see, such an evaluation approach has the benefit that these questions lend themselves well to traditional information retrieval evaluation approaches. A major drawback, however, is that high performance on micro-level metrics does not necessarily translate into a CIS system being effective for satisfying users' needs.

An alternative is to break down by evaluation approaches in a user-centric manner: *Does the user trust the system? What is the cognitive load of interactions? How fluent and efficient is the system in communication in general?* Within the context of a particular information need, one can seek metrics to evaluate based on properties such as *Is the user satisfied with the outcome of the conversation? How much effort and/or time was required to satisfy the information need? Is the information need ultimately resolved? Was the user frustrated in the process?* For such metrics, subjectivity is a common concern. Additionally, while such evaluation does assess the overall quality of a CIS system, such metrics are particularly difficult to optimize.

7.2 Offline Evaluation

As a staple of information retrieval evaluation, offline evaluation permits reproducible evaluations that can reliably compare different systems. We start with a discussion of some of the existing datasets commonly used to evaluate CIS systems. Following a summary of each category of dataset, we present open challenges with respect to offline evaluation of CIS tasks.

7.2.1 Conversational Datasets

Conversational datasets are transcripts of actual conversations that have occurred between two or more parties, either as part of natural information seeking or through a role-play conversation exercise.

We begin by observing that some conversational datasets are *synchronous* (*e.g.*, Budzianowski *et al.*, 2018), while others are *asynchronous* (such as datasets derived from Reddit (Henderson *et al.*, 2019)). Although, in principle, the content of these can be similar, subtle timing effects can lead to meaningful practical differences. For instance, asynchronous conversations may contain fewer disfluencies and unintentional errors as participants take time to consider their utterances (Serban *et al.*, 2018). Asynchronicity also makes it possible to carry out time-consuming tasks such as consulting external sources between conversational turns. Of particular importance to studies of mixed initiative, the role of initiative and conversational turn taking is very different in synchronous and asynchronous conversations (Gibson, 2009; Boye *et al.*, 2000).

An example of a widely used conversational dataset is Multi-WOZ (Budzianowski *et al.*, 2018). Consisting of synchronous naturalistic task-oriented dialogues designed to simulate a possible conversation between a tourist and information agent, it focuses on recommendation tasks with well-defined slots and values. To create these, one person is presented with search criteria, while a second ("wizard") has access to a search system that allows them to identify recommendations that satisfy the "user's" constraints. However, by presenting such specific requirements that perfectly match the wizard's known fields, it may be argued that the conversations can be somewhat unnatural. The TaskMaster dataset (Byrne *et al.*, 2019) generalizes on the Multi-WOZ idea, with dialogues around making orders and setting up appointments, such as ordering a pizza or creating an auto repair appointment. In addition to synchronous Wizard-of-Oz dialogues similar to those from Multi-WOZ, the authors also include asynchronous self-dialogues where a single person types both sides of a conversation, focusing on given needs. To make the conversations more natural, the authors also instructed raters to intentionally include understanding errors and other

types of dialog glitches, with some conversations created to be intentionally unsuccessful. This type of dataset is predominantly used for the evaluation of slot-filling algorithms. As an alternative to task-oriented dialogues, Radlinski *et al.* (2019) presented Coached Conversational Preference Elicitation, intending to obtain realistic synchronous dialogues by instructing a "wizard" to simply motivate a "user" to describe their preferences, without setting a detailed goal for either.

Another category of conversational datasets is used for conversational question answering (Iyyer *et al.*, 2017; Choi *et al.*, 2018; Reddy *et al.*, 2019) or TREC CAsT Track (Dalton *et al.*, 2019; Dalton *et al.*, 2020b). Here the major challenge addressed is co-reference resolution, evaluating the systems ability to answer questions in sequence, particularly when a given question may refer to earlier questions or their answers (for example, "Who won the superbowl?" followed by "Who is their quarterback?"). Such dialogues can be sampled from search engine interactions, known answers, or manually constructed.

Two more types of conversational datasets are commonly used in developing CIS systems. Asynchronous discussions on a given topic, typically from the Reddit forum (for example, Henderson *et al.*, 2019; Qu *et al.*, 2018), are often used to model open-ended conversations. As a massive corpus of free-form dialogues, these exchanges can be used to train and evaluate conversational agents with a goal of responding reasonably to any utterance on any topic without an assumption of a particular task. Of course, it is important to note in the context of web forums that careful attention must be paid to the representativeness of the authors of the corpus being used. For instance, training or evaluating CIS systems based on a forum with a particular type of contributor may lead to bias in a CIS system evaluation, and may lead to undesirable conversational behaviors being learned if they mirror the behavior of the authors who contributed to that forum. For instance, language ranging from microaggressions to insults or worse is often observed (Bagga *et al.*, 2021). For this reason, the use of massive web corpora must be done with care. To obtain open-ended synchronous conversations with higher quality than may be expected in an open forum, transcripts of movie and television dialogues are frequently used (Müller and Volk, 2013; Henderson *et al.*, 2019).

There are numerous challenges in creating and using conversational datasets for offline evaluation. One of the key challenges is that the motivation of the participants can greatly influence the dialogues observed. In a Wizard-of-Oz setting, if the wizard is provided with a particular interface to obtain answers for user requests, this is likely to influence their utterances (Radlinski *et al.*, 2019). If the user is given detailed instructions, especially if these do not align with the person's actual interests, this again can result in unnatural dialogue (Serban *et al.*, 2018). If several Wizard-of-Oz datasets are used together for evaluation, they may uncover slight differences in the study setup impacting the conversations (Trippas and Thomas, 2019). Moreover, if users are asked to complete predefined tasks, there is a risk that they do not approach these tasks as someone who actually *wants* to perform that task (Serban *et al.*, 2018). For example, suppose a user is tasked with purchasing something under a given price. A real user may exhibit certain flexibility regarding the price, or may ask questions relating to value for money, rather than solely around price – and examples of realistic behavior around pricing may end up missing from the collected corpus. A second major challenge in evaluation with offline datasets lies in how the datasets are interpreted. Where dialogues are taken to contain *correct* responses in a particular context, we can suffer from false negatives: A perfectly capable system may be judged to perform poorly when it is simply performing the task *differently* (Finch and Choi, 2020).

7.2.2 Single-Step Datasets

As a step towards fully conversational systems, a number of challenges have been proposed to address the necessary sub-tasks. Here we refer to them as single-step datasets, as the focus is on a single step within the many that a conversational system must perform. We note that they do not focus on *single dialogue turns* (as is the case with Conversational QA datasets), but even more fundamental steps of information processing.

One recent example is generating the natural text from structured information to describe a particular search result, as the conversational equivalent of search snippet generation (Turpin *et al.*, 2007). For instance, suppose a conversational agent needs to explain a specific

restaurant to a user, showing how it satisfies their request. The agent may possess rich structured information about the restaurant – its name, address, the type of food offered, pricing information, and other key attributes. However, just presenting these facets of information to the user may not be suitable. The End-to-End NLG Challenge (Dušek *et al.*, 2018) produced a dataset mapping a set of attributes to natural language descriptions, allowing a challenge for generating text from structured information – a critical single step of many CIS systems.

A second example where single-step datasets are used is for applications where generating full text is unnecessary. This common task treats conversational information seeking as the ranking of possible (existing) responses that an agent could give at a particular time. For instance, Yang *et al.* (2018a) described datasets derived from transcripts of past technical support dialogues: They assume that for any given user utterance, the system should select from previous agent utterances (as most technical support problems are not novel). Such specialized single-step datasets will address this single-turn ranking problem.

As a third example, when an agent asks a question, it must be able to interpret the user's answers. Taking the seemingly simple case of yes/no questions, a user may answer indirectly. For instance, if an agent asks if a user would be interested in an evening activity, the user may say "I'd prefer to go to bed" rather than simply "no". The Circa dataset (Louis *et al.*, 2020) was developed to contain natural questions and answers to train and evaluate reliable answer interpretation by CIS systems. The approach used multiple phases of crowdworker tasks first to develop natural questions and then, in turn, natural answers while attempting to minimize bias and maximize the diversity and naturalness of answers.

7.2.3 Simulated Users

A recent alternative to static conversational datasets is relying on simulators (Zhang and Balog, 2020; Ie *et al.*, 2019; Aliannejadi *et al.*, 2021a). For instance, Zhang and Balog (2020) argued that a simulator "should enable to compute an automatic assessment of the agent such that it is predictive of its performance with real users". In this way, rather

than evaluating with a fixed dataset, an agent could be assessed dynamically against a (fixed) simulator to obtain the benefits of effective offline evaluation. This recent work showed a high correlation between simulation-based evaluation and an online evaluation approach. Simulation also addresses challenges in fixed datasets, particularly relating to user privacy (Slokom, 2018).

7.2.4 Datasets Beyond the Text

Several authors have considered evaluating CIS tasks beyond simply the text of interactions between a user and a CIS system. Typically this involves additional annotation of the conversational dialogue to indicate relevant aspects, although it can also involve other content modalities in addition to the conversation.

One example is the annotation of the high-level role of individual utterances. This may be at the level of sentences within a conversation (annotated as to whether they are asking a question, sharing an opinion, thanking, or so forth) (Yu and Yu, 2019), or may be at the level of the high-level structure of conversations as in the case of sub-goal or sub-task prediction. Alternatively, user-centric metrics may be annotated, such as indicators of customer frustration at specific points in customer service conversations (Oraby *et al.*, 2017). Note that these evaluation annotations are in contrast and complementary to datasets which have been annotated to investigate **how** interactions between the user and CIS system are structured (Vakulenko *et al.*, 2021; Trippas *et al.*, 2020).

A fundamentally different type of CIS dataset involves multiple modalities. The conversation may include both text, images, or gestures to illustrate the user's need in a recommendation setting (Nie *et al.*, 2019; Deldjoo *et al.*, 2021), or even include navigation within a virtual or physical environment as part of the conversational task (Ku *et al.*, 2020).

7.3 Online Evaluation

In contrast to offline evaluation, CIS systems may also be evaluated *online*: deploying a system that real users interact with, dynamically

obtaining user utterances and the system's responses.

Online evaluation allows systems to be evaluated much more robustly, as the consequences of earlier system actions can be seen in how users respond, which in turn determines what options the system has and how these are handled. In this way, online evaluations are much more predictive of real-world system performance, and is more likely to identify limitations in current solutions. Online evaluation can be done in one of two ways.

7.3.1 Lab or Crowdsourced Studies

It is often desirable to evaluate components of a system that is not end-to-end complete (such as when developing specific aspects of a CIS system), or where it is necessary to control certain conditions (such as when performance for specific use cases is of particular interest). In this situation, paid users or volunteers are often employed.

For instance, Christakopoulou *et al.* (2016) studied different approaches for eliciting user preferences in a restaurant recommendation setting. As the authors' goal was to assess how well different ways of asking questions efficiently established users' interests, the authors chose to perform a lab study. Participants were presented with preference questions that a conversational system might want to ask. The results were used to inform algorithms for learning about users interests. This type of evaluation was appropriate as the information could not be collected through an offline corpus (as rating data in offline studies is usually incomplete), nor in a real-world system (as preference elicitation studied here is but one part of the overall CIS recommendation challenge).

Similarly, Aliannejadi *et al.* (2019) introduced a crowdsourced approach for evaluating clarification question selection. They started with a variety of queries, crowdsourced a collection of possible clarifying questions, then collected possible answers to these questions. Despite simplifying assumptions, the approach allowed a clarifying question selection model to be evaluated based on the retrieval performance, giving possible answers to the system's potential questions. For the same task, Zamani *et al.* (2020b) provided guidelines for manual an-

notation of clarifying questions and their candidate answers based on their fluency, grammar, usefulness for clarification, comprehensiveness, coverage, understandability, diversity, coherency, and so forth.

Evaluating a different aspect of CIS behavior, Balog and Radlinski (2020) studied the role of explanations in recommendation tasks. As one may expect explanations of results presented to be part of CIS, the authors focused on assessing what constitutes a valuable explanation. Using a non-conversational approach, crowdworkers were first asked to express their preferences in a given domain. They were then presented with recommendations along with explanations. These explanations were assessed using a focused questionnaire addressing different reactions the participants may have to the explanations.

As another example, Jiang *et al.* (2015) recruited participants to complete specific tasks using a well established CIS agent, including specific search tasks. After the tasks were complete, participants were asked to answer specific questions about their experiences. Based on the answers to these questions and a record of the participants' interactions with the CIS system, the authors developed an automated approach for predicting satisfaction and natural language understanding.

As these examples show, controlled studies can allow investigation of the performance of particular aspects of CIS. A detailed treatment of designing user studies for interactive IR systems is presented by Kelly (2009).

7.3.2 Real-World Studies

When a CIS system is complete and a fully realistic evaluation of the users' overall experience is desired, a real-world study is the gold standard. This involves observing actual natural interactions between users and the CIS system, particularly with users motivated by relevant information needs. On one hand, users bringing their own information needs lead to more realistic interactions. On the other hand, as such an evaluation depends on actual interactions, only limited feedback from users may be available.

As an example of such a study, Park *et al.* (2020) presented a study of a commercial CIS agent where, for some intents (such as asking for

weather), the agent asked for user feedback. In particular, the agent asked users "Did I answer your question?". Responses to this question were used to assess the quality of the end-to-end CIS system.

A similar approach is used in the Alexa Prize Challenge (Ram *et al.*, 2018). Here, real users may request to interact with a conversational system. At the end of the interaction, the user is asked to rate their experience. Such online evaluation can assess the quality of the conversational abilities of the system according to predetermined criteria (here, user-declared satisfaction, and level of engagement based on time spent).

7.4 Metrics

Having considered evaluation approaches, here we briefly discuss an essential aspect of CIS evaluation separately, namely that of metrics. While a complete treatment of metrics suitable for conversational information seeking is beyond our scope, we provide a high-level overview of the metric types used in different cases, and some of the appropriate considerations that are required when determining the right ones. We refer the reader to Liu *et al.* (2021) for a more extended treatment of conversational systems' single-turn and multi-turn metrics.

7.4.1 Metrics for Individual Steps

At individual steps, it is possible to evaluate whether the system understood a user's utterance, whether the search system respected a constraint, or whether a system utterance was fluent among other things. Most often, these aspects can be measured with metrics that can be computed offline.

As an example, we take conversational question answering (ConvQA), discussed in depth in Section 5.1. Often used to assess clarification approaches in the NLP community (Rao and Daumé III, 2019), common metrics include BLEU (Papineni *et al.*, 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). At a high level, these metrics match the similarity between a given string and reference strings. While effective for some applications, these metrics do not correlate highly

with user satisfaction in conversational systems (Liu *et al.*, 2016). More recently, machine learned metrics have achieved significantly higher correlation with manual human ratings for such language tasks (Ma *et al.*, 2018; Sellam *et al.*, 2020).

When assessing the relevance of recommendations that terminate a conversational exchange, classic information retrieval metrics are used (Croft *et al.*, 2010). For instance, Normalized Discounted Cumulative Gain (nDCG), Mean Reciprocal Rank (MRR), and Precision are often used to assess if recommended items match information needs of users given a particular user representation, *e.g.*, (Christakopoulou *et al.*, 2016), if a system is able to rank possible clarifying question, *e.g.*, (Aliannejadi *et al.*, 2019), or if a system accurately provides answers to specific requests, *e.g.*, (Christmann *et al.*, 2019). As with language metrics, such metrics do not necessarily agree with user experience of an end-to-end system (Jiang and Allan, 2016).

As an example of more nuanced refinements of relevance in a conversational setting, consider work by Rashkin *et al.* (2021). Here, the authors propose a metric that assesses whether a CIS system only presents verifiable information, rather than hallucinated or factually unverifiable information.

7.4.2 Metrics for End-To-End Evaluation

An essential characteristic of conversational information seeking systems is the multi-turn nature of conversations. As such, it is vital that evaluation considers an end-to-end interaction. For example, consider catastrophic failure in the middle of a long conversation, where an agent may lose the state after a user has provided significant information to a CIS system. Kiseleva *et al.* (2016) showed how one failure in a more extended conversation often leads to dissatisfaction. This can happen even if the vast majority of individual conversational steps are successful.

The richness of conversational interactions thus means that CIS systems can be assessed along many different dimensions. Trivially, one may consider whether users were successful at their task (Chuklin *et al.*, 2018; Dubiel *et al.*, 2018) or achieved success quickly (Thomas *et al.*,

2018; Trippas *et al.*, 2017). Despite this, a shorter time to success is not necessarily sufficient. For instance, in a non-conversational recommendation setting, Schnabel *et al.* (2016) showed that more successful recommendations may be obtained using systems that require more prolonged user interactions, leading to overall higher user satisfaction. In conversational settings, a system may trade-off long-term and short-term utility (Radlinski and Craswell, 2017). It is important to note that it is also possible to succeed while leaving users frustrated, as studied by Feild *et al.* (2010). A particular end-to-end evaluation approach was recently presented by Lipani *et al.* (2021), based on the flow of different subtopics within a conversation.

Two other classes of metrics are critical to consider. First, *trust* between the user and a CIS system. For example, Daronnat *et al.* (2020) studied how trust affects users satisfaction. Trust usually requires factuality. It has been noted that some modern neural conversational systems can produce utterances that are false (often termed hallucinations). A detailed treatment of hallucination, and references to further work, can be found in Shuster *et al.* (2021). Trust may also be affected by explanations being incorporated into CIS systems. Explainable AI is, in general, an extremely active and critical area of study (Adadi and Berrada, 2018). In a conversational recommendation setting, explanations have recently received attention as well, for example, see (Chen *et al.*, 2020b; Balog and Radlinski, 2020).

The second critical concept to consider in CIS systems is that of *fairness*. While often not treated as a key metric of effectiveness, many researchers have recognized this as a principal desirable aspect of AI systems in general and recommendation systems in particular. A CIS system that provides recommendations in the course of a conversation, for instance, may aim to do so in a fair manner. Thus biases that may be present within the conversational system warrant careful consideration. We refer interested readers to Beutel *et al.* (2019) and Ge *et al.* (2021) and their citations for definitions, approaches and relevant metrics.

7.5 Summary

This section presented an overview of key concepts in the evaluation of conversational information seeking systems. We provided an overview of offline as well as online evaluation techniques, discussing common methodologies in both cases. Benefits and drawbacks of the two high-level approaches were discussed. Finally, we provided an overview of common metrics used to evaluate CIS systems, as well as references to broader topics that should be considered when measuring the performance of CIS systems, such as trust and fairness.

8

Conclusions and Open Research Directions

The goal of this survey was to provide an overview of research in the area of conversational information seeking (CIS), summarizing current research and presenting an introduction to researchers new to this area. We addressed CIS from both a user- and system-centred approach, with the aim of not singling out one view but providing a holistic outlook. CIS could be naively approached as a linear and straightforward pipeline of all components such as user input (*e.g.*, automatic speech recognition), which transcribes the spoken query as input, information retrieval, which identifies and retrieves the relevant items to the query, or information visualization, which presents the found information to the user. However, many more components are needed to make CIS truly conversational, including features that can capture and utilize interaction and preference history, adapt results presentations to the user's need or context, or track the conversation flow in long-term representations. Indeed, we argue that the interconnectedness of all the CIS building blocks are intrinsically interrelated and needs to be investigated beyond the sum of its parts. Furthermore, we show that CIS is more than system evaluation and retrieval effectiveness but requires a broad range of techniques.

CIS is a new interaction paradigm beyond the basic query-response

approach. This means that existing knowledge and assumptions of traditional IR may be challenged, reviewed, and expanded. Furthermore, CIS research aims to investigate and develop systems that users will use and perceive as genuinely helpful. The more users will interact with CIS systems, the more information on supporting users in their information seeking may become apparent. A such, creating more usable CIS systems will help users adopt and adapt conversational and intelligent methods to search for information.

Current research mostly makes several simplifying assumptions about user interactions and system capabilities. Given these assumptions, this monograph showed that large-scale pre-trained language models have many applications in developing different part of CIS systems that deal with natural language, *e.g.*conversational search, question answering, preference elicitation, and clarification. However, deciding about the interaction type, modality, initiative, explanation, etc. involves many components that need to work cooperatively with these language models for optimal dialogue understanding and generation.

We provided an overview of evaluation methodologies for CIS research. Due to the interactivity nature of CIS systems, developing reusable datasets and user simulation strategies for model training and offline evaluation is incredibly important and challenging. Again, most existing benchmarks and evaluation methodologies suggest numerous simplifications to the CIS tasks. Currently, online evaluation and collecting human annotations are the most robust and reliable approach for evaluating CIS systems.

It can be challenging to negotiate all different components for CIS, being ethical and rigorous in the research while maintaining a vision of an information system that does not hinder access to information. We hope that the overview of the broad range of research topics within CIS reflect the various research disciplines that should be part of the conversation studying and developing CIS.

8.1 Open Research Directions

Many open questions and directions of research have been mentioned throughout this monograph. In this section, we bring many of them

together with the aim of providing a unified starting point for researchers and graduate students currently investigating conversational information seeking. While not intended to be exhaustive, we believe that these key areas for future work are particularly likely to have deep impact in the field. The content of this section can be seen as complementary to directions suggested by the recent Dagstuhl Seminar on Conversational Search (Anand *et al.*, 2020) and the SWIRL 2018 report (Culpepper *et al.*, 2018).

Although some of these topics could be grouped under multiple headings, we divide this section into four main topics, (1) *modeling and producing conversational interactions*, which covers the foundation of conversational systems to understand and produce user-system interactions and the information transfer between them, (2) *result presentation* with different interaction modality and devices, (3) *types of conversational tasks* that are mostly under-explored, and (4) *measuring interaction success and evaluation*, with a focus on interactivity, ethics and privacy in conversational systems, and lastly looking at evaluation as a more extensive and broader topic than measuring success.

8.1.1 Modeling and Producing Conversational Interactions

Interactivity, the process of two or more agents (human or machine) working together, is a crucial characteristic of information seeking. Modeling interactions and deciding the next action or interaction is at the core of CIS research. In this context, even though much research has been devoted recently to **mixed-initiative interactions**, most mixed-initiative strategies have not been fully explored. In fact, our understanding of when a system can take the initiative without disrupting the flow of natural information seeking conversation needs significant further exploration. We believe that systems should more accurately identify opportune moments to initiate the conversation, introduce new topics, or support disambiguation. Similarly, the ability for systems to model **uncertainty** in user needs requires further study to effectively and efficiently clarify needs. We argue that supporting all these interactions will enhance the user experience, enable improved information seeking interactions, and thus positively impact this collaborative

process.

On top of the aforementioned open research directions for interactions, **long-term conversational interactions** may need specialized attention. In general, when investigating CIS it is often assumed that the user is interacting with the system *only* at the time of information need. However, supporting users in long-term information needs, be it multi-session tasks or the ability for a conversation to be continued and repeated at a much later date, need further research. This implies that the history and memory of conversations may be stored and used in future user-system interactions. Thus, further research needs to be done on how users want this *memory* to work, including **privacy and transparency** of what is stored and how the system retrieves and identifies relevant past interactions responsibly.

8.1.2 Result Presentation

Presenting results which the user can incorporate into their personal "knowledge space," and how the user interacts with them, can be seen as part of a broader challenge of information transfer. Results presentation has not received commensurate attention in the CIS research community relative to its impact. This includes what and how the information needs to be presented. Furthermore, with the increased interest in **multi-modal and cross-device CIS**, further research on when, how, and on which device users want to receive information is crucial. Questions such as how CIS systems can/should use **sensor data** to optimize result presentation is an open problems (*e.g.*, if a user is close to a screen, instead of using a smart speaker, should the information be presented visually?). As part of results presentation, further research on interactions between multiple devices will be pivotal. Thus, research on how to include more user context to predict how the user is going to interact with the available devices is warranted.

8.1.3 Types of Conversational Information Seeking Tasks

Many users will have different goals for why they engage in CIS tasks, with these varying based on the time, context and social situation of the need. Supporting each user's goals means recognizing these differences.

For instance, users interacting with a CIS may choose this mode of search to seek advice, look for a detailed summary of a complex topic, or verify a fact. Developing CIS systems that can **integrate different kinds of information seeking tasks** and produce a human-like dialogue needs to be better understood. Further, different scenarios or settings may require distinct ways of interaction. For instance, searching for information in enterprise settings contrasts with "every day" search. Conversations may also be structured differently, depending on the number of actors in the CIS process, thus making **collaborative CIS** an important topic for further exploration.

8.1.4 Measuring Interaction Success and Evaluation

We distinguish measuring success from evaluation to emphasize the importance of interactivity in CIS. Even though interactivity has always been a major part of information seeking, interactivity becomes even more critical with the paradigm shift from the basic query-response approach to CIS. For example, further research is needed to investigate a more robust **definition of success in CIS** across different user populations, contexts, and modalities. Thus highlighting the difficulty of defining success since it is changeable depending on the context or modality. Furthermore, that success may be incredibly personal, and metrics are only helpful when measuring what is desirable for a particular user. As such, further research on **personalized evaluation** of CIS is needed.

Another option of measuring interaction success includes keeping track of how well a user has understood the system and vice versa. This measurement enables systems to increase their capability to explain answers identified, and increase the confidence of successful information transfer. As already mentioned in Section 8.1.3, this kind of conversation calibration can help with the **transparency** of how systems are (1) retrieving, presenting, or augmenting information, (2) handling references to past user-system interactions beyond the current session, or (3) aiming to mitigate potential biases generated during the CIS process.

Furthermore, much more work on evaluating the entire CIS process

is needed beyond measuring interaction success. Indeed, evaluating any highly interactive process is challenging. Recently, many reusable test sets have been created for CIS tasks. However, many of these make simplifying assumptions about user behaviours and system capabilities. This means that ongoing efforts on **creating datasets that enable training and offline evaluation** are needed. However, to compare datasets and ensure researchers are producing more useful corpora in general, **dataset protocols** are desirable. We observe that the ongoing trend in **simulating users** could be helpful here.

Tools for dataset creation and evaluation may also benefit from further research effort. For instance, many researchers build their own wizard-of-oz frameworks with limited reuse capabilities. Other tools which are more intuitive to use, similar to WYSIWYG (What You See Is What You Get) website builders to test particular interactions, may accelerate all research in CIS.

Acknowledgments

We would like to thank the researchers who contributed directly or indirectly to the field of conversational information seeking. We are also grateful to our colleagues, W. Bruce Croft, Reena Jana, and David Reitter, who gave us invaluable suggestions and helped us better position this monograph. We also thank the anonymous reviewers and the editors of Foundations and Trends in Information Retrieval, especially Maarten de Rijke for his support throughout the process.

This work was supported in part by the Center for Intelligent Information Retrieval, in part by an Alexa Prize grant from Amazon, and in part by an Engineering and Physical Sciences Research Council fellowship titled "Neural Conversational Information Seeking Assistant" grant number EP/V025708/1. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

Appendices

A

Historical Context

In this appendix, we briefly provide a historical context to information retrieval and dialogue systems research related to conversational information seeking systems. Readers that are not familiar with early IR research are especially encouraged to read this appendix.

A.1 Interactive Information Retrieval Background

Conversational information seeking systems have roots in interactive information retrieval (IIR) research. The study of interaction has a long history in information retrieval research, starting in the 1960s (Kelly and Sugimoto, 2013). Much of the earlier research studied how users interacted with intermediaries (*e.g.*, librarians) during information seeking dialogues but this rapidly shifted to studying how users interacted with operational retrieval systems, including proposals for how to improve the interaction. Information retrieval systems based on this research were also implemented. Brooks and Belkin (1983) studied information seeking dialogues between a user and an intermediary and introduced a annotation coding scheme for discourse analysis of the dialogues.

Oddy (1977) developed an interactive information retrieval system with rule-based dialogue interactions in 1977, called THOMAS. Ex-

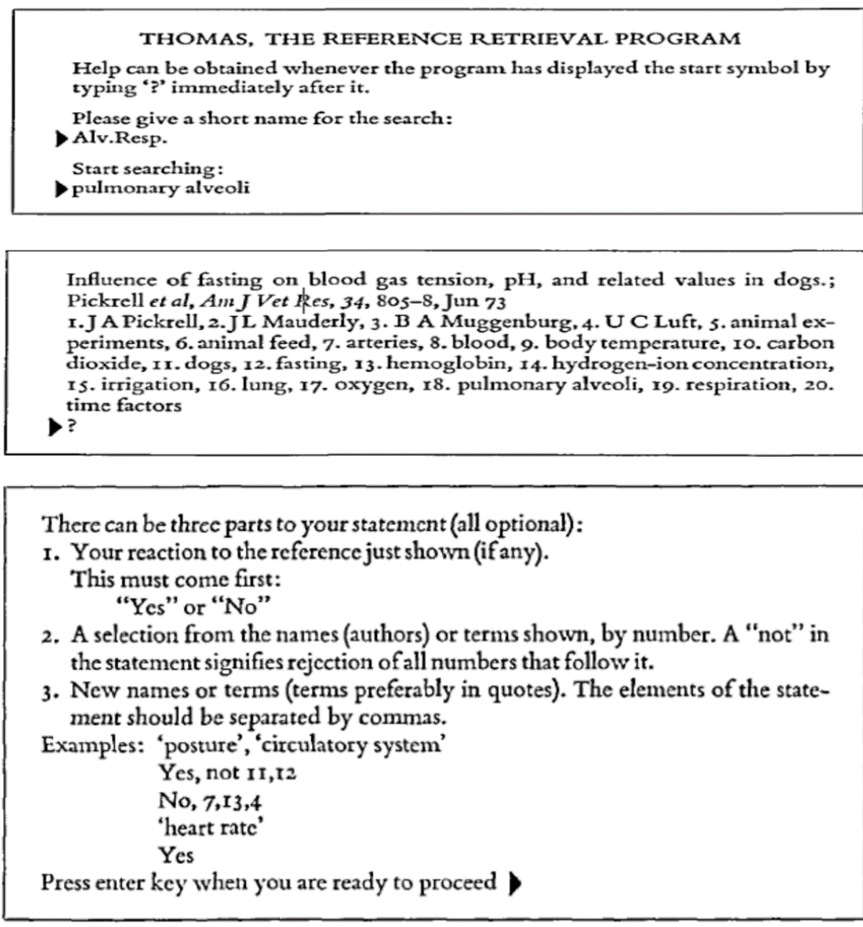

Figure A.1: Snapshots from the THOMAS system (Oddy, 1977).

ample snapshots of user interactions with THOMAS are presented in Figure A.1. As shown in the figure, THOMAS includes a few pre-defined interaction types. Even though THOMAS handles a sequence of interactions, it does not model users which is essential for IIR systems. Croft and Thompson (1987) closed this gap by proposing the I^3R system – the first IIR system with a user modeling component. I^3R uses a mixture of experts architecture. It assists users by accepting Boolean queries, typical text queries, and documents (query by examples). It enables users to provide explicit relevance feedback to the system. Example snapshots of user interactions with I^3R is presented in Figure A.2. Later on, Belkin *et al.* (1995) focused on user interactions with IIR systems and characterized information seeking strategies for interactive IR, offering users choices in a search session based on case-based reasoning. They defined a multi-dimensional space of information seeking strategies and applied their model to the MERIT system, a prototype IIR system that implements these multi-dimensional design principles.

Since the development of web search engines, research has mostly focused heavily on understanding user interaction with search engines based on an analysis of the search logs available to commercial search engine providers, *e.g.*, see Dumais *et al.* (2014), Buscher *et al.* (2009), Teevan *et al.* (2007), and Murray and Teevan (2007). Since then, explicit modeling of information seeking dialogues or conversations with the aim of improving the effectiveness of retrieval has not been a focus of research until recently. Among them, session search is perhaps the closest research area to CIS (see Section A.3).

A.2 Formal Modeling of IIR Systems

The proposition that IR systems are fundamentally interactive and should be evaluated from the users' perspective is not new (Kelly, 2009). This has been highlighted by many pioneers in the field since the 1960s (Cleverdon and Kean, 1968; Salton, 1970). However, today's search engines are mostly based on algorithms designed for retrieving documents for a single query. A main reason for this is due to the complexity of IIR modeling and evaluation. Recently, there has been some promising progress in formal modeling of IIR problems, including

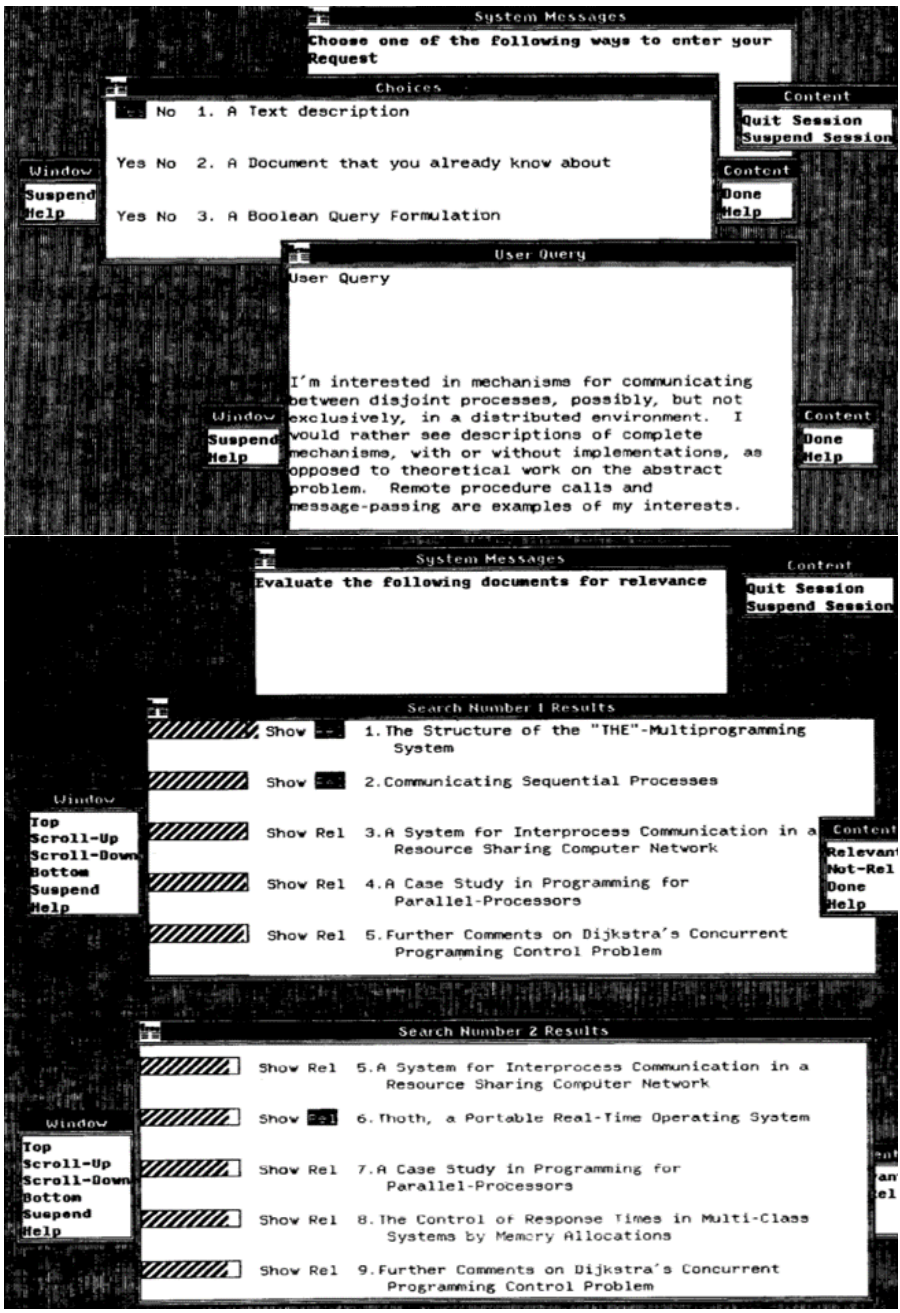

Figure A.2: Snapshots from the I^{3}R system (Croft and Thompson, 1987).

the probability ranking principle for IIR (Fuhr, 2008), the economics models for IIR (Azzopardi, 2011), the game theoretic framework for IR (Zhai, 2016), and the interface card model (Zhang and Zhai, 2015). Conversational information seeking is an emerging application of IIR systems and many of the developed IIR models and evaluation methodologies can be extended to CIS systems too. For further reading on approaches for designing and evaluating IIR systems, we refer the reader to the comprehensive survey by Kelly (2009) and the recent tutorial by Zhai (2020).

A.3 Session-based Information Retrieval

One can put information retrieval tasks in context based on the user's short-term history (Bennett *et al.*, 2012), the user's long-term history (Keenoy and Levene, 2003), or the user's situation (Zamani *et al.*, 2017).

Short-term history is often formulated by the user interactions with the search engine in a short period of time (*e.g.*, a few minutes), referred to as a *search session*. Sessions are different from conversations in that one can pick up and continue a past conversation, while this is not possible in sessions.

Interactions in a session include past queries, retrieved documents, and clicked documents. Therefore, a session can be considered as a period consisting of all interactions for the same information need (Shen *et al.*, 2005). However, this is a strong assumption. In reality, sessions are complex and they are not all alike. Some sessions contain various interactions and query reformulations for a single information need, while some other sessions may involve a series of related simple tasks. Therefore, sessions should be treated differently. This makes modeling search sessions challenging. Existing methods oftentimes relax the assumptions. For instance, Shen *et al.* (2005) assumed that all queries in a session represent the same information need and proposed a model based on the language modeling framework (Ponte and Croft, 1998) for

session search tasks. In more detail, they provide a more accurate query language model by interpolating the distribution estimated from the current query, with the ones estimated from the past queries and clicked documents. Bennett *et al.* (2012) introduced a learning to rank approach for session search and defined a number of features that can be used for improving the session search performance in web search. TREC Session Track (Carterette *et al.*, 2016) focused on the development of query formulation during a search session and improving retrieval performance by incorporating knowledge of the session context. Session information can also be used for a number of other information retrieval tasks, such as query suggestion (Sordoni *et al.*, 2015; Dehghani *et al.*, 2017) and clarification (Zamani *et al.*, 2020a).

Whole session evaluation of IR systems is also challenging. Järvelin *et al.* (2008) proposed sDCG, an extension of the nDCG (Järvelin and Kekäläinen, 2002) metric to session search tasks. sDCG basically sums up the nDCG values of all the queries in the session and gives higher weight to the earlier queries. Kanoulas *et al.* (2011) later introduced a normalized variation of sDCG, called nsDCG. Jiang and Allan (2016) conducted a user study to measure the correlation between these metrics and user's opinion. They found that nsDCG has a significant yet weak correlation with the user metrics. They also showed that user's opinions are highly correlated with the performance of the worst and the last queries in the session. More recently, Lipani *et al.* (2019) proposed a user model for session search in which users at each step make a decision to assess the documents in the result list or submit a new query. This user model led to the development of the sRBP metric.

It is clear that session search provides a logical foundation for conversational search tasks, however, there are some fundamental differences that necessitates developing novel models and evaluation methodologies for the conversational search tasks. For instance, since most conversational systems are using limited-bandwidth interfaces, the underlying user models of the aforementioned metrics cannot be extended to conversational search. From the modeling perspective, the type of queries in conversational systems are closer to natural language compared to the session search tasks. In addition, unlike in session search, co-reference and ellipsis resolutions play a central role in conversational search. That

being said, we believe that the rich history of IR research on session search would be sometimes quite useful in developing and evaluating conversational search systems.

A.4 Exploratory Search

A significant research effort in interactive IR has focused on *exploratory search* tasks. Exploratory search is an information retrieval task in which the user is unfamiliar with the search task, unsure about the goal, or even unsure about how to complete the task. Users engage in exploratory search with the aim of learning about and exploring a topic – as opposed to known-item/look-up tasks in which users are focused on finding a fact or answering a specific question. Exploratory search refers to a broad set of real-world search tasks that involve learning, investigation, planning, discovery, aggregation, and synthesis (Marchionini, 2006). Exploratory search tasks can be generally categorized as (1) exploratory browsing and (2) focused searching (White and Roth, 2009). Previous work on exploratory search has examined interface features to support users with query refinement and filtering (*e.g.*, faceted search) (Hearst, 2006); tools to help gather and synthesize information (Morris *et al.*, 2008; Donato *et al.*, 2010; Hearst and Degler, 2013); and tools to support collaboration (Golovchinsky *et al.*, 2009).

Natural language conversation is a convenient way for exploratory search tasks. In many exploratory search tasks, users experience difficulties describing their information needs using accurate keyword queries. This is mainly due to a misconception of the topics and/or the document collection. Information seeking conversations would be the natural solution for this problem as natural language conversation is perhaps the most convenient way of human communication and users can express their exploratory search needs quite easily.

Interestingly, many conversations in the TREC CAsT Tracks (Dalton *et al.*, 2019; Dalton *et al.*, 2020b) are basically addressing exploratory

information seeking tasks through natural language conversation.

A.5 Dialogue Systems

CIS is also related to dialogue systems. Many concepts used in developing CIS systems were also explored in the context of dialogue systems. Dialogue systems, or conversational agents, refer to computer systems that are intended to converse with humans through natural language. That being said, dialogue systems are not limited to natural language interactions and can benefit from one or more of text, speech, graphics, haptics, gestures, and other modalities. Dialogue systems are mainly categorized as either chatbots (a.k.a. chit-chat dialogue systems) or task-oriented dialogue systems. The former is designed to mimic human conversations mostly for entertainment, while the latter is developed to help the user accomplish a task, *e.g.*, hotel reservation. Task-oriented dialogues are closer to CIS yet with fundamental differences.

Designing and developing dialogue systems require a deep understanding of human conversation. Therefore, the dialogue community spent considerable efforts on extracting and modeling conversations. Jurafsky and Martin (2021) reviewed these properties in detail. For instance, *turn* is perhaps the simplest property – a dialogue is a sequence of turns, each a single contribution from one speaker. *Dialogue acts* is another important property – each dialogue utterance is a type of action performed by the speaker. Different modules in real-world dialogue systems are designed because of this property, such as dialogue act classification. *Grounding* is yet another property of dialogues – acknowledging that dialogue participants understand each other. *Initiative* is the final property we review here. As mentioned in Section 6, it is common in human conversations for initiative to shift back and forth between the participants. For example, in response to a question, a participant can ask for a clarification instead of immediately answering the question. Such interactions are called mixed-initiative. For learning more about dialogue properties and detailed explanations, we refer the reader to (Jurafsky and Martin, 2021, Chapter 24) and (McTear *et al.*, 2016, Chapter 3).

Dialogue systems have been studied for decades. ELIZA is an early

successful chatbot system, developed by Weizenbaum (1966) in the 1960s. ELIZA is a rule-based dialogue system designed to simulate a Rogerian psychologist. It involves drawing the patient out by reflecting patient's statements back at them. ELIZA selects the best match rule for every utterance (regular expression matching) and uses it for producing the next utterance. PARRY is an updated version of ELIZA developed by Colby *et al.* (1971) with a clinical psychology focus, used to study schizophrenia. Besides regular expressions, PARRY models fear and anger and uses these variables to generate utterances. PARRY was the first known system to pass the Turing test, meaning that psychologists could not distinguish PARRY's outputs from transcripts of interviews with real paranoids (Colby *et al.*, 1972).

Another successful implementation of dialogue systems in early years was done by the SHRDLU system (Winograd, 1972). SHRDLU provides a natural language interface to a virtual space filled with different blocks. Therefore, SHRDLU users could select and move objects in the virtual space. Given the few number of object types, the action space and vocabulary in SHRDLU conversations are highly limited. The AT&T How May I Help You? (HMIHY) system (Gorin *et al.*, 1997) is also a notable example of dialogue systems developed in the 1990s. HMIHY involved speech recognition, named entity extraction, and intent classification with the goal of call routing. It used a wizard-of-oz approach for data collection and training. It also implemented an active learning algorithm for language understanding.

Dialogue research was later accelerated by the DARPA Communicator Program. For instance, Xu and Rudnicky (2000) developed a language modeling framework for dialogue systems during the Communicator Program. It was designed to support the creation of speech-enabled interfaces that scale across modalities, from speech-only to interfaces that include graphics, maps, pointing and gesture. Recent chatbot systems often use large-scale language models, such as GPT-3 (Brown *et al.*, 2020), in addition to corpus-based approaches that retrieve information from an external corpus in order to produce more sensible utterances.

For task-oriented dialogue systems, Bobrow *et al.* (1977) introduced the GUS architecture in the 1970s. GUS is a frame-based architecture for dialogue systems, where a frame is a kind of knowledge structure

representing the information and intention that the system can extract from the user utterances. Thus, frames consist of many slots and dialogue systems need to extract and generate the values of these slots based on the conversation. Architectures similar to or inspired by GUS are still used in real dialogue systems. An alternative to such a modular architecture is end-to-end dialogue systems that do not explicitly model slots and are based on text generation models. We refer the reader to (Gao *et al.*, 2019, Chapter 4) for recent advances on task-oriented dialogue systems using neural models.

Evaluating dialogue systems is a challenging and widely explored topic. N-grams matching metrics, such as BLEU (Papineni *et al.*, 2002) and ROUGE (Lin, 2004), have been used for dialogue system evaluation. Semantic similarity-based metrics, such as BERT-Score (Zhang *et al.*, 2020b), have also been used. However, research shows that these metrics have several shortcomings (Liu *et al.*, 2016). Using human annotators to evaluate the output of the system and/or using implicit or explicit feedback provided by real users are perhaps the most reliable forms of evaluation for dialogue systems. The PARADISE framework (Walker *et al.*, 1997) for measure overall system success. Developing and evaluating dialogue systems are still active areas of research, we refer the reader to Finch and Choi (2020) for recent work.

A.6 Summary

In this appendix, we briefly reviewed decades of research related to systems and formal models for interactive information retrieval systems, exploratory search, and dialogue systems. Even though the natural language nature of interaction in CIS makes it more complex and many simplifying assumptions made by prior work on IIR cannot be overlooked in the context of CIS systems, many of the concepts that have been developed for IIR can be directly applied to or extended to CIS tasks. Same argument holds for past research on dialogue systems that has been briefly reviewed in the last subsection. Therefore, instead of re-inventing the wheel for various problems in CIS systems, we urge the reader to have a thorough review of the rich literature on IIR and dialogue research, some of which are pointed out in this appendix.

References

Abdollahpouri, H., G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, and L. Pizzato. 2020. "Multistakeholder recommendation: Survey and research directions". *User Modeling and User-Adapted Interaction*. 30(1): 127–158.

Adadi, A. and M. Berrada. 2018. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". *IEEE access*. 6: 52138–52160.

Adiwardana, D. D. F., M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le. 2020. "Towards a Human-like Open-Domain Chatbot". *ArXiv*. abs/2001.09977.

Aliannejadi, M., L. Azzopardi, H. Zamani, E. Kanoulas, P. Thomas, and N. Craswell. 2021a. "Analysing Mixed Initiatives and Search Strategies during Conversational Search". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management. CIKM '21*. New York, NY, USA: Association for Computing Machinery. 16–26. ISBN: 9781450384469. URL: https://doi.org/10.1145/3459637.3482231.

Aliannejadi, M., M. Chakraborty, E. A. Ríssola, and F. Crestani. 2020. "Harnessing Evolution of Multi-Turn Conversations for Effective Answer Retrieval". *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*.

Aliannejadi, M., H. Zamani, F. Crestani, and W. B. Croft. 2018a. "In Situ and Context-Aware Target Apps Selection for Unified Mobile Search". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM '18*. Torino, Italy: Association for Computing Machinery. 1383–1392. ISBN: 9781450360142. DOI: 10.1145/3269206.3271679. URL: https://doi.org/10.1145/3269206.3271679.

Aliannejadi, M., H. Zamani, F. Crestani, and W. B. Croft. 2018b. "Target Apps Selection: Towards a Unified Search Framework for Mobile Devices". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18*. Ann Arbor, MI, USA: Association for Computing Machinery. 215–224. ISBN: 9781450356572. DOI: 10.1145/3209978.3210039. URL: https://doi.org/10.1145/3209978.3210039.

Aliannejadi, M., H. Zamani, F. Crestani, and W. B. Croft. 2019. "Asking Clarifying Questions in Open-Domain Information Seeking Conversations". In: *Proceedings of the 42nd ACM SIGIR International Conference on Research and Development on Information Retrieval (to appear). SIGIR '19*.

Aliannejadi, M., H. Zamani, F. Crestani, and W. B. Croft. 2021b. "Context-Aware Target Apps Selection and Recommendation for Enhancing Personal Mobile Assistants". *ACM Trans. Inf. Syst.* 39(3). ISSN: 1046-8188. DOI: 10.1145/3447678. URL: https://doi.org/10.1145/3447678.

Allen, J. E., C. I. Guinn, and E. Horvitz. 1999. "Mixed-Initiative Interaction". *IEEE Intelligent Systems and their Applications*. 14(5): 14–23.

Anand, A., L. Cavedon, H. Joho, M. Sanderson, and B. Stein. 2020. "Conversational Search (Dagstuhl Seminar 19461)". In: Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Anantha, R., S. Vakulenko, Z. Tu, S. Longpre, S. Pulman, and S. Chappidi. 2021. "Open-Domain Question Answering Goes Conversational via Question Rewriting". In: *NAACL*.

Andolina, S., V. Orso, H. Schneider, K. Klouche, T. Ruotsalo, L. Gamberini, and G. Jacucci. 2018. "Investigating Proactive Search Support in Conversations". In: *Proceedings of the 2018 Designing Interactive Systems Conference. DIS '18*. Hong Kong, China: Association for Computing Machinery. 1295–1307. ISBN: 9781450351980. DOI: 10.1145/3196709.3196734. URL: https://doi-org.ezp.lib.unimelb.edu.au/10.1145/3196709.3196734.

Arnold, A., G. Dupont, C. Kobus, F. Lancelot, and Y.-H. Liu. 2020. "Perceived Usefulness of Conversational Agents Predicts Search Performance in Aerospace Domain". In: *Proceedings of the 2nd Conference on Conversational User Interfaces. CUI '20*. Bilbao, Spain: Association for Computing Machinery. ISBN: 9781450375443. DOI: 10.1145/3405755.3406172. URL: https://doi.org/10.1145/3405755.3406172.

Arons, B. 1997. "SpeechSkimmer: a system for interactively skimming recorded speech". *ACM Transactions on Computer-Human Interaction (TOCHI)*. 4(1): 3–38.

Avula, S. 2020. "Characterizing and Understanding User Perception of System Initiative for Conversational Systems to Support Collaborative Search". *PhD thesis*. The University of North Carolina at Chapel Hill.

Avula, S. and J. Arguello. 2020. "Wizard of Oz Interface to Study System Initiative for Conversational Search". In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. CHIIR '20*. Vancouver BC, Canada: Association for Computing Machinery. 447–451. ISBN: 9781450368926. DOI: 10.1145/3343413.3377941. URL: https://doi-org.ezp.lib.unimelb.edu.au/10.1145/3343413.3377941.

Avula, S., G. Chadwick, J. Arguello, and R. Capra. 2018. "SearchBots: User Engagement with ChatBots during Collaborative Search". In: *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval. CHIIR '18*. New Brunswick, NJ, USA: Association for Computing Machinery. 52–61. ISBN: 9781450349253. DOI: 10.1145/3176349.3176380. URL: https://doi.org/10.1145/3176349.3176380.

Azzopardi, L., M. Dubiel, M. Halvey, and J. Dalton. 2018. "Conceptualizing agent-human interactions during the conversational search process". In:

Azzopardi, L. 2011. "The Economics in Interactive Information Retrieval". In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '11*. Beijing, China: Association for Computing Machinery. 15–24. ISBN: 9781450307574. DOI: 10.1145/2009916.2009923. URL: https://doi.org/10.1145/2009916.2009923.

Azzopardi, L. 2021. "Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval". In: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. CHIIR '21*. Canberra ACT, Australia: Association for Computing Machinery. 27–37. ISBN: 9781450380553. DOI: 10.1145/3406522.3446023. URL: https://doi.org/10.1145/3406522.3446023.

Bagga, S., A. Piper, and D. Ruths. 2021. ""Are you kidding me?": Detecting Unpalatable Questions on Reddit". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2083–2099.

Baheti, A., A. Ritter, and K. Small. 2020. "Fluent Response Generation for Conversational Question Answering". In: *ACL*.

Balasuriya, S. S., L. Sitbon, A. A. Bayor, M. Hoogstrate, and M. Brereton. 2018. "Use of Voice Activated Interfaces by People with Intellectual Disability". In: *Proceedings of the 30th Australian Conference on Computer-Human Interaction. OzCHI '18*. Melbourne, Australia: Association for Computing Machinery. 102–112. ISBN: 9781450361880. DOI: 10.1145/3292147.3292161. URL: https://doi.org/10.1145/3292147.3292161.

Baldauf, M., R. Bösch, C. Frei, F. Hautle, and M. Jenny. 2018. "Exploring Requirements and Opportunities of Conversational User Interfaces for the Cognitively Impaired". In: *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct. MobileHCI '18*. Barcelona, Spain: Association for Computing Machinery. 119–126. ISBN: 9781450359412. DOI: 10.1145/3236112.3236128.

Balog, K. and F. Radlinski. 2020. "Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*.

Banerjee, S. and A. Lavie. 2005. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics. 65–72. URL: https://www.aclweb.org/anthology/W05-0909.

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

Belkin, N. J. and W. B. Croft. 1992. "Information Filtering and Information Retrieval: Two Sides of the Same Coin?" *Commun. ACM*. 35(12): 29–38. ISSN: 0001-0782. DOI: 10.1145/138859.138861. URL: https://doi.org/10.1145/138859.138861.

Beltagy, I., M. E. Peters, and A. Cohan. 2020. "Longformer: The Long-Document Transformer". *arXiv:2004.05150*.

Bendersky, M. and W. Croft. 2008. "Discovering key concepts in verbose queries". In: *SIGIR '08*.

Bennett, P. N., R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui. 2012. "Modeling the Impact of Short- and Long-term Behavior on Search Personalization". In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '12*. Portland, Oregon, USA: ACM. 185–194. ISBN: 978-1-4503-1472-5. DOI: 10.1145/2348283.2348312. URL: http://doi.acm.org/10.1145/2348283.2348312.

Benzeghiba, M., R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. 2007. "Automatic Speech Recognition and Speech Variability: A Review". *Speech Commun.* 49(10–11): 763–786. ISSN: 0167-6393. DOI: 10.1016/j.specom.2007.02.006. URL: https://doi.org/10.1016/j.specom.2007.02.006.

Beutel, A., J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, *et al.* 2019. "Fairness in recommendation ranking through pairwise comparisons". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2212–2220.

Bi, K., Q. Ai, and W. B. Croft. 2021. "Asking Clarifying Questions Based on Negative Feedback in Conversational Search". In: *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. New York, NY, USA: Association for Computing Machinery. 157–166. ISBN: 9781450386111. URL: https://doi.org/10.1145/3471158.3472232.

Bickmore, T. W., D. Utami, R. Matsuyama, and M. K. Paasche-Orlow. 2016. "Improving access to online health information with conversational agents: a randomized controlled experiment". *Journal of medical Internet research*. 18(1): e5239.

Blanco, H. and F. Ricci. 2013. "Acquiring User Profiles from Implicit Feedback in a Conversational Recommender System". In: *Proceedings of the 7th ACM Conference on Recommender Systems. RecSys '13*. Hong Kong, China: Association for Computing Machinery. 307–310. ISBN: 9781450324090. DOI: 10.1145/2507157.2507217. URL: https://doi.org/10.1145/2507157.2507217.

Bobrow, D. G., R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd. 1977. "GUS, a Frame-Driven Dialog System". *Artif. Intell.* 8(2): 155–173. ISSN: 0004-3702. DOI: 10.1016/0004-3702(77)90018-2.

Boye, J., B. A. Hockey, and M. Rayner. 2000. "Asynchronous dialogue management: Two case-studies". In: *Gotalog: Fourth Workshop on the Semantics and Pragmatics of Dialogue*.

Braslavski, P., D. Savenkov, E. Agichtein, and A. Dubatovka. 2017. "What Do You Mean Exactly? Analyzing Clarification Questions in CQA". In: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. CHIIR '17*. Oslo, Norway: Association for Computing Machinery. 345–348. ISBN: 9781450346771. DOI: 10.1145/3020165.3022149. URL: https://doi.org/10.1145/3020165.3022149.

Brennan, S. E. 2012. "Conversation and dialogue". In: *Encyclopedia of the Mind*. Ed. by H. Pashler. SAGE Publications. 202–205.

Brooks, H. M. and N. J. Belkin. 1983. "Using Discourse Analysis for the Design of Information Retrieval Interaction Mechanisms". *SIGIR Forum*. 17(4): 31–47. ISSN: 0163-5840. DOI: 10.1145/1013230.511800. URL: https://doi.org/10.1145/1013230.511800.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. 2020. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Budzianowski, P. and I. Vulić. 2019. "Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems". In: *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Hong Kong: Association for Computational Linguistics. 15–22. DOI: 10.18653/v1/D19-5602. URL: https://www.aclweb.org/anthology/D19-5602.

Budzianowski, P., T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gasic. 2018. "MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 5016–5026.

Bunt, H., V. Petukhova, D. Traum, and J. Alexandersson. 2017. "Dialogue Act Annotation with the ISO 24617-2 Standard". In:

Buscher, G., J. Gwizdka, J. Teevan, N. J. Belkin, R. Bierig, L. van Elst, and J. Jose. 2009. "SIGIR 2009 Workshop on Understanding the User: Logging and Interpreting User Interactions in Information Search and Retrieval". *SIGIR Forum*. 43(2): 57–62. ISSN: 0163-5840. DOI: 10.1145/1670564.1670574. URL: https://doi.org/10.1145/1670564.1670574.

Byrne, B., K. Krishnamoorthi, C. Sankar, A. Neelakantan, B. Goodrich, D. Duckworth, S. Yavuz, A. Dubey, K.-Y. Kim, and A. Cedilnik. 2019. "Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4506–4517.

Car, L. T., D. A. Dhinagaran, B. M. Kyaw, T. Kowatsch, S. Joty, Y.-L. Theng, and R. Atun. 2020. "Conversational agents in health care: scoping review and conceptual analysis". *Journal of medical Internet research*. 22(8): e17158.

Carterette, B., P. Clough, M. Hall, E. Kanoulas, and M. Sanderson. 2016. "Evaluating Retrieval over Sessions: The TREC Session Track 2011-2014". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '16*. Pisa, Italy: ACM. 685–688. ISBN: 978-1-4503-4069-4. DOI: 10.1145/2911451.2914675. URL: http://doi.acm.org/10.1145/2911451.2914675.

Chen, D., A. Fisch, J. Weston, and A. Bordes. 2017. "Reading Wikipedia to Answer Open-Domain Questions". In: *ACL*.

Chen, L., G. Chen, and F. Wang. 2015. "Recommender Systems Based on User Reviews: The State of the Art". *User Modeling and User-Adapted Interaction*. 25(2): 99–154. ISSN: 0924-1868. DOI: 10.1007/s11257-015-9155-5. URL: https://doi.org/10.1007/s11257-015-9155-5.

Chen, Y., L. Wu, and M. J. Zaki. 2020a. "GraphFlow: Exploiting Conversation Flow with Graph Neural Networks for Conversational Machine Comprehension". In: *IJCAI*.

Chen, Z., X. Wang, X. Xie, M. Parsana, A. Soni, X. Ao, and E. Chen. 2020b. "Towards Explainable Conversational Recommendation." In: *IJCAI*. 2994–3000.

Chiang, T.-R., H.-T. Ye, and Y.-N. Chen. 2020. "An Empirical Study of Content Understanding in Conversational Question Answering". *ArXiv*. abs/1909.10743.

Choi, B.-J., J. Hong, D. Park, and S. W. Lee. 2020. "F^2-Softmax: Diversifying Neural Text Generation via Frequency Factorized Softmax". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. EMNLP '20*. Online: Association for Computational Linguistics. 9167–9182. DOI: 10.18653/v1/2020.emnlp-main.737.

Choi, E., H. He, M. Iyyer, M. Yatskar, W.-t. Yih, Y. Choi, P. Liang, and L. Zettlemoyer. 2018. "QuAC: Question Answering in Context". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. EMNLP '18*. Brussels, Belgium: Association for Computational Linguistics. 2174–2184.

Christakopoulou, K., F. Radlinski, and K. Hofmann. 2016. "Towards Conversational Recommender Systems". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16*. San Francisco, California, USA: Association for Computing Machinery. 815–824. ISBN: 9781450342322. DOI: 10.1145/2939672.2939746. URL: https://doi.org/10.1145/2939672.2939746.

Christmann, P., R. Saha Roy, A. Abujabal, J. Singh, and G. Weikum. 2019. "Look before You Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management. CIKM '19*. Beijing, China: Association for Computing Machinery. 729–738. ISBN: 9781450369763. DOI: 10.1145/3357384.3358016.

Chuklin, A., A. Severyn, J. R. Trippas, E. Alfonseca, H. Silen, and D. Spina. 2018. "Prosody Modifications for Question-Answering in Voice-Only Settings". *arXiv preprint arXiv:1806.03957*: 1–5.

Chuklin, A., A. Severyn, J. R. Trippas, E. Alfonseca, H. Silen, and D. Spina. 2019. "Using Audio Transformations to Improve Comprehension in Voice Question Answering". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, and N. Ferro. Cham: Springer International Publishing. 164–170. ISBN: 978-3-030-28577-7.

Church, K. and N. Oliver. 2011. "Understanding Mobile Web and Mobile Search Use in Today's Dynamic Mobile Landscape". In: *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services. MobileHCI '11*. Stockholm, Sweden: Association for Computing Machinery. 67–76. ISBN: 9781450305419. DOI: 10.1145/2037373.2037385. URL: https://doi.org/10.1145/2037373.2037385.

Clark, D. 1988. "The Design Philosophy of the DARPA Internet Protocols". *SIGCOMM Comput. Commun. Rev.* 18(4): 106–114. ISSN: 0146-4833. DOI: 10.1145/52325.52336. URL: https://doi.org/10.1145/52325.52336.

Clarke, C. 2019. "WaterlooClarke at the TREC 2019 Conversational Assistant Track". In: *TREC*.

Clarke, C. L. A., E. Agichtein, S. Dumais, and R. W. White. 2007. "The Influence of Caption Features on Clickthrough Patterns in Web Search". In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '07*. Amsterdam, The Netherlands: Association for Computing Machinery. 135–142. ISBN: 9781595935977. DOI: 10.1145/1277741.1277767. URL: https://doi.org/10.1145/1277741.1277767.

Cleverdon, C. and M. Kean. 1968. "Factors Determining the Performance of Indexing Systems". Aslib Cranfield Research Project, Cranfield, England.

Coden, A., D. Gruhl, N. Lewis, and P. N. Mendes. 2015. "Did you mean A or B? Supporting Clarification Dialog for Entity Disambiguation". In: *Joint Proceedings of the 1st International Workshop on Summarizing and Presenting Entities and Ontologies and the 3rd International Workshop on Human Semantic Web Interfaces (SumPre 2015, HSWI 2015) co-located with the 12th Extended Semantic Web Conference (ESWC 2015)*. Vol. 1556. *CEUR Workshop Proceedings*. Portoroz, Slovenia: CEUR-WS.org.

Colby, K. M., F. D. Hilf, S. Weber, and H. C. Kraemer. 1972. "Turing-like Indistinguishability Tests for the Validation of a Computer Simulation of Paranoid Processes". *Artif. Intell.* 3(1): 199–221. ISSN: 0004-3702. DOI: 10.1016/0004-3702(72)90049-5.

Colby, K. M., S. Weber, and F. D. Hilf. 1971. "Artificial Paranoia". *Artificial Intelligence*. 2(1): 1–25. ISSN: 0004-3702. DOI: https://doi.org/10.1016/0004-3702(71)90002-6.

Cooper, A. 2008. "A Survey of Query Log Privacy-Enhancing Techniques from a Policy Perspective". *ACM Trans. Web*. 2(4). ISSN: 1559-1131. DOI: 10.1145/1409220.1409222. URL: https://doi.org/10.1145/1409220.1409222.

Croft, W. B., D. Metzler, and T. Strohman. 2010. *Search engines: Information retrieval in practice*. Vol. 520. Addison-Wesley Reading.

Croft, W. B. and R. H. Thompson. 1987. "I³R: A New Approach to the Design of Document Retrieval Systems". *J. Am. Soc. Inf. Sci.* 38(6): 389–404. ISSN: 0002-8231.

Cross, S., A. Mourad, G. Zuccon, and B. Koopman. 2021. "Search Engines vs. Symptom Checkers: A Comparison of Their Effectiveness for Online Health Advice". In: *Proceedings of the Web Conference 2021. WWW '21*. Ljubljana, Slovenia: Association for Computing Machinery. 206–216. ISBN: 9781450383127. DOI: 10.1145/3442381.3450140. URL: https://doi.org/10.1145/3442381.3450140.

Cucerzan, S. 2007. "Large-Scale Named Entity Disambiguation Based on Wikipedia Data". In: *EMNLP-CoNLL*.

Culpepper, J. S., F. Diaz, and M. D. Smucker. 2018. "Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018)". *SIGIR Forum*. 52(1): 34–90. ISSN: 0163-5840. DOI: 10.1145/3274784.3274788. URL: https://doi-org.ezp.lib.unimelb.edu.au/10.1145/3274784.3274788.

Cutrell, E. and Z. Guan. 2007. "What Are You Looking for? An Eye-Tracking Study of Information Usage in Web Search". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '07*. San Jose, California, USA: Association for Computing Machinery. 407–416. ISBN: 9781595935939. DOI: 10.1145/1240624.1240690. URL: https://doi.org/10.1145/1240624.1240690.

Dalton, J., C. Xiong, V. Kumar, and J. Callan. 2020a. "CAsT-19: A Dataset for Conversational Information Seeking". *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Dalton, J., C. Xiong, and J. Callan. 2019. "TREC CAsT 2019: The Conversational Assistance Track Overview".

Dalton, J., C. Xiong, and J. Callan. 2020b. "TREC CAsT 2020: The Conversational Assistance Track Overview".

Dalton, J., C. Xiong, and J. Callan. 2020c. "TREC CAsT 2021: The Conversational Assistance Track Overview".

Daronnat, S., L. Azzopardi, M. Halvey, and M. Dubiel. 2020. "Impact of agent reliability and predictability on trust in real time human-agent collaboration". In: *Proceedings of the 8th International Conference on Human-Agent Interaction*. 131–139.

Dehghani, M., H. Azarbonyad, J. Kamps, and M. de Rijke. 2019. "Learning to Transform, Combine, and Reason in Open-Domain Question Answering". In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. WSDM '19*. Melbourne VIC, Australia: Association for Computing Machinery. 681–689. ISBN: 9781450359405. DOI: 10.1145/3289600.3291012. URL: https://doi.org/10.1145/3289600.3291012.

Dehghani, M., S. Rothe, E. Alfonseca, and P. Fleury. 2017. "Learning to Attend, Copy, and Generate for Session-Based Query Suggestion". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. CIKM '17*. Singapore, Singapore: Association for Computing Machinery. 1747–1756. ISBN: 9781450349185. DOI: 10.1145/3132847.3133010. URL: https://doi.org/10.1145/3132847.3133010.

Deits, R., S. Tellex, P. Thaker, D. Simeonov, T. Kollar, and N. Roy. 2013. "Clarifying Commands with Information-Theoretic Human-Robot Dialog". *J. Hum.-Robot Interact.* 2(2): 58–79. DOI: 10.5898/JHRI.2.2.Deits. URL: https://doi-org.ezp.lib.unimelb.edu.au/10.5898/JHRI.2.2.Deits.

Deldjoo, Y., J. R. Trippas, and H. Zamani. 2021. "Towards Multi-Modal Conversational Information Seeking". In: *Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR'21)*. 1–11.

Derboven, J., J. Huyghe, and D. De Grooff. 2014. "Designing Voice Interaction for People with Physical and Speech Impairments". In: *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational. NordiCHI '14*. Helsinki, Finland: Association for Computing Machinery. 217–226. ISBN: 9781450325424. DOI: 10.1145/2639189.2639252. URL: https://doi.org/10.1145/2639189.2639252.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL-HLT*.

Diaz, F. and D. Metzler. 2006. "Improving the estimation of relevance models using large external corpora". *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*.

Dinan, E., V. Logacheva, V. Malykh, A. H. Miller, K. Shuster, J. Urbanek, D. Kiela, A. Szlam, I. Serban, R. Lowe, S. Prabhumoye, A. Black, A. I. Rudnicky, J. Williams, J. Pineau, M. Burtsev, and J. Weston. 2019a. "The Second Conversational Intelligence Challenge (ConvAI2)". *ArXiv*. abs/1902.00098.

Dinan, E., S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston. 2019b. "Wizard of Wikipedia: Knowledge-Powered Conversational Agents". In: *International Conference on Learning Representations. ICLR '19*. URL: https://openreview.net/forum?id=r1l73iRqKm.

Donato, D., F. Bonchi, T. Chi, and Y. Maarek. 2010. "Do You Want to Take Notes?: Identifying Research Missions in Yahoo! Search Pad". In: *Proceedings of the 19th International Conference on World Wide Web. WWW '10*. ACM. 321–330.

Du, J., Z. Zhang, J. Yan, Y. Cui, and Z. Chen. 2010. "Using search session context for named entity recognition in query". *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*.

Dubiel, M., M. Halvey, L. Azzopardi, D. Anderson, and S. Daronnat. 2020a. "Conversational strategies: impact on search performance in a goal-oriented task". In: *The Third International Workshop on Conversational Approaches to Information Retrieval*.

Dubiel, M., M. Halvey, L. Azzopardi, and S. Daronnat. 2018. "Investigating how conversational search agents affect user's behaviour, performance and search experience". In: *The Second International Workshop on Conversational Approaches to Information Retrieval*.

Dubiel, M., M. Halvey, P. O. Gallegos, and S. King. 2020b. "Persuasive Synthetic Speech: Voice Perception and User Behaviour". In: *Proceedings of the 2nd Conference on Conversational User Interfaces. CUI '20*. Bilbao, Spain: Association for Computing Machinery. ISBN: 9781450375443. DOI: 10.1145/3405755.3406120. URL: https://doi.org/10.1145/3405755.3406120.

Dumais, S., R. Jeffries, D. M. Russell, D. Tang, and J. Teevan. 2014. "Understanding User Behavior Through Log Data and Analysis". In: *Ways of Knowing in HCI*. Ed. by J. S. Olson and W. A. Kellogg. New York, NY: Springer New York. 349–372. ISBN: 978-1-4939-0378-8. DOI: 10.1007/978-1-4939-0378-8_14. URL: https://doi.org/10.1007/978-1-4939-0378-8_14.

Dušek, O., J. Novikova, and V. Rieser. 2018. "Findings of the E2E NLG Challenge". In: *Proceedings of the 11th International Conference on Natural Language Generation*. Tilburg University, The Netherlands: Association for Computational Linguistics. 322–328. DOI: 10.18653/v1/W18-6539. URL: https://www.aclweb.org/anthology/W18-6539.

Dwivedi, S. K. and V. Singh. 2013. "Research and reviews in question answering system". *Procedia Technology*. 10: 417–424.

Dziri, N., H. Rashkin, T. Linzen, and D. Reitter. 2021. "Evaluating Groundedness in Dialogue Systems: The BEGIN Benchmark". *ArXiv*. abs/2105.00071.

Elgohary, A., D. Peskov, and J. L. Boyd-Graber. 2019. "Can You Unpack That? Learning to Rewrite Questions-in-Context". In: *EMNLP/IJCNLP*.

Feild, H. A., J. Allan, and R. Jones. 2010. "Predicting Searcher Frustration". In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '10*. Geneva, Switzerland: Association for Computing Machinery. 34–41. ISBN: 9781450301534. DOI: 10.1145/1835449.1835458. URL: https://doi.org/10.1145/1835449.1835458.

Ferguson, G. and J. F. Allen. 1998. "TRIPS: An Integrated Intelligent Problem-Solving Assistant". In: *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference. AAAI '98, IAAI '98*. Madison, Wisconsin, USA: AAAI Press / The MIT Press. 567–572.

Finch, S. E. and J. D. Choi. 2020. "Towards Unified Dialogue System Evaluation: A Comprehensive Analysis of Current Evaluation Protocols". In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 236–245.

Firdaus, M., N. Thakur, and A. Ekbal. 2021. "Aspect-Aware Response Generation for Multimodal Dialogue System". *ACM Trans. Intell. Syst. Technol.* 12(2). ISSN: 2157-6904. DOI: 10.1145/3430752. URL: https://doi.org/10.1145/3430752.

Fono, D. and R. Baecker. 2006. "Structuring and Supporting Persistent Chat Conversations". In: *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work. CSCW '06*. Banff, Alberta, Canada: Association for Computing Machinery. 455–458. ISBN: 1595932496. DOI: 10.1145/1180875.1180944. URL: https://doi.org/10.1145/1180875.1180944.

Foster, M. E., R. Alami, O. Gestranius, O. Lemon, M. Niemelä, J.-M. Odobez, and A. K. Pandey. 2016. "The MuMMER Project: Engaging Human-Robot Interaction in Real-World Public Spaces". In: *Social Robotics*. Ed. by A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, and H. He. Cham: Springer International Publishing. 753–763. ISBN: 978-3-319-47437-3.

Fraser, N. 1998. "Assessment of interactive systems". In: *Handbook of standards and resources for spoken language systems*. Ed. by D. Gibbon, R. Moore, and R. Winski. Mouton de Gruyter. 564–615.

Freed, A. R. 2021. *Conversational AI: Chatbots that work*. Manning Publications Co.

Fuhr, N. 2008. "A Probability Ranking Principle for Interactive Information Retrieval". *Inf. Retr.* 11(3): 251–265. ISSN: 1386-4564. DOI: 10.1007/s10791-008-9045-0. URL: https://doi.org/10.1007/s10791-008-9045-0.

"Multimodal Interfaces". 2008. In: *Encyclopedia of Multimedia*. Ed. by B. Furht. Boston, MA: Springer US. 651–652. ISBN: 978-0-387-78414-4. DOI: 10.1007/978-0-387-78414-4_159. URL: https://doi.org/10.1007/978-0-387-78414-4_159.

Gao, C., W. Lei, X. He, M. de Rijke, and T.-S. Chua. 2021a. "Advances and challenges in conversational recommender systems: A survey". *AI Open*. 2: 100–126. ISSN: 2666-6510. DOI: https://doi.org/10.1016/j.aiopen.2021.06.002. URL: https://www.sciencedirect.com/science/article/pii/S2666651021000164.

Gao, J., M. Galley, and L. Li. 2019. "Neural Approaches to Conversational AI". *Foundations and Trends® in Information Retrieval*. 13(2-3): 127–298. ISSN: 1554-0669. DOI: 10.1561/1500000074. URL: http://dx.doi.org/10.1561/1500000074.

Gao, Y., J. Li, M. R. Lyu, and I. King. 2021b. "Open-Retrieval Conversational Machine Reading". *ArXiv*. abs/2102.08633.

Gao, Y., C.-S. Wu, J. Li, S. R. Joty, S. C. H. Hoi, C. Xiong, I. King, and M. R. Lyu. 2020. "Discern: Discourse-Aware Entailment Reasoning Network for Conversational Machine Reading". *ArXiv*. abs/2010.01838.

Ge, Y., S. Liu, R. Gao, Y. Xian, Y. Li, X. Zhao, C. Pei, F. Sun, J. Ge, W. Ou, *et al.* 2021. "Towards Long-term Fairness in Recommendation". In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 445–453.

Gemmell, C. and J. Dalton. 2020. "Glasgow Representation and Information Learning Lab (GRILL) at the Conversational Assistance Track 2020". In: *TREC*.

Gerritse, E., F. Hasibi, and A. D. Vries. 2020. "Bias in Conversational Search: The Double-Edged Sword of the Personalized Knowledge Graph". *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*.

Gibbon, D., R. Moore, and R. Winski. 1997. *Handbook of Standards and Resources for Spoken Language Systems*. de Gruyter.

Gibson, W. 2009. "Intercultural communication online: Conversation analysis and the investigation of asynchronous written discourse". In: *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*. Vol. 10. No. 1.

Goes, P., N. Ilk, W. T. Yue, and J. L. Zhao. 2012. "Live-Chat Agent Assignments to Heterogeneous e-Customers under Imperfect Classification". *ACM Trans. Manage. Inf. Syst.* 2(4). ISSN: 2158-656X. DOI: 10.1145/2070710.2070715. URL: https://doi.org/10.1145/2070710.2070715.

Golovchinsky, G., P. Qvarfordt, and J. Pickens. 2009. "Collaborative Information Seeking". *Computer*. 42(3): 47–51. DOI: 10.1109/MC.2009.73. URL: https://doi.org/10.1109/MC.2009.73.

Gooda Sahib, N., A. Tombros, and T. Stockman. 2015. "Evaluating a Search Interface for Visually Impaired Searchers". *J. Assoc. Inf. Sci. Technol.* 66(11): 2235–2248. ISSN: 2330-1635.

Gorin, A. L., G. Riccardi, and J. H. Wright. 1997. "How May I Help You?" *Speech Commun.* 23(1–2): 113–127. ISSN: 0167-6393. DOI: 10.1016/S0167-6393(97)00040-X. URL: https://doi.org/10.1016/S0167-6393(97)00040-X.

Gosper, S., J. R. Trippas, H. Richards, F. Allison, C. Sear, S. Khorasani, and F. Mattioli. 2021. "Understanding the Utility of Digital Flight Assistants: A Preliminary Analysis". In: *CUI 2021 - 3rd Conference on Conversational User Interfaces. CUI '21.* Bilbao (online), Spain: Association for Computing Machinery. ISBN: 9781450389983. DOI: 10.1145/3469595.3469627. URL: https://doi.org/10.1145/3469595.3469627.

Green, B. F., A. K. Wolf, C. Chomsky, and K. Laughery. 1961. "Baseball: An Automatic Question-Answerer". In: *Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference. IRE-AIEE-ACM '61 (Western).* Los Angeles, California: Association for Computing Machinery. 219–224. ISBN: 9781450378727. DOI: 10.1145/1460690.1460714. URL: https://doi.org/10.1145/1460690.1460714.

Gunasekara, R. C., S. Kim, L. F. D'Haro, A. Rastogi, Y.-N. Chen, M. Eric, B. Hedayatnia, K. Gopalakrishnan, Y. Liu, C.-W. Huang, D. Hakkani-Tür, J. Li, Q. Zhu, L. Luo, L. Liden, K. Huang, S. Shayandeh, R. Liang, B. Peng, Z. Zhang, S. Shukla, M. Huang, J. Gao, S. Mehri, Y. Feng, C. Gordon, S. H. Alavi, D. R. Traum, M. Eskénazi, A. Beirami, E. Cho, P. A. Crook, A. De, A. Geramifard, S. Kottur, S. Moon, S. Poddar, and R. Subba. 2020. "Overview of the Ninth Dialog System Technology Challenge: DSTC9". *CoRR*. abs/2011.06486. arXiv: 2011.06486. URL: https://arxiv.org/abs/2011.06486.

Guo, D., D. Tang, N. Duan, M. Zhou, and J. Yin. 2018. "Dialog-to-Action: Conversational Question Answering Over a Large-Scale Knowledge Base". In: *NeurIPS*.

Guo, J., Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, and X. Cheng. 2020. "A Deep Look into neural ranking models for information retrieval". *Information Processing & Management*. 57(6): 102067. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2019.102067.

Hashemi, H., H. Zamani, and W. B. Croft. 2020. "Guided Transformer: Leveraging Multiple External Sources for Representation Learning in Conversational Search". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20*. Virtual Event, China: Association for Computing Machinery. 1131–1140. ISBN: 9781450380164. DOI: 10.1145/3397271.3401061. URL: https://doi.org/10.1145/3397271.3401061.

Hassan Awadallah, A., R. Gurunath Kulkarni, U. Ozertem, and R. Jones. 2015. "Characterizing and Predicting Voice Query Reformulation". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. CIKM '15*. Melbourne, Australia: ACM. 543–552. ISBN: 978-1-4503-3794-6. DOI: 10.1145/2806416.2806491. URL: http://doi.acm.org/10.1145/2806416.2806491.

Hauptmann, A., J. Magalhaes, R. G. Sousa, and J. P. Costeira. 2020. "MuCAI'20: 1st International Workshop on Multimodal Conversational AI". In: *Proceedings of the 28th ACM International Conference on Multimedia. MM '20*. Seattle, WA, USA: Association for Computing Machinery. 4767–4768. ISBN: 9781450379885. DOI: 10.1145/3394171.3421900. URL: https://doi.org/10.1145/3394171.3421900.

Hawking, D. 2004. "Challenges in Enterprise Search". In: *Proceedings of the 15th Australasian Database Conference - Volume 27. ADC '04*. Dunedin, New Zealand: Australian Computer Society, Inc. 15–24.

He, J., P. Duboue, and J.-Y. Nie. 2012. "Bridging the gap between intrinsic and perceived relevance in snippet generation". In: *Proceedings of COLING 2012*. 1129–1146.

Hearst, M. A. 2006. "Clustering versus faceted categories for information exploration". *Communications of the ACM*. 49(4): 59–61.

Hearst, M. A. 2009. *Search User Interfaces*. Cambridge University Press.

Hearst, M. A. and D. Degler. 2013. "Sewing the Seams of Sensemaking: A Practical Interface for Tagging and Organizing Saved Search Results". In: *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval. HCIR '13*. Vancouver BC, Canada: Association for Computing Machinery. ISBN: 9781450325707. DOI: 10.1145/2528394.2528398. URL: https://doi.org/10.1145/2528394.2528398.

Henderson, M., P. Budzianowski, I. Casanueva, S. Coope, D. Gerz, G. Kumar, N. Mrkšić, G. Spithourakis, P.-H. Su, I. Vulić, and T.-H. Wen. 2019. "A Repository of Conversational Datasets". arXiv: 1904.06472 [cs.CL].

Henderson, M., I. Casanueva, N. Mrkvsi'c, P. Su, Tsung-Hsien, and I. Vulic. 2020. "ConveRT: Efficient and Accurate Conversational Representations from Transformers". *ArXiv*. abs/1911.03688.

Hochreiter, S. and J. Schmidhuber. 1997. "Long short-term memory". *Neural computation*. 9(8): 1735–1780.

Hofstätter, S., B. Mitra, H. Zamani, N. Craswell, and A. Hanbury. 2021. "Intra-Document Cascading: Learning to Select Passages for Neural Document Ranking". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. Virtual Event, Canada: Association for Computing Machinery. 1349–1358. ISBN: 9781450380379. DOI: 10.1145/3404835.3462889. URL: https://doi.org/10.1145/3404835.3462889.

Horvitz, E. 1999. "Principles of Mixed-initiative User Interfaces". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '99*. Pittsburgh, Pennsylvania, USA: ACM. 159–166. ISBN: 0-201-48559-1. DOI: 10.1145/302979.303030. URL: http://doi.acm.org/10.1145/302979.303030.

Huang, H., C. Zhu, Y. Shen, and W. Chen. 2018. "FusionNet: Fusing via Fully-Aware Attention with Application to Machine Comprehension". *ArXiv*. abs/1711.07341.

Huang, H.-Y., E. Choi, and W.-T. Yih. 2019. "FlowQA: Grasping Flow in History for Conversational Machine Comprehension". In: *International Conference on Learning Representations. ICLR '19*. URL: https://openreview.net/forum?id=ByftGnR9KX.

Humeau, S., K. Shuster, M.-A. Lachaux, and J. Weston. 2020. "Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring". In: *ICLR*.

Ie, E., C.-w. Hsu, M. Mladenov, V. Jain, S. Narvekar, J. Wang, R. Wu, and C. Boutilier. 2019. "RecSim: A Configurable Simulation Platform for Recommender Systems". arXiv: 1909.04847 [cs.LG].

Iovine, A. 2020. "Conversational Agents for Recommender Systems". In: *Fourteenth ACM Conference on Recommender Systems. RecSys '20*. Virtual Event, Brazil: Association for Computing Machinery. 758–763. ISBN: 9781450375832. DOI: 10.1145/3383313.3411453. URL: https://doi.org/10.1145/3383313.3411453.

Iyyer, M., W.-t. Yih, and M.-W. Chang. 2017. "Search-based neural structured learning for sequential question answering". In: *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1821–1831.

Izacard, G. and E. Grave. 2020. "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering". *ArXiv*. abs/2007.01282.

Jannach, D., L. Lerche, and M. Zanker. 2018. "Recommending Based on Implicit Feedback". In: *Social Information Access: Systems and Technologies*. Ed. by P. Brusilovsky and D. He. Cham: Springer International Publishing. 510–569. ISBN: 978-3-319-90092-6. DOI: 10.1007/978-3-319-90092-6_14. URL: https://doi.org/10.1007/978-3-319-90092-6_14.

Jannach, D., A. Manzoor, W. Cai, and L. Chen. 2021. "A Survey on Conversational Recommender Systems". *ACM Comput. Surv.* 54(5). ISSN: 0360-0300. DOI: 10.1145/3453154. URL: https://doi.org/10.1145/3453154.

Järvelin, K. and J. Kekäläinen. 2002. "Cumulated Gain-Based Evaluation of IR Techniques". *ACM Trans. Inf. Syst.* 20(4): 422–446. ISSN: 1046-8188. DOI: 10.1145/582415.582418. URL: https://doi.org/10.1145/582415.582418.

Järvelin, K., S. L. Price, L. M. L. Delcambre, and M. L. Nielsen. 2008. "Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions". In: *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval. ECIR'08*. Glasgow, UK: Springer-Verlag. 4–15. ISBN: 3540786457.

Jiang, J. and J. Allan. 2016. "Correlation Between System and User Metrics in a Session". In: *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval. CHIIR '16*. Carrboro, North Carolina, USA: Association for Computing Machinery. 285–288. ISBN: 9781450337519. DOI: 10.1145/2854946.2855005. URL: https://doi.org/10.1145/2854946.2855005.

Jiang, J., A. Hassan Awadallah, R. Jones, U. Ozertem, I. Zitouni, R. Gurunath Kulkarni, and O. Z. Khan. 2015. "Automatic Online Evaluation of Intelligent Assistants". In: *Proceedings of the 24th International Conference on World Wide Web. WWW '15*. Florence, Italy: International World Wide Web Conferences Steering Committee. 506–516. ISBN: 978-1-4503-3469-3. DOI: 10.1145/2736277.2741669. URL: https://doi.org/10.1145/2736277.2741669.

Jiang, J., W. Jeng, and D. He. 2013. "How Do Users Respond to Voice Input Errors? Lexical and Phonetic Query Reformulation in Voice Search". In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '13*. Dublin, Ireland: Association for Computing Machinery. 143–152. ISBN: 9781450320344. DOI: 10.1145/2484028.2484092. URL: https://doi.org/10.1145/2484028.2484092.

Joachims, T., L. Granka, B. Pan, H. Hembrooke, and G. Gay. 2005. "Accurately Interpreting Clickthrough Data as Implicit Feedback". In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '05*. Salvador, Brazil: Association for Computing Machinery. 154–161. ISBN: 1595930345. DOI: 10.1145/1076034.1076063. URL: https://doi.org/10.1145/1076034.1076063.

Joko, H., F. Hasibi, K. Balog, and A. P. de Vries. 2021. "Conversational Entity Linking: Problem Definition and Datasets". *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jones, M., G. Marsden, N. Mohd-Nasir, K. Boone, and G. Buchanan. 1999. "Improving Web Interaction on Small Displays". *Comput. Netw.* 31(11–16): 1129–1137. ISSN: 1389-1286. DOI: 10.1016/S1389-1286(99)00013-4. URL: https://doi.org/10.1016/S1389-1286(99)00013-4.

Ju, Y., F. Zhao, S. Chen, B. Zheng, X. Yang, and Y. Liu. 2019. "Technical report on Conversational Question Answering". *ArXiv*. abs/1909.10772.

Jurafsky, D. and J. H. Martin. 2021. *Speech and Language Processing (3rd Edition)*. USA: Prentice-Hall, Inc.

Kacupaj, E., J. Plepi, K. Singh, H. Thakkar, J. Lehmann, and M. Maleshkova. 2021. "Conversational Question Answering over Knowledge Graphs with Transformer and Graph Attention Networks". In: *EACL*.

Kaiser, M., R. S. Roy, and G. Weikum. 2020. "Conversational Question Answering over Passages by Leveraging Word Proximity Networks". SIGIR '20.

Kaiser, M., R. S. Roy, and G. Weikum. 2021. "Reinforcement Learning from Reformulations in Conversational Question Answering over Knowledge Graphs". *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Kaisser, M., M. A. Hearst, and J. B. Lowe. 2008. "Improving search results quality by customizing summary lengths". In: *Proceedings of ACL-08: HLT*. 701–709.

Kanoulas, E., B. Carterette, P. D. Clough, and M. Sanderson. 2011. "Evaluating Multi-Query Sessions". In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '11*. Beijing, China: Association for Computing Machinery. 1053–1062. ISBN: 9781450307574. DOI: 10.1145/2009916.2010056. URL: https://doi.org/10.1145/2009916.2010056.

Kaushik, A., V. Bhat Ramachandra, and G. J. F. Jones. 2020. "An Interface for Agent Supported Conversational Search". In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. CHIIR '20*. Vancouver BC, Canada: Association for Computing Machinery. 452–456. ISBN: 9781450368926. DOI: 10.1145/3343413.3377942. URL: https://doi.org/10.1145/3343413.3377942.

Keenoy, K. and M. Levene. 2003. "Personalisation of Web Search". In: *Proceedings of the 2003 International Conference on Intelligent Techniques for Web Personalization. ITWP'03*. Acapulco, Mexico: Springer-Verlag. 201–228. ISBN: 3540298460. DOI: 10.1007/11577935_11. URL: https://doi-org.ezp.lib.unimelb.edu.au/10.1007/11577935_11.

Kelly, D. 2009. "Methods for Evaluating Interactive Information Retrieval Systems with Users". *Found. Trends Inf. Retr.* 3(1—2): 1–224. ISSN: 1554-0669. DOI: 10.1561/1500000012. URL: http://dx.doi.org/10.1561/1500000012.

Kelly, D. and C. R. Sugimoto. 2013. "A systematic review of interactive information retrieval evaluation studies, 1967-2006". *J. Assoc. Inf. Sci. Technol.* 64(4): 745–770.

Khattab, O., C. Potts, and M. Zaharia. 2020. "Relevance-guided Supervision for OpenQA with ColBERT". *ArXiv*. abs/2007.00814.

Khattab, O., C. Potts, and M. Zaharia. 2021. “Baleen: Robust Multi-Hop Reasoning at Scale via Condensed Retrieval”. *ArXiv*. abs/2101.00436.

Kiesel, J., L. Meyer, M. Potthast, and B. Stein. 2021a. “Meta-Information in Conversational Search”. *ACM Transactions on Information Systems (ACM TOIS)*. 39(4). Ed. by C. Hauff, J. Kiseleva, M. Sanderson, H. Zamani, and Y. Zhang. ISSN: 1046-8188. DOI: 10.1145/3468868. URL: https://doi.org/10.1145/3468868.

Kiesel, J., D. Spina, H. Wachsmuth, and B. Stein. 2021b. “The Meant, the Said, and the Understood: Conversational Argument Search and Cognitive Biases”. In: *CUI 2021 - 3rd Conference on Conversational User Interfaces. CUI ’21*. Bilbao (online), Spain: Association for Computing Machinery. ISBN: 9781450389983. DOI: 10.1145/3469595.3469615. URL: https://doi.org/10.1145/3469595.3469615.

Kim, G., H. Kim, J. Park, and J. Kang. 2021. “Learn to Resolve Conversational Dependency: A Consistency Training Framework for Conversational Question Answering”. In: *ACL/IJCNLP*.

Kim, J., P. Thomas, R. Sankaranarayana, T. Gedeon, and H.-J. Yoon. 2015. “Eye-Tracking Analysis of User Behavior and Performance in Web Search on Large and Small Screens”. *J. Assoc. Inf. Sci. Technol.* 66(3): 526–544. ISSN: 2330-1635. DOI: 10.1002/asi.23187. URL: https://doi.org/10.1002/asi.23187.

Kim, J., P. Thomas, R. Sankaranarayana, T. Gedeon, and H.-J. Yoon. 2017. “What Snippet Size is Needed in Mobile Web Search?” In: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. CHIIR ’17*. Oslo, Norway: Association for Computing Machinery. 97–106. ISBN: 9781450346771. DOI: 10.1145/3020165.3020173. URL: https://doi.org/10.1145/3020165.3020173.

Kim, S., M. Eric, K. Gopalakrishnan, B. Hedayatnia, Y. Liu, and D. Z. Hakkani-Tür. 2020. “Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access”. *ArXiv*. abs/2006.03533.

Kiseleva, J., K. Williams, A. Hassan Awadallah, A. C. Crook, I. Zitouni, and T. Anastasakos. 2016. "Predicting User Satisfaction with Intelligent Assistants". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '16*. Pisa, Italy: ACM. 45–54. ISBN: 978-1-4503-4069-4. DOI: 10.1145/2911451.2911521. URL: http://doi.acm.org/10.1145/2911451.2911521.

Kitano, H. and C. Van Ess-Dykema. 1991. "Toward a Plan-Based Understanding Model for Mixed-Initiative Dialogues". In: *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics. ACL '91*. Berkeley, California: Association for Computational Linguistics. 25–32. DOI: 10.3115/981344.981348.

Kolomiyets, O. and M.-F. Moens. 2011. "A Survey on Question Answering Technology from an Information Retrieval Perspective". *Inf. Sci.* 181(24): 5412–5434. ISSN: 0020-0255. DOI: 10.1016/j.ins.2011.07.047. URL: https://doi.org/10.1016/j.ins.2011.07.047.

Koman, J., K. Fauvelle, S. Schuck, N. Texier, and A. Mebarki. 2020. "Physicians' Perceptions of the Use of a Chatbot for Information Seeking: Qualitative Study". *Journal of medical Internet research.* 22(11): e15185.

Konstan, J. A. and J. Riedl. 2012. "Recommender Systems: From Algorithms to User Experience". *User Modeling and User-Adapted Interaction.* 22(1–2): 101–123. ISSN: 0924-1868. DOI: 10.1007/s11257-011-9112-x. URL: https://doi-org.ezp.lib.unimelb.edu.au/10.1007/s11257-011-9112-x.

Krishna, K., A. Roy, and M. Iyyer. 2021. "Hurdles to Progress in Long-form Question Answering". *ArXiv.* abs/2103.06332.

Krum, U., H. Holzapfel, and A. Waibel. 2005. "Clarification questions to improve dialogue flow and speech recognition in spoken dialogue systems". In: *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*. ISCA. 3417–3420.

Ku, A., P. Anderson, R. Patel, E. Ie, and J. Baldridge. 2020. "Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics. 4392–4412. DOI: 10.18653/v1/2020.emnlp-main.356. URL: https://www.aclweb.org/anthology/2020.emnlp-main.356.

Kumar, V. and J. Callan. 2020. "Making Information Seeking Easier: An Improved Pipeline for Conversational Search". In: *EMNLP*.

Laban, G. and T. Araujo. 2020. "The Effect of Personalization Techniques in Users' Perceptions of Conversational Recommender Systems". In: *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents. IVA '20*. Virtual Event, Scotland, UK: Association for Computing Machinery. ISBN: 9781450375863. DOI: 10.1145/3383652.3423890. URL: https://doi.org/10.1145/3383652.3423890.

Lai, J., C. Karat, and N. Yankelovich. 2009. "Conversational speech interfaces and technologies". *Human-Computer Interaction: Design Issues, Solutions, and Applications*: 53–63.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2020. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *International Conference on Learning Representations. ICLR '20*. URL: https://openreview.net/forum?id=H1eA7AEtvS.

Lei, W., X. He, Y. Miao, Q. Wu, R. Hong, M.-Y. Kan, and T.-S. Chua. 2020a. "Estimation-Action-Reflection: Towards Deep Interaction Between Conversational and Recommender Systems". In: *Proceedings of the 13th International Conference on Web Search and Data Mining. WSDM '20*. Houston, TX, USA: Association for Computing Machinery. 304–312. ISBN: 9781450368223. DOI: 10.1145/3336191.3371769. URL: https://doi.org/10.1145/3336191.3371769.

Lei, W., G. Zhang, X. He, Y. Miao, X. Wang, L. Chen, and T.-S. Chua. 2020b. "Interactive Path Reasoning on Graph for Conversational Recommendation". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '20*. Virtual Event, CA, USA: Association for Computing Machinery. 2073–2083. ISBN: 9781450379984. DOI: 10.1145/3394486.3403258. URL: https://doi.org/10.1145/3394486.3403258.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". *ArXiv*. abs/2005.11401.

Li, S., W. Lei, Q. Wu, X. He, P. Jiang, and T.-S. Chua. 2021. "Seamlessly Unifying Attributes and Items: Conversational Recommendation for Cold-Start Users". *ACM Trans. Inf. Syst.* 39(4). ISSN: 1046-8188. DOI: 10.1145/3446427. URL: https://doi.org/10.1145/3446427.

Li, X., G. Tür, D. Z. Hakkani-Tür, and Q. Li. 2014. "Personal knowledge graph population from user utterances in conversational understanding". *2014 IEEE Spoken Language Technology Workshop (SLT)*: 224–229.

Liao, L., L. H. Long, Z. Zhang, M. Huang, and T.-S. Chua. 2021. "MMConv: an Environment for Multimodal Conversational Search across Multiple Domains". In: *Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR'21)*. 1–10.

Lin, C.-Y. 2004. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics. 74–81. URL: https://www.aclweb.org/anthology/W04-1013.

Lin, J., R. Nogueira, and A. Yates. 2020a. "Pretrained Transformers for Text Ranking: BERT and Beyond". *CoRR*. abs/2010.06467. URL: https://arxiv.org/abs/2010.06467.

Lin, S.-C., J.-H. Yang, R. Nogueira, M.-F. Tsai, C.-J. Wang, and J. Lin. 2020b. "Query Reformulation using Query History for Passage Retrieval in Conversational Search". *ArXiv*. abs/2005.02230.

Lin, S.-C., J.-H. Yang, R. Nogueira, M.-F. Tsai, C.-J. Wang, and J. J. Lin. 2020c. "Query Reformulation using Query History for Passage Retrieval in Conversational Search". *ArXiv*. abs/2005.02230.

Lin, S.-C., J.-H. Yang, R. Nogueira, M.-F. Tsai, C.-J. Wang, and J. J. Lin. 2021. "Multi-Stage Conversational Passage Retrieval: An Approach to Fusing Term Importance Estimation and Neural Query Rewriting". *ACM Transactions on Information Systems (TOIS)*. 39: 1–29.

Liono, J., M. S. Rahaman, F. D. Salim, Y. Ren, D. Spina, F. Scholer, J. R. Trippas, M. Sanderson, P. N. Bennett, and R. W. White. 2020. "Intelligent Task Recognition: Towards Enabling Productivity Assistance in Daily Life". In: *International Conference on Multimedia Retrieval (ICMR'20)*.

Lipani, A., B. Carterette, and E. Yilmaz. 2019. "From a User Model for Query Sessions to Session Rank Biased Precision (SRBP)". In: *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval. ICTIR '19*. Santa Clara, CA, USA: Association for Computing Machinery. 109–116. ISBN: 9781450368810. DOI: 10.1145/3341981.3344216. URL: https://doi.org/10.1145/3341981.3344216.

Lipani, A., B. Carterette, and E. Yilmaz. 2021. "How Am I Doing?: Evaluating Conversational Search Systems Offline". *ACM Transactions on Information Systems*.

Liu, C.-W., R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau. 2016. "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics. 2122–2132. DOI: 10.18653/v1/D16-1230. URL: https://www.aclweb.org/anthology/D16-1230.

Liu, Z., K. Zhou, and M. L. Wilson. 2021. "Meta-Evaluation of Conversational Search Evaluation Metrics". *ACM Trans. Inf. Syst.* 39(4). ISSN: 1046-8188. DOI: 10.1145/3445029. URL: https://doi.org/10.1145/3445029.

Louis, A., F. Radlinski, and D. Roth. 2020. ""I'd rather just go to bed": Understanding Indirect Answers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Lu, X., S. Pramanik, R. Saha Roy, A. Abujabal, Y. Wang, and G. Weikum. 2019. "Answering Complex Questions by Joining Multi-Document Evidence with Quasi Knowledge Graphs". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR'19*. Paris, France: Association for Computing Machinery. 105–114. ISBN: 9781450361729. DOI: 10.1145/3331184.3331252.

Ma, Q., O. Bojar, and Y. Graham. 2018. "Results of the WMT18 Metrics Shared Task: Both characters and embeddings achieve good performance". In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. 671–688. DOI: 10.18653/v1/W18-6450.

Mahmood, T. and F. Ricci. 2009. "Improving Recommender Systems with Adaptive Conversational Strategies". In: *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia. HT '09*. Torino, Italy: Association for Computing Machinery. 73–82. ISBN: 9781605584867. DOI: 10.1145/1557914.1557930. URL: https://doi.org/10.1145/1557914.1557930.

Mallinson, J., A. Severyn, E. Malmi, and G. Garrido. 2020. "Felix: Flexible Text Editing Through Tagging and Insertion". In: *EMNLP*.

Malmi, E., S. Krause, S. Rothe, D. Mirylenka, and A. Severyn. 2019. "Encode, Tag, Realize: High-Precision Text Editing". In: *EMNLP/IJCNLP*.

Mandya, A., J. O'Neill, D. Bollegala, and F. Coenen. 2020. "Do not let the history haunt you - Mitigating Compounding Errors in Conversational Question Answering". In: *LREC*.

Marchionini, G. 2006. "Exploratory Search: From Finding to Understanding". *Commun. ACM*. 49(4): 41–46. ISSN: 0001-0782. DOI: 10.1145/1121949.1121979. URL: http://doi.acm.org/10.1145/1121949.1121979.

Marion, P., P. Nowak, and F. Piccinno. 2021. "Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs". *ArXiv*. abs/2109.00269.

Matteson, M. L., J. Salamon, and L. Brewster. 2011. "A Systematic Review of Research on Live Chat Service". *Reference & User Services Quarterly*. 51(2): 172–190. ISSN: 10949054, 21635242. URL: http://www.jstor.org/stable/refuseserq.51.2.172.

Maxwell, D., L. Azzopardi, and Y. Moshfeghi. 2017. "A Study of Snippet Length and Informativeness: Behaviour, Performance and User Experience". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '17*. Shinjuku, Tokyo, Japan: Association for Computing Machinery. 135–144. ISBN: 9781450350228. DOI: 10.1145/3077136.3080824. URL: https://doi.org/10.1145/3077136.3080824.

McTear, M., Z. Callejas, and D. Griol. 2016. *The Conversational Interface: Talking to Smart Devices*. Springer.

McTear, M. F. 2017. "The Rise of the Conversational Interface: A New Kid on the Block?" In: *Future and Emerging Trends in Language Technology. Machine Learning and Big Data*. Ed. by J. F. Quesada, F.-J. Martín Mateos, and T. López Soto. Springer International Publishing. 38–49. ISBN: 978-3-319-69365-1.

Mehri, S., M. Eric, and D. Hakkani-Tur. 2020. "DialoGLUE: A Natural Language Understanding Benchmark for Task-Oriented Dialogue". *ArXiv*. abs/2009.13570.

Mele, I., C. I. Muntean, F. Nardini, R. Perego, N. Tonellotto, and O. Frieder. 2020. "Topic Propagation in Conversational Search". *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Mirzadeh, N., F. Ricci, and M. Bansal. 2005. "Feature selection methods for conversational recommender systems". In: *2005 IEEE International Conference on e-Technology, e-Commerce and e-Service*. 772–777. DOI: 10.1109/EEE.2005.75.

Mitra, B. and N. Craswell. 2018. "An Introduction to Neural Information Retrieval". *Foundations and Trends® in Information Retrieval*. 13(1): 1–126. ISSN: 1554-0669. DOI: 10.1561/1500000061. URL: http://dx.doi.org/10.1561/1500000061.

Mitra, B., S. Hofstätter, H. Zamani, and N. Craswell. 2021. "Improving Transformer-Kernel Ranking Model Using Conformer and Query Term Independence". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. Virtual Event, Canada: Association for Computing Machinery. 1697–1702. ISBN: 9781450380379. DOI: 10.1145/3404835.3463049. URL: https://doi.org/10.1145/3404835.3463049.

Mori, M., K. F. MacDorman, and N. Kageki. 2012. "The Uncanny Valley [From the Field]". *IEEE Robotics Automation Magazine*. 19(2): 98–100. DOI: 10.1109/MRA.2012.2192811.

Morris, D., M. Ringel Morris, and G. Venolia. 2008. "SearchBar: A Search-Centric Web History for Task Resumption and Information Re-Finding". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '08*. Florence, Italy: Association for Computing Machinery. 1207–1216. ISBN: 9781605580111. DOI: 10.1145/1357054.1357242. URL: https://doi.org/10.1145/1357054.1357242.

Müller, M. and M. Volk. 2013. "Statistical Machine Translation of Subtitles: From OpenSubtitles to TED". In: *GSCL*.

Murray, G. C. and J. Teevan. 2007. "Query Log Analysis: Social and Technological Challenges". *SIGIR Forum*. 41(2): 112–120. ISSN: 0163-5840. DOI: 10.1145/1328964.1328985. URL: https://doi.org/10.1145/1328964.1328985.

Myers, C., A. Furqan, J. Nebolsky, K. Caro, and J. Zhu. 2018. "Patterns for How Users Overcome Obstacles in Voice User Interfaces". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery. 1–7. ISBN: 9781450356206. URL: https://doi.org/10.1145/3173574.3173580.

Nag, P. and Ö. N. Yalçın. 2020. "Gender Stereotypes in Virtual Agents". In: *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents. IVA '20*. Virtual Event, Scotland, UK: Association for Computing Machinery. ISBN: 9781450375863. DOI: 10.1145/3383652.3423876. URL: https://doi.org/10.1145/3383652.3423876.

Nass, C. and S. Brave. 2005. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. The MIT Press. ISBN: 0262140926.

Nie, L., W. Wang, R. Hong, M. Wang, and Q. Tian. 2019. "Multimodal dialog system: Generating responses via adaptive decoders". In: *Proceedings of the 27th ACM International Conference on Multimedia*. 1098–1106.

Novick, D. G. and S. A. Douglas. 1988. "Control of Mixed-Initiative Discourse through Meta-Locutionary Acts: A Computational Model". *Tech. rep.* AAI8911322. USA.

Oard, D. W. and J. Kim. 1998. "Implicit Feedback for Recommender Systems". In: *Proceedings of the AAAI Workshop on Recommender Systems*. AAAI.

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

Ohsugi, Y., I. Saito, K. Nishida, H. Asano, and J. Tomita. 2019. "A Simple but Effective Method to Incorporate Multi-turn Context with BERT for Conversational Machine Comprehension". *ArXiv*. abs/1905.12848.

Ong, K., K. Järvelin, M. Sanderson, and F. Scholer. 2017. "Using Information Scent to Understand Mobile and Desktop Web Search Behavior". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '17*. Shinjuku, Tokyo, Japan: Association for Computing Machinery. 295–304. ISBN: 9781450350228. DOI: 10.1145/3077136.3080817.

Oraby, S., P. Gundecha, J. Mahmud, M. Bhuiyan, and R. Akkiraju. 2017. ""How May I Help You?" Modeling Twitter Customer Service Conversations Using Fine-Grained Dialogue Acts". In: *Proceedings of the 22nd international conference on intelligent user interfaces*. 343–355.

Ouyang, S., Z. Zhang, and H. Zhao. 2021. "Dialogue Graph Modeling for Conversational Machine Reading". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics. 3158–3169. DOI: 10.18653/v1/2021.findings-acl.279. URL: https://aclanthology.org/2021.findings-acl.279.

Oviatt, S. and P. R. Cohen. 2015. *The Paradigm Shift to Multimodality in Contemporary Computer Interfaces*. Morgan & Claypool Publishers. ISBN: 162705751X.

Paek, T., S. Dumais, and R. Logan. 2004. "WaveLens: A New View onto Internet Search Results". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '04*. Vienna, Austria: Association for Computing Machinery. 727–734. ISBN: 1581137028. DOI: 10.1145/985692.985784. URL: https://doi.org/10.1145/985692.985784.

Pajukoski, J. 2018. "Impact of chat layout on usability in customer service chat multitasking". *MA thesis*. Aalto University. 85.

Papenmeier, A., D. Kern, D. Hienert, A. Sliwa, A. Aker, and N. Fuhr. 2021. "Dataset of Natural Language Queries for E-Commerce". In: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. CHIIR '21*. Canberra ACT, Australia: Association for Computing Machinery. 307–311. ISBN: 9781450380553. DOI: 10.1145/3406522.3446043. URL: https://doi.org/10.1145/3406522.3446043.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. 311–318. DOI: 10.3115/1073083.1073135. URL: https://www.aclweb.org/anthology/P02-1040.

Parikh, A. P., O. Täckström, D. Das, and J. Uszkoreit. 2016. "A Decomposable Attention Model for Natural Language Inference". In: *EMNLP*.

Park, D., H. Yuan, D. Kim, Y. Zhang, M. Spyros, Y.-B. Kim, R. Sarikaya, E. Guo, Y. Ling, K. Quinn, P. Hung, B. Yao, and S. Lee. 2020. "Large-scale Hybrid Approach for Predicting User Satisfaction with Conversational Agents". arXiv: 2006.07113.

Peckham, J. 1991. "Speech Understanding and Dialogue over the telephone: an overview of the ESPRIT SUNDIAL project." In: *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

Peshterliev, S., B. Oğuz, D. Chatterjee, H. Inan, and V. Bhardwaj. 2021. "Conversational Answer Generation and Factuality for Reading Comprehension Question-Answering". In:

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. "Deep contextualized word representations". In: *NAACL-HLT*.

Plepi, J., E. Kacupaj, K. Singh, H. Thakkar, and J. Lehmann. 2021. "Context Transformer with Stacked Pointer Networks for Conversational Question Answering over Knowledge Graphs". In: *ESWC*.

Pommeranz, A., J. Broekens, P. Wiggers, W.-P. Brinkman, and C. M. Jonker. 2012. "Designing Interfaces for Explicit Preference Elicitation: A User-Centered Investigation of Preference Representation and Elicitation Process". *User Modeling and User-Adapted Interaction*. 22(4–5): 357–397. ISSN: 0924-1868. DOI: 10.1007/s11257-011-9116-6. URL: https://doi.org/10.1007/s11257-011-9116-6.

Ponte, J. M. and W. B. Croft. 1998. "A Language Modeling Approach to Information Retrieval". In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '98*. Melbourne, Australia: Association for Computing Machinery. 275–281. ISBN: 1581130155. DOI: 10.1145/290941.291008. URL: https://doi.org/10.1145/290941.291008.

Prakash, P., J. Killingback, and H. Zamani. 2021. "Learning Robust Dense Retrieval Models from Incomplete Relevance Labels". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21*. Virtual Event, Canada: Association for Computing Machinery. 1728–1732. ISBN: 9781450380379. DOI: 10.1145/3404835.3463106. URL: https://doi.org/10.1145/3404835.3463106.

Qu, C., L. Yang, C. Chen, M. Qiu, W. Croft, and M. Iyyer. 2020. "Open-Retrieval Conversational Question Answering". *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Qu, C., L. Yang, C. Chen, W. Croft, K. Krishna, and M. Iyyer. 2021. "Weakly-Supervised Open-Retrieval Conversational Question Answering". *ArXiv*. abs/2103.02537.

Qu, C., L. Yang, W. B. Croft, J. R. Trippas, Y. Zhang, and M. Qiu. 2018. "Analyzing and Characterizing User Intent in Information-seeking Conversations". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18*. Ann Arbor, MI, USA: ACM. 989–992. ISBN: 978-1-4503-5657-2. DOI: 10.1145/3209978.3210124. URL: http://doi.acm.org/10.1145/3209978.3210124.

Qu, C., L. Yang, M. Qiu, W. B. Croft, Y. Zhang, and M. Iyyer. 2019a. "BERT with History Answer Embedding for Conversational Question Answering". In: *Proceedings of the 42nd ACM SIGIR International Conference on Research and Development on Information Retrieval (to appear). SIGIR '19*.

Qu, C., L. Yang, M. Qiu, Y. Zhang, C. Chen, W. B. Croft, and M. Iyyer. 2019b. "Attentive History Selection for Conversational Question Answering". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management. CIKM '19*. Beijing, China: Association for Computing Machinery. 1391–1400. ISBN: 9781450369763. DOI: 10.1145/3357384.3357905.

Radlinski, F., K. Balog, B. Byrne, and K. Krishnamoorthi. 2019. "Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences". In: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Stockholm, Sweden: Association for Computational Linguistics. 353–360. DOI: 10.18653/v1/W19-5941. URL: https://www.aclweb.org/anthology/W19-5941.

Radlinski, F. and N. Craswell. 2017. "A Theoretical Framework for Conversational Search". In: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. CHIIR '17*. Oslo, Norway: Association for Computing Machinery. 117–126. ISBN: 9781450346771. DOI: 10.1145/3020165.3020183. URL: https://doi.org/10.1145/3020165.3020183.

Rae, J. W., A. Potapenko, S. M. Jayakumar, and T. Lillicrap. 2020. "Compressive Transformers for Long-Range Sequence Modelling". *ArXiv*. abs/1911.05507.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". *J. Mach. Learn. Res.* 21: 140:1–140:67.

Ram, A., R. Prasad, C. Khatri, A. Venkatesh, R. Gabriel, Q. Liu, J. Nunn, B. Hedayatnia, M. Cheng, A. Nagar, E. King, K. Bland, A. Wartick, Y. Pan, H. Song, S. Jayadevan, G. Hwang, and A. Pettigrue. 2018. "Conversational AI: The Science Behind the Alexa Prize". *ArXiv*. abs/1801.03604.

Ranzato, M., S. Chopra, M. Auli, and W. Zaremba. 2016. "Sequence Level Training with Recurrent Neural Networks". In: *Proceedings of the 4th International Conference on Learning Representations*. Ed. by Y. Bengio and Y. LeCun. *ICLR '16*. San Juan, Puerto Rico.

Rao, S. and H. Daumé III. 2018. "Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics. 2737–2746. URL: https://www.aclweb.org/anthology/P18-1255.

Rao, S. and H. Daumé III. 2019. "Answer-based Adversarial Training for Generating Clarification Questions". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics. 143–155. URL: https://www.aclweb.org/anthology/N19-1013.

Rashkin, H., V. Nikolaev, M. Lamm, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, and D. Reitter. 2021. "Measuring Attribution in Natural Language Generation Models". arXiv: 2112.12870 [cs.CL].

Rastogi, A., X. Zang, S. Sunkara, R. Gupta, and P. Khaitan. 2020. "Schema-Guided Dialogue State Tracking Task at DSTC8". *ArXiv*. abs/2002.01359.

Rastogi, P., A. Gupta, T. Chen, and L. Mathias. 2019. "Scaling Multi-Domain Dialogue State Tracking via Query Reformulation". In: *NAACL-HLT*.

Reddy, S., D. Chen, and C. D. Manning. 2019. "CoQA: A Conversational Question Answering Challenge". *Transactions of the Association for Computational Linguistics*. 7(Mar.): 249–266. DOI: 10.1162/tacl_a_00266. URL: https://www.aclweb.org/anthology/Q19-1016.

Reichman, R. 1985. *Getting Computers to Talk Like You and Me: Discourse Context, Focus, and Semantics:(an ATN Model)*. MIT press.

Ren, P., Z. Chen, Z. Ren, E. Kanoulas, C. Monz, and M. Rijke. 2020a. "Conversations with Search Engines: SERP-based Conversational Response Generation". *ArXiv*. abs/2004.14162.

Ren, P., Z. Liu, X. Song, H. Tian, Z. Chen, Z. Ren, and M. de Rijke. 2021. "Wizard of Search Engine: Access to Information Through Conversations with Search Engines". *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Ren, X., H. Yin, T. Chen, H. Wang, N. Q. V. Hung, Z. Huang, and X. Zhang. 2020b. "CRSAL: Conversational Recommender Systems with Adversarial Learning". *ACM Trans. Inf. Syst.* 38(4). ISSN: 1046-8188. DOI: 10.1145/3394592. URL: https://doi.org/10.1145/3394592.

Resnick, P., N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. 1994. "GroupLens: An Open Architecture for Collaborative Filtering of Netnews". In: *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. CSCW '94*. Chapel Hill, North Carolina, USA: Association for Computing Machinery. 175–186. ISBN: 0897916891. DOI: 10.1145/192844.192905. URL: https://doi.org/10.1145/192844.192905.

Resnick, P. and H. R. Varian. 1997. "Recommender Systems". *Commun. ACM*. 40(3): 56–58. ISSN: 0001-0782. DOI: 10.1145/245108.245121. URL: https://doi.org/10.1145/245108.245121.

Ricci, F., L. Rokach, B. Shapira, and P. B. Kantor. 2010. *Recommender Systems Handbook*. 1st. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387858199.

Rieser, V. and J. Moore. 2005. "Implications for Generating Clarification Requests in Task-Oriented Dialogues". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, Michigan: Association for Computational Linguistics. 239–246. DOI: 10.3115/1219840.1219870. URL: https://www.aclweb.org/anthology/P05-1030.

Roller, S., E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith, Y.-L. Boureau, and J. Weston. 2020. "Recipes for building an open-domain chatbot". *ArXiv*. abs/2004.13637.

Rose, D. E., D. Orr, and R. G. P. Kantamneni. 2007. "Summary Attributes and Perceived Search Quality". In: *Proceedings of the 16th International Conference on World Wide Web. WWW '07*. Banff, Alberta, Canada: Association for Computing Machinery. 1201–1202. ISBN: 9781595936547. DOI: 10.1145/1242572.1242765. URL: https://doi.org/10.1145/1242572.1242765.

Rosset, C., C. Xiong, X. Song, D. Campos, N. Craswell, S. Tiwary, and P. Bennett. 2020. "Leading Conversational Search by Suggesting Useful Questions". In: *Proceedings of The Web Conference 2020. WWW '20*. Taipei, Taiwan: Association for Computing Machinery. 1160–1170. ISBN: 9781450370233. DOI: 10.1145/3366423.3380193. URL: https://doi.org/10.1145/3366423.3380193.

Rousseau, C., Y. Bellik, F. Vernier, and D. Bazalgette. 2006. "A Framework for the Intelligent Multimodal Presentation of Information". *Signal Process.* 86(12): 3696–3713. ISSN: 0165-1684. DOI: 10.1016/j.sigpro.2006.02.041. URL: https://doi.org/10.1016/j.sigpro.2006.02.041.

Rudnicky, A. I. 2005. "Multimodal Dialogue Systems". In: *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*. Ed. by W. Minker, D. Bühler, and L. Dybkjær. Dordrecht: Springer Netherlands. 3–11. ISBN: 978-1-4020-3075-8. DOI: 10.1007/1-4020-3075-4_1. URL: https://doi.org/10.1007/1-4020-3075-4_1.

Sachse, J. 2019. "The influence of snippet length on user behavior in mobile web search". *Aslib Journal of Information Management*.

Saeidi, M., M. Bartolo, P. Lewis, S. Singh, T. Rocktäschel, M. Sheldon, G. Bouchard, and S. Riedel. 2018. "Interpretation of Natural Language Rules in Conversational Machine Reading". In: *EMNLP*.

Saha, A., V. Pahuja, M. M. Khapra, K. Sankaranarayanan, and A. P. S. Chandar. 2018. "Complex Sequential Question Answering: Towards Learning to Converse Over Linked Question Answer Pairs with a Knowledge Graph". In: *AAAI*.

Salton, G. 1970. "Evaluation problems in interactive information retrieval". *Information Storage and Retrieval*. 6(1): 29–44. ISSN: 0020-0271. DOI: https://doi.org/10.1016/0020-0271(70)90011-2.

Schaffer, S. and N. Reithinger. 2019. "Conversation is Multimodal: Thus Conversational User Interfaces Should Be as Well". In: *Proceedings of the 1st International Conference on Conversational User Interfaces. CUI '19*. Dublin, Ireland: Association for Computing Machinery. ISBN: 9781450371872. DOI: 10.1145/3342775.3342801. URL: https://doi.org/10.1145/3342775.3342801.

Schnabel, T., P. N. Bennett, S. T. Dumais, and T. Joachims. 2016. "Using shortlists to support decision making and improve recommender system performance". In: *Proceedings of the 25th International Conference on World Wide Web*. 987–997.

See, A., P. J. Liu, and C. D. Manning. 2017. "Get To The Point: Summarization with Pointer-Generator Networks". In: *ACL*.

Sellam, T., D. Das, and A. Parikh. 2020. "BLEURT: Learning Robust Metrics for Text Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7881–7892.

Seo, M., A. Kembhavi, A. Farhadi, and H. Hajishirzi. 2017. "Bidirectional Attention Flow for Machine Comprehension". *ArXiv*. abs/1611.01603.

Sepliarskaia, A., J. Kiseleva, F. Radlinski, and M. de Rijke. 2018. "Preference Elicitation as an Optimization Problem". In: *Proceedings of the 12th ACM Conference on Recommender Systems. RecSys '18*. Vancouver, British Columbia, Canada: Association for Computing Machinery. 172–180. ISBN: 9781450359016. DOI: 10.1145/3240323.3240352. URL: https://doi.org/10.1145/3240323.3240352.

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

Shen, T., X. Geng, T. Qin, D. Guo, D. Tang, N. Duan, G. Long, and D. Jiang. 2019. "Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base". In: *EMNLP/IJCNLP*.

Shen, X., B. Tan, and C. Zhai. 2005. "Context-sensitive Information Retrieval Using Implicit Feedback". In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '05*. Salvador, Brazil: ACM. 43–50. ISBN: 1-59593-034-5. DOI: 10.1145/1076034.1076045. URL: http://doi.acm.org/10.1145/1076034.1076045.

Shuster, K., S. Poff, M. Chen, D. Kiela, and J. Weston. 2021. "Retrieval Augmentation Reduces Hallucination in Conversation". *ArXiv*. abs/2104.07567.

Skantze, G. 2007. "Error Handling in Spoken Dialogue Systems-Managing Uncertainty, Grounding and Miscommunication". *PhD thesis*. KTH, Stockholm.

Slokom, M. 2018. "Comparing Recommender Systems Using Synthetic Data". In: *Proceedings of the 12th ACM Conference on Recommender Systems. RecSys '18*. 548–552.

Sordoni, A., Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, and J.-Y. Nie. 2015. "A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. CIKM '15*. Melbourne, Australia: Association for Computing Machinery. 553–562. ISBN: 9781450337946. DOI: 10.1145/2806416.2806493. URL: https://doi.org/10.1145/2806416.2806493.

Spina, D., J. R. Trippas, P. Thomas, H. Joho, K. Byström, L. Clark, N. Craswell, M. Czerwinski, D. Elsweiler, A. Frummet, S. Ghosh, J. Kiesel, I. Lopatovska, D. McDuff, S. Meyer, A. Mourad, P. Owoicho, S. P. Cherumanal, D. Russell, and L. Sitbon. 2021. "Report on the Future Conversations Workshop at CHIIR 2021". *SIGIR Forum*. 55(1). ISSN: 0163-5840. DOI: 10.1145/3476415.3476421. URL: https://doi.org/10.1145/3476415.3476421.

Staliunaite, I. and I. Iacobacci. 2020. "Compositional and Lexical Semantics in RoBERTa, BERT and DistilBERT: A Case Study on CoQA". In: *EMNLP*.

Stephen, J., K. Collie, D. McLeod, A. Rojubally, K. Fergus, M. Speca, J. Turner, J. Taylor-Brown, S. Sellick, K. Burrus, and M. Elramly. 2014. "Talking with text: Communication in therapist-led, live chat cancer support groups". *Social Science & Medicine*. 104: 178–186. ISSN: 0277-9536. DOI: https://doi.org/10.1016/j.socscimed.2013.12.001. URL: https://www.sciencedirect.com/science/article/pii/S0277953613006692.

Stoyanchev, S., A. Liu, and J. Hirschberg. 2014. "Towards Natural Clarification Questions in Dialogue Systems". In: *AISB '14*. Vol. 20. London, UK.

Su, H., X. Shen, R. Zhang, F. Sun, P. Hu, C. Niu, and J. Zhou. 2019. "Improving Multi-turn Dialogue Modelling with Utterance ReWriter". In: *ACL*.

Su, Z., J. A. Schneider, and S. D. Young. 2021. "The Role of Conversational Agents for Substance Use Disorder in Social Distancing Contexts". *Substance Use & Misuse*. 56(11): 1732–1735.

Sun, Y. and Y. Zhang. 2018. "Conversational Recommender System". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18*. Ann Arbor, MI, USA: Association for Computing Machinery. 235–244. ISBN: 9781450356572. DOI: 10.1145/3209978.3210002. URL: https://doi.org/10.1145/3209978.3210002.

Sutskever, I., O. Vinyals, and Q. V. Le. 2014. "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. *NeurIPS '14*. Curran Associates, Inc. 3104–3112. URL: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

Sutton, S. and R. Cole. 1997. "The CSLU Toolkit: Rapid Prototyping of Spoken Language Systems". In: *Proceedings of the 10th Annual ACM Symposium on User Interface Software and Technology. UIST '97*. Banff, Alberta, Canada: Association for Computing Machinery. 85–86. ISBN: 0897918819. DOI: 10.1145/263407.263517. URL: https://doi.org/10.1145/263407.263517.

Tabassum, M., T. Kosiundefinedski, A. Frik, N. Malkin, P. Wijesekera, S. Egelman, and H. R. Lipford. 2019. "Investigating Users' Preferences and Expectations for Always-Listening Voice Assistants". *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3(4). DOI: 10.1145/3369807. URL: https://doi-org.ezp.lib.unimelb.edu.au/10.1145/3369807.

Tao, C., J. Feng, R. Yan, W. Wu, and D. Jiang. 2021. "A Survey on Response Selection for Retrieval-based Dialogues". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Z.-H. Zhou. Survey Track. International Joint Conferences on Artificial Intelligence Organization. 4619–4626. URL: https://doi.org/10.24963/ijcai.2021/627.

Tavakoli, L., H. Zamani, F. Scholer, W. B. Croft, and M. Sanderson. 2021. "Analyzing clarification in asynchronous information-seeking conversations". *Journal of the Association for Information Science and Technology*. DOI: https://doi.org/10.1002/asi.24562.

Teevan, J. 2020. "Conversational Search in the Enterprise". In: *Conversational Search (Dagstuhl Seminar 19461)*. Ed. by A. Anand, L. Cavedon, H. Joho, M. Sanderson, and B. Stein. Dagstuhl. 47.

Teevan, J., E. Adar, R. Jones, and M. A. S. Potts. 2007. "Information Re-Retrieval: Repeat Queries in Yahoo's Logs". In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '07*. Amsterdam, The Netherlands: Association for Computing Machinery. 151–158. ISBN: 9781595935977. DOI: 10.1145/1277741.1277770. URL: https://doi.org/10.1145/1277741.1277770.

Teevan, J., E. Cutrell, D. Fisher, S. M. Drucker, G. Ramos, P. André, and C. Hu. 2009. "Visual Snippets: Summarizing Web Pages for Search and Revisitation". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '09*. Boston, MA, USA: Association for Computing Machinery. 2023–2032. ISBN: 9781605582467. DOI: 10.1145/1518701.1519008. URL: https://doi.org/10.1145/1518701.1519008.

Tenney, I., D. Das, and E. Pavlick. 2019. "BERT Rediscovers the Classical NLP Pipeline". In: *ACL*.

Thomas, P., M. Czerwinski, D. McDuff, N. Craswell, and G. Mark. 2018. "Style and Alignment in Information-Seeking Conversation". In: *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval. CHIIR '18*. New Brunswick, NJ, USA: Association for Computing Machinery. 42–51. ISBN: 9781450349253. DOI: 10.1145/3176349.3176388. URL: https://doi.org/10.1145/3176349.3176388.

Thompson, C. A., M. H. Göker, and P. Langley. 2004. "A Personalized System for Conversational Recommendations". *J. Artif. Int. Res.* 21(1): 393–428. ISSN: 1076-9757.

Traum, D. and P. Heeman. 1996. "Utterance Units in Spoken Dialogue". In: *ECAI Workshop on Dialogue Processing in Spoken Language Systems*.

Traum, D. R. and S. Larsson. 2003. "The information state approach to dialogue management". In: *Current and New Directions in Discourse and Dialogue*. Springer. 325–353.

Tredici, M. D., G. Barlacchi, X. Shen, W. Cheng, and A. Gispert. 2021. "Question Rewriting for Open-Domain Conversational QA: Best Practices and Limitations". *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

Trienes, J. and K. Balog. 2019. "Identifying Unclear Questions in Community Question Answering Websites". In: *Advances in Information Retrieval*. Ed. by L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, and D. Hiemstra. Cham: Springer International Publishing. 276–289. ISBN: 978-3-030-15712-8.

Trippas, J. R. 2019. "Spoken Conversational Search: Audio-only Interactive Information Retrieval". *PhD thesis*. RMIT, Melbourne.

Trippas, J. R., D. Spina, L. Cavedon, H. Joho, and M. Sanderson. 2018. "Informing the Design of Spoken Conversational Search: Perspective Paper". In: *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval. CHIIR '18*. New Brunswick, NJ, USA: Association for Computing Machinery. 32–41. ISBN: 9781450349253. DOI: 10.1145/3176349.3176387. URL: https://doi-org.ezp.lib.unimelb.edu.au/10.1145/3176349.3176387.

Trippas, J. R., D. Spina, L. Cavedon, and M. Sanderson. 2017. "How Do People Interact in Conversational Speech-Only Search Tasks: A Preliminary Analysis". In: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. CHIIR '17*. Oslo, Norway: ACM. 325–328. ISBN: 978-1-4503-4677-1. DOI: 10.1145/3020165.3022144. URL: http://doi.acm.org/10.1145/3020165.3022144.

Trippas, J. R., D. Spina, M. Sanderson, and L. Cavedon. 2015a. "Results Presentation Methods for a Spoken Conversational Search System". In: *Proceedings of the First International Workshop on Novel Web Search Interfaces and Systems. NWSearch '15*. Melbourne, Australia: ACM. 13–15. ISBN: 978-1-4503-3789-2. DOI: 10.1145/2810355.2810356. URL: http://doi.acm.org/10.1145/2810355.2810356.

Trippas, J. R., D. Spina, M. Sanderson, and L. Cavedon. 2015b. "Towards Understanding the Impact of Length in Web Search Result Summaries over a Speech-only Communication Channel". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '15*. Santiago, Chile: ACM. 991–994. ISBN: 978-1-4503-3621-5. DOI: 10.1145/2766462.2767826. URL: http://doi.acm.org/10.1145/2766462.2767826.

Trippas, J. R., D. Spina, M. Sanderson, and L. Cavedon. 2021. "Accessing Media Via an Audio-Only Communication Channel: A Log Analysis". In: *CUI 2021 - 3rd Conference on Conversational User Interfaces. CUI '21*. Bilbao (online), Spain: Association for Computing Machinery. ISBN: 9781450389983. DOI: 10.1145/3469595.3469623. URL: https://doi.org/10.1145/3469595.3469623.

Trippas, J. R., D. Spina, F. Scholer, A. H. Awadallah, P. Bailey, P. N. Bennett, R. W. White, J. Liono, Y. Ren, F. D. Salim, and M. Sanderson. 2019. "Learning About Work Tasks to Inform Intelligent Assistant Design". In: *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval. CHIIR '19*. Glasgow, Scotland UK: Association for Computing Machinery. 5–14. ISBN: 9781450360258. DOI: 10.1145/3295750.3298934.

Trippas, J. R., D. Spina, P. Thomas, M. Sanderson, H. Joho, and L. Cavedon. 2020. "Towards a Model for Spoken Conversational Search". *Information Processing & Management*. 57(2): 102162. ISSN: 0306-4573.

Trippas, J. R. and P. Thomas. 2019. "Data Sets for Spoken Conversational Search". In: *Proceedings of the CHIIR 2019 Workshop on Barriers to Interactive IR Resources Re-use. BIIRRR@CHIIR '19*. Glasgow, UK. 14–18.

Turpin, A., Y. Tsegay, D. Hawking, and H. E. Williams. 2007. "Fast generation of result snippets in web search". In: *Proceedings of the ACM SIGIR conference on Research and development in information retrieval*. 127–134.

Vakulenko, S., E. Kanoulas, and M. De Rijke. 2021. "A Large-Scale Analysis of Mixed Initiative in Information-Seeking Dialogues for Conversational Search". *ACM Trans. Inf. Syst.* 39(4). ISSN: 1046-8188. DOI: 10.1145/3466796. URL: https://doi.org/10.1145/3466796.

Vakulenko, S., S. Longpre, Z. Tu, and R. Anantha. 2020. "Question Rewriting for Conversational Question Answering". *ArXiv*. abs/2004.14652.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. *NeurIPS '17*. Curran Associates, Inc. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Voorhees, E. 2004. "Overview of the TREC 2004 Question Answering Track". In: *TREC*.

Voorhees, E. 2005. "Overview of the TREC 2005 Question Answering Track". In: *TREC*.

Voorhees, E. M. *et al.* 1999. "The TREC-8 question answering track report". In: *Trec*. Vol. 99. 77–82.

Voskarides, N., D. Li, P. Ren, E. Kanoulas, and M. Rijke. 2020. "Query Resolution for Conversational Search with Limited Supervision". *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Vtyurina, A., C. L. Clarke, E. Law, J. R. Trippas, and H. Bota. 2020. "A Mixed-Method Analysis of Text and Audio Search Interfaces with Varying Task Complexity". In: *Proc. ICTIR*.

Vuong, T., G. Jacucci, and T. Ruotsalo. 2018. "Naturalistic Digital Task Modeling for Personal Information Assistance via Continuous Screen Monitoring". In: *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. UbiComp '18*. Singapore, Singapore: Association for Computing Machinery. 778–785. ISBN: 9781450359665. DOI: 10.1145/3267305.3274130. URL: https://doi-org.ezp.lib.unimelb.edu.au/10.1145/3267305.3274130.

Wadhwa, S. and H. Zamani. 2021. "Towards System-Initiative Conversational Information Seeking". In: *Proceedings of the Second International Conference on Design of Experimental Search and Information Retrieval Systems. DESIRES '21*. Padua, Italy: CSUR. 102–116.

Walker, M. and S. Whittaker. 1990. "Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation". In: *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics. ACL '90*. Pittsburgh, Pennsylvania: Association for Computational Linguistics. 70–78. DOI: 10.3115/981823.981833.

Walker, M. A., D. J. Litman, C. A. Kamm, and A. Abella. 1997. "PARADISE: A Framework for Evaluating Spoken Dialogue Agents". In: *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain: Association for Computational Linguistics. 271–280. DOI: 10.3115/976909.979652. URL: https://aclanthology.org/P97-1035.

Wang, Z. and Q. Ai. 2021. "Controlling the Risk of Conversational Search via Reinforcement Learning". In: *Proceedings of the Web Conference 2021. WWW '21*. Ljubljana, Slovenia: Association for Computing Machinery. 1968–1977. ISBN: 9781450383127. DOI: 10.1145/3442381.3449893. URL: https://doi.org/10.1145/3442381.3449893.

Warren, D. H. D. and F. C. N. Pereira. 1982. "An Efficient Easily Adaptable System for Interpreting Natural Language Queries". *Comput. Linguist.* 8(3–4): 110–122. ISSN: 0891-2017.

Weeratunga, A. M., S. A. U. Jayawardana, P. M. A. K. Hasindu, W. P. M. Prashan, and S. Thelijjagoda. 2015. "Project Nethra - an intelligent assistant for the visually disabled to interact with internet services". In: *2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS)*. 55–59. DOI: 10.1109/ICIINFS.2015.7398985.

Weizenbaum, J. 1966. "ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine". *Commun. ACM*. 9(1): 36–45. ISSN: 0001-0782. DOI: 10.1145/365153.365168. URL: https://doi.org/10.1145/365153.365168.

Welleck, S., I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston. 2020. "Neural Text Generation With Unlikelihood Training". In: *International Conference on Learning Representations. ICLR '20*.

Weston, J., E. Dinan, and A. H. Miller. 2018. "Retrieve and Refine: Improved Sequence Generation Models For Dialogue". In: *SCAI@EMNLP*.

White, R. W. and R. A. Roth. 2009. *Exploratory Search: Beyond the Query-Response Paradigm. Synthesis Lectures on Information Concepts, Retrieval, and Services*. Morgan & Claypool Publishers.

Williams, J., A. Raux, and M. Henderson. 2016. "The Dialog State Tracking Challenge Series: A Review". *Dialogue Discourse*. 7: 4–33.

Williams, R. J. 1992. "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning". *Mach. Learn.* 8(3–4): 229–256. ISSN: 0885-6125. DOI: 10.1007/BF00992696. URL: https://doi.org/10.1007/BF00992696.

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

Winograd, T. 1972. "Understanding natural language". *Cognitive Psychology*. 3(1): 1–191. ISSN: 0010-0285. DOI: https://doi.org/10.1016/0010-0285(72)90002-3. URL: https://www.sciencedirect.com/science/article/pii/0010028572900023.

Winograd, T. 1974. "Five lectures on artificial intelligence". *Tech. rep.* Stanford University, computer science department.

Winters, R. M., N. Joshi, E. Cutrell, and M. R. Morris. 2019. "Strategies for Auditory Display of Social Media". *Ergonomics in Design*. 27(1): 11–15. DOI: 10.1177/1064804618788098. eprint: https://doi.org/10.1177/1064804618788098. URL: https://doi.org/10.1177/1064804618788098.

Woodruff, A., R. Rosenholtz, J. B. Morrison, A. Faulring, and P. Pirolli. 2002. "A Comparison of the Use of Text Summaries, Plain Thumbnails, and Enhanced Thumbnails for Web Search Tasks". *J. Am. Soc. Inf. Sci. Technol.* 53(2): 172–185. ISSN: 1532-2882. DOI: 10.1002/asi.10029. URL: https://doi.org/10.1002/asi.10029.

Woods, W. A., R. M. Kaplan, and B. Nash-Webber. 1972. *The Lunar Sciences Natural Language Information System Final Report*. NASA.

Wu, C.-S., A. Madotto, E. Hosseini-Asl, C. Xiong, R. Socher, and P. Fung. 2019. "Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics. 808–819. DOI: 10.18653/v1/P19-1078. URL: https://www.aclweb.org/anthology/P19-1078.

Xiong, L., C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. N. Bennett, J. Ahmed, and A. Overwijk. 2021. "Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval". In: *International Conference on Learning Representations. ICLR '21*.

Xiong, W., X. Li, S. Iyer, J. Du, P. Lewis, W. Y. Wang, Y. Mehdad, W.-t. Yih, S. Riedel, D. Kiela, and B. Ouguz. 2020. "Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval". *ArXiv*. abs/2009.12756.

Xu, J., A. D. Szlam, and J. Weston. 2021. "Beyond Goldfish Memory: Long-Term Open-Domain Conversation". *ArXiv*. abs/2107.07567.

Xu, W. and A. Rudnicky. 2000. "Language modeling for dialog system." In: *INTERSPEECH*. ISCA. 118–121.

Xu, X. and J. Lockwood. 2021. "What's going on in the chat flow? A move analysis of e-commerce customer service webchat exchange". *English for Specific Purposes*. 61: 84–96. ISSN: 0889-4906. DOI: https://doi.org/10.1016/j.esp.2020.09.002. URL: https://www.sciencedirect.com/science/article/pii/S0889490620300533.

Yang, J.-H., S.-C. Lin, C.-J. Wang, J. Lin, and M.-F. Tsai. 2019. "Query and Answer Expansion from Conversation History". In: *TREC*.

Yang, L., M. Qiu, C. Qu, C. Chen, J. Guo, Y. Zhang, W. Croft, and H. Chen. 2020. "IART: Intent-aware Response Ranking with Transformers in Information-seeking Conversation Systems". *Proceedings of The Web Conference 2020*.

Yang, L., M. Qiu, C. Qu, J. Guo, Y. Zhang, W. B. Croft, J. Huang, and H. Chen. 2018a. "Response Ranking with Deep Matching Networks and External Knowledge in Information-seeking Conversation Systems". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18*. Ann Arbor, MI, USA: ACM. 245–254. ISBN: 978-1-4503-5657-2. DOI: 10.1145/3209978.3210011. URL: http://doi.acm.org/10.1145/3209978.3210011.

Yang, L., H. Zamani, Y. Zhang, J. Guo, and W. B. Croft. 2017. "Neural Matching Models for Question Retrieval and Next Question Prediction in Conversation". *CoRR*. abs/1707.05409. URL: http://arxiv.org/abs/1707.05409.

Yang, Y., Y. Gong, and X. Chen. 2018b. "Query Tracking for E-Commerce Conversational Search: A Machine Comprehension Perspective". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM '18*. Torino, Italy: Association for Computing Machinery. 1755–1758. ISBN: 9781450360142. DOI: 10.1145/3269206.3269326. URL: https://doi-org.ezp.lib.unimelb.edu.au/10.1145/3269206.3269326.

Yeh, Y.-T. and Y.-N. Chen. 2019. "FlowDelta: Modeling Flow Information Gain in Reasoning for Conversational Machine Comprehension". *ArXiv*. abs/1908.05117.

Yenala, H., A. Jhanwar, and G. Chinnakotla M.K. an Goyal. 2018. "Deep learning for detecting inappropriate content in text". *Int J Data Sci Anal*. 6: 273–286. DOI: 10.1007/s41060-017-0088-4.

Yu, D. and Z. Yu. 2019. "MIDAS: A dialog act annotation scheme for open domain human machine spoken conversations". *arXiv preprint arXiv:1908.10023*.

Yu, S., Z. Liu, C. Xiong, T. Feng, and Z. Liu. 2021. "Few-Shot Conversational Dense Retrieval". *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yu, S., J. Liu, J. Yang, C. Xiong, P. Bennett, J. Gao, and Z. Liu. 2020. "Few-Shot Generative Conversational Query Rewriting". *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zamani, H., M. Bendersky, X. Wang, and M. Zhang. 2017. "Situational Context for Ranking in Personal Search". In: *Proceedings of the 26th International Conference on World Wide Web. WWW '17*. Perth, Australia: International World Wide Web Conferences Steering Committee. 1531–1540. ISBN: 9781450349130. DOI: 10.1145/3038912.3052648. URL: https://doi.org/10.1145/3038912.3052648.

Zamani, H. and N. Craswell. 2020. "Macaw: An Extensible Conversational Information Seeking Platform". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20*. Virtual Event, China: Association for Computing Machinery. 2193–2196. ISBN: 9781450380164. DOI: 10.1145/3397271.3401415. URL: https://doi.org/10.1145/3397271.3401415.

Zamani, H. and W. B. Croft. 2020a. "Joint Modeling and Optimization of Search and Recommendation". In: *Proceedings of the First International Conference on Design of Experimental Search and Information Retrieval Systems. DESIRES '18*. Bertinoro, Italy: CSUR. 36–41.

Zamani, H. and W. B. Croft. 2020b. "Learning a Joint Search and Recommendation Model from User-Item Interactions". In: *Proceedings of the 13th International Conference on Web Search and Data Mining. WSDM '20*. Houston, TX, USA: Association for Computing Machinery. 717–725. ISBN: 9781450368223. DOI: 10.1145/3336191.3371818. URL: https://doi.org/10.1145/3336191.3371818.

Zamani, H., S. T. Dumais, N. Craswell, P. N. Bennett, and G. Lueck. 2020a. "Generating Clarifying Questions for Information Retrieval". In: *Proceedings of the 29th International Conference on World Wide Web. WWW '20*. Taipei, Taiwan.

Zamani, H., G. Lueck, E. Chen, R. Quispe, F. Luu, and N. Craswell. 2020b. "MIMICS: A Large-Scale Data Collection for Search Clarification". *arXiv preprint arXiv:2006.10174*.

Zamani, H., B. Mitra, E. Chen, G. Lueck, F. Diaz, P. N. Bennet, N. Craswell, and S. T. Dumais. 2020c. "Analyzing and Learning from User Interactions for Search Clarification". In: *The 43rd International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '20*. Virtual Event, China.

Zamora, J. 2017. "I'm Sorry, Dave, I'm Afraid I Can't Do That: Chatbot Perception and Expectations". In: *Proceedings of the 5th International Conference on Human Agent Interaction. HAI '17*. Bielefeld, Germany: Association for Computing Machinery. 253–260. ISBN: 9781450351133. DOI: 10.1145/3125739.3125766. URL: https://doi.org/10.1145/3125739.3125766.

Zhai, C. 2016. "Towards a game-theoretic framework for text data retrieval". *IEEE Data Eng. Bull.* 39(3): 51–62. URL: http://sites.computer.org/debull/A16sept/p51.pdf.

Zhai, C. 2020. "Interactive Information Retrieval: Models, Algorithms, and Evaluation". In: *SIGIR '20*. Virtual Event, China: Association for Computing Machinery. 2444–2447. ISBN: 9781450380164. DOI: 10.1145/3397271.3401424. URL: https://doi.org/10.1145/3397271.3401424.

Zhang, S. and K. Balog. 2020. "Evaluating Conversational Recommender Systems via User Simulation". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery. 1512–1520. ISBN: 9781450379984. URL: https://doi.org/10.1145/3394486.3403202.

Zhang, S., Z. Dai, K. Balog, and J. Callan. 2020a. "Summarizing and Exploring Tabular Data in Conversational Search". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20*. Virtual Event, China: Association for Computing Machinery. 1537–1540. ISBN: 9781450380164. DOI: 10.1145/3397271.3401205. URL: https://doi.org/10.1145/3397271.3401205.

Zhang, S., H. Yang, and L. Singh. 2016. "Anonymizing Query Logs by Differential Privacy". In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 753–756.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020b. "BERTScore: Evaluating Text Generation with BERT". In: *Proceedings of the 2020 International Conference on Learning Representations. ICLR '20*.

Zhang, X., H. Xie, H. Li, and J. C.S. Lui. 2020c. "Conversational Contextual Bandit: Algorithm and Application". In: *Proceedings of The Web Conference 2020. WWW '20*. Taipei, Taiwan: Association for Computing Machinery. 662–672. ISBN: 9781450370233. DOI: 10.1145/3366423.3380148. URL: https://doi.org/10.1145/3366423.3380148.

Zhang, Y. and C. Zhai. 2015. "Information Retrieval as Card Playing: A Formal Model for Optimizing Interactive Retrieval Interface". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '15*. Santiago, Chile: Association for Computing Machinery. 685–694. ISBN: 9781450336215. DOI: 10.1145/2766462.2767761. URL: https://doi.org/10.1145/2766462.2767761.

Zhang, Y., X. Chen, Q. Ai, L. Yang, and W. B. Croft. 2018. "Towards Conversational Search and Recommendation: System Ask, User Respond". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM '18*. Torino, Italy: ACM. 177–186. ISBN: 978-1-4503-6014-2. DOI: 10.1145/3269206.3271776. URL: http://doi.acm.org/10.1145/3269206.3271776.

Zhao, T., K. Xie, and M. Eskenazi. 2019. "Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics. 1208–1218. DOI: 10.18653/v1/N19-1123. URL: https://www.aclweb.org/anthology/N19-1123.

Zhou, L., J. Gao, D. Li, and H.-Y. Shum. 2020. "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot". *Computational Linguistics*. 46(1): 53–93. ISSN: 0891-2017. DOI: 10.1162/coli_a_00368. URL: https://doi.org/10.1162/coli%5C_a%5C_00368.

Zhou, X., L. Li, D. Dong, Y. Liu, Y. Chen, W. X. Zhao, D. Yu, and H. Wu. 2018. "Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network". In: *ACL*.

Zue, V. W. and J. R. Glass. 2000. "Conversational interfaces: advances and challenges". *Proceedings of the IEEE*. 88(8): 1166–1180. DOI: 10.1109/5.880078.