A Survey on Conversational Search and Applications in Biomedicine

Naga Sai Krishna Adatrao  
Kennesaw State University  
Kennesaw, Georgia, USA  
saikrishna.adatrao@gmail.com

Gowtham Reddy Gadireddy  
Kennesaw State University  
Kennesaw, Georgia, USA  
ggadired@students.kennesaw.edu

Jiho Noh  
Kennesaw State University  
Kennesaw, Georgia, USA  
jnoh3@kennesaw.edu

ABSTRACT
This paper aims to provide a radical rundown on Conversational Search (ConvSearch), an approach to enhance the information retrieval (IR) method where users engage in a dialogue for the information-seeking tasks. In this survey, we predominantly focused on the human interactive characteristics of the ConvSearch systems, highlighting the operations of the action modules, likely the retrieval system, question-answering, and recommender system. We labeled various ConvSearch research problems in knowledge bases, natural language processing, and dialogue management systems with action modules. We further categorized the framework to ConvSearch, and the application is directed toward biomedical and healthcare fields for the utilization of clinical social technology. Finally, we conclude by talking through the challenges and issues of ConvSearch, particularly in Bio-Medicine. Our main aim is to provide an integrated and unified vision of the ConvSearch components from different fields, which benefit the information-seeking process in healthcare systems.

CCS CONCEPTS
• Information systems → Information retrieval; Users and interactive retrieval; Information retrieval query processing; Retrieval tasks and goals.

KEYWORDS
Information Retrieval, Conversational Search, Question Answering, Knowledge Base, Dialogue Management Systems, Recommender Systems, Generative Language Models, Biomedical ConvSearch, Privacy Concerns

ACM Reference Format:
Naga Sai Krishna Adatrao, Gowtham Reddy Gadireddy, and Jiho Noh. 2023. A Survey on Conversational Search and Applications in Biomedicine. In 2023 ACM Southeast Conference (ACMSE 2023), April 12–14, 2023, Virtual Event, USA. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3564746.3587001

1 INTRODUCTION
1.1 Information-Seeking Process
The information-seeking process via human-computer interactions has become more “interactive” than ever before. For instance, modern web search engines provide features like query suggestions that attempt to specify the user’s original search intent, featured snippets provide relevant text snippets that may directly answer the user’s query, and knowledge panels that present pertinent information to a query in a structured format. According to a market research company [69], 65% of middle-aged adults speak to their voice-enabled devices at least once a day, which requires conversational features. Designing the features of Conversational Search (ConvSearch) has recently become a more critical factor in developing information retrieval (IR) systems and advances in natural language understanding.

The lexical definition of conversation is the act of interchangeing thoughts, information, etc., by spoken words between multiple parties. We apply the exact definition to the ConvSearch systems and focus on the properties of conversation that include (1) the use of natural language, (2) multi-turn interactions, and (3) the aspect of the information-seeking process.

1.2 History of Conversational Search
The idea of designing a search engine that allows users or the system to interact with the other party through dialog to seek users’ information needs is not a new thing. Croft and Thompson [19] stated that “the IR system should enter into the dialogue with a user to verify inferred concepts and request more information on domains that are not well specified.” From the late 1990s, researchers studied concrete methodologies to implement dialogue components for IR [7] and recommendation systems [31]. Recently, Croft [18] stresses on the importance of user-system interaction in the information-seeking scenario, particularly in the limited-bandwidth environments provided by mobile devices and voice-based assistants.

The volume of scientific literature on ConvSearch has grown exponentially in the last two years (2021-22). Research topics expanded towards multi-modality search [62] and conversational agents for info-seeking in pharmacologic knowledge bases [68], and so on.

2 BACKGROUND
In the frame of the information-seeking process, IR is closely related to recommender systems and question-answering systems. This section discusses the shared objectives and differences between these related research topics.
A Survey on Conversational Search and Applications in Biomedicine
K. Adatrao, G. Gadireddy, J. Noh

2.1 Information Retrieval
Conventional IR systems evaluate ad-hoc queries with documents based on a lexical exact-match model such as BM25 [76] and query likelihood models. Typically, a user specifies the information need with a single query that initiates a search. Then the user receives the system-computed list of relevant items and determines whether to accept the results or refine the initial query for a follow-up search. That sequence completes one session of the information retrieval process. To progress qualitative stand of IR in research, it is important to use a range of techniques and tools, which includes advanced search algorithms, metadata management, and NLP techniques. Moreover, researchers must be skilled in utilizing these tools and understand the core principles of IR to extract and analyze data effectively.

ConvSearch differs from conventional IR systems in their interaction mode; ConvSearch allows users to describe their information needs in a less strict natural language format. During the conversation, the system can request for additional information from the user when the system’s understanding of the initial query is uncertain. The effort to communicate back and forth to reduce the uncertainty characterizes the conversational search systems.

2.2 IR vs. Recommender Systems
Recommender systems (RecSys) qualify available items (e.g., products, documents) and make suggestions that may suit the user’s expectations or information needs. A recommender system computationally models users’ expectations, preferences, item features, and interactions between users and items. Unlike IR, RecSys does not require a user query; suggestions can be made based on item features, user profiles, or interactions.

Recently, conversational approaches to recommender systems (ConvRecSys) allow personalized recommendations through natural language dialog with users, which has become a trending research topic. Notably, the research focuses include (1) refining a user’s preference by clarifying preferred item features via conversation [99, 109, 111] and (2) natural language understanding and generation in a conversational RecSys context [52, 87].

2.3 IR vs. Question-Answering
Within the fields of IR and NLP, question-answering (QA) is the task of computationally answering questions posed by humans in natural language. Like ConvRecSys, Conversational Question-Answering (ConvQA) emphasizes the multi-turn dialogue interactions between users and the system.

Due to the shared objectives between QA and IR, many QA techniques and benchmarks are adopted in IR research; both systems are expected to respond to the user’s information needs. Unlike IR, QA seeks to find relevant passages from a document collection or infer over a knowledge graph to answer a question. On the other hand, IR considers a set of relevant documents an answer to the user’s query.

ConvQA takes a more vital role in ConvSearch because we aim to provide integrated and dynamic system responses in ConvSearch. In that scenario, a direct answer to a user’s query (e.g., a factoid question) is a desirable system response.

2.4 Scope of Survey
2.4.1 Interactive Information Retrieval. ConvSearch is interactive information retrieval (IR) because of the mixed-initiative communication between the user and the system [73]. From the Human-Computer Interaction (HCI) perspective, the medium of IR communication can be non-verbal (e.g., gestures as sensory or images/videos as visual) or verbal (e.g., spoken as audio or written as textual). We consider speech a verbal communication that requires recognition and translation into text [17, 59] for text-based information retrieval systems. This survey will mainly focus on verbal communication, as Anand et al. [3] described a conversational search system as an IR system with speech and language processing capabilities.

2.4.2 Task-Oriented Conversation. ConvSearch is a task-oriented dialogue system where the task is specific to IR. So, ConvSearch has the inherent problems of dialogue systems, such as dealing with phenomena like anaphora (use of a word whose interpretation depends upon another expression in context) and ellipsis (omission or suppression of words that appeared earlier) [93] or modeling and controlling dialogue structure [15].

This survey highlights the research efforts on dialogue systems with IR capabilities. However, we do not exclude the approaches toward questions-answering and recommendation tasks.

2.4.3 Inclusion and Exclusion. This paper surveys recent advances in ConvSearch technologies and their applications in the biomedical domain. Section 3 portrays ConvSearch as an elaboration process in the information-seeking process, which attempts to maximize the certainty to users’ information needs. In Section 4, we dissect conversational search systems and present representative research efforts on each component of the conceptual architecture, such as conversational QA, recommender systems, and knowledge bases. In Section 5, we demonstrate information-seeking applications in the biomedical domain. In addition, we discuss the aims of biomedical conversational search systems, likely semantic representations, trustworthiness, and personalization issues.

We identified relevant works by first querying several scholarly document search systems (i.e., Google Scholar, Semantic Scholar, arXiv, Citeseer, and ResearchGate) with commonly used keywords for conversational search and related research areas, such as “conversational search,” “ConvSearch,” “conversational information retrieval,” and “CIR search.” For biomedical applications, we used keywords such as “biomedical information retrieval” and “biomedical document search.” Further, we expanded the collection by manually selecting representative and relevant papers cited from the documents we have already collected. We have collected 346 citations, and approximately half of the documents are considered in this study for review.

3 INFORMATION SEEKING PROCESS
This section contextualizes the role of conversation in the theme of the information-seeking process. Table 1 presents the staged process of information-seeking activity, in particular to the information retrieval systems, where conversation acts as a means of clarifying the user’s information need and adding details to the query. We adopt a model of information-seeking behavior proposed
by Kuhlthau [46]. Her model comprises six stages: initiation, selection, exploration, formulation, collection, and presentation. From the information retrieval perspective, we modified it to a simple four-staged process: initiation, mutual understanding, elaboration, and presentation. The following subsections explain each stage and demonstrate examples.

3.1 Initiation
In the initiation stage, a user begins a search process by recognizing the means of communication for their information-seeking activities. The user interface plays a vital role in this stage and the following presentation stage as well. First, interface recognition aids in the users’ expression of their information needs. Users formulate a query, select appropriate information sources, filter results by specific options using facets, and later review the search results in the presentation stage.

A standard interface for search engines is an ad-hoc search box that allows a single query string from a user. Typically, users write a short sequence of keywords that describes their information needs. The system may provide query suggestions in this search box, but we consider it as an elaboration action in the following stage.

The emergence of conversational systems demands a different mode of communication, such as voice user interfaces (VUIs) or chatbots [32, 55]. As we defined the two characteristics of ConvSearch, this conversational interface has two attributes: the use of natural language and multi-turn interactions.

3.2 Mutual Understanding
Provided with the interface features, a user and the system first attempt to explore the topic of interest in the search space. In this stage, a user specifies the information needs, and the system reveals what is available in the data collection. We assume that users may not be familiar with the searching task and even unclear about the target information due to the misconception of the topics or the data collection itself. Depending on the search system, users endeavor to find the best query by which the system can retrieve relevant data and satisfy the user. The system analyzes the initial query utilizing various NLP and IR techniques such as query prediction [78] and query expansion [64]. Then, the system can ask for clarification or return the initial retrieval results upon the analysis.

3.3 Elaboration
Elaboration is an iterative process, often illustrated as a loop between a user and the system in IR process diagrams. The outcome of the interaction is a refined query. We narrow down the search focus through the interactions between a user and the system in this stage. We add additional information to a user query (e.g., query expansion) or reformulate a query given by the user’s feedback (e.g., explicit/implicit relevance feedback). We also consider interactions using visual aids as an elaboration action. We aim to leverage conversations for the elaboration task toward conversational search systems.

3.4 Presentation
The presentation stage is highly linked with communication and the initiation stage. The scope of user interfaces for search has been expanded recently, including textual and audio/video communications, which requires the system to accommodate different interfaces, such as multi-modal (e.g., a conversation using visual information), multi-dimensional (e.g., knowledge panel), and interactive (e.g., visual analytics), including conversational agents (e.g., chatbots).

The content of search responses should also contain answers to questions and recommendations. As the search systems become more flexible with their communication strategies, systems response should include the retrieval results and other types of interactions. This demand creates another set of challenges in search system design.

4 CONVERSATIONAL SEARCH
In many real-world information-seeking scenarios, seekers lack detailed knowledge of the underlying information sources. Consequently, a query formulation effort (“an intelligent guess”) primarily relies on the user’s prior knowledge. However, ConvSearch changes this scenario according to the initiatives and user intents to make information more accessible and satisfying in retrieval practice [106].

Suppose the goal of ConvSearch aims at satisfying a user’s information needs in a sequence (S) of multi-turn dialogue between the system and a user.

\[ S = Q_0, (A_1 R_1, A_2 R_2, \ldots, A_{k-1} R_{k-1}), A_k \]

A conversation is initiated by a user’s first attempt to express his/her information needs (Q0); this can be a keyword-based query or a complete interrogative sentence. The conversation continues with
multiple pairs of a system-oriented action \((A_i)\) and the user’s response \((R_i)\). The goal of ConvSearch is to satisfy the user with \((A_k|A_1, \ldots, A_{k-1})\) and ideally terminate the conversation when the user is satisfied with \(A_k\).

We do not strictly define or categorize system actions \((A_i)\). Still, we include at least the following action types: \(A_i \in \{D_R, D_S, q, r\}\), where \(D_R\) is a set of relevant data items, \(D_S\) is a recommendation with suggested items, \(q\) is a question for search-oriented clarification, and \(r\) is pertinent information to any of the previous user-initiated questions (e.g., a direct answer or a textual snippet that can contain an answer to \(R_{c,i}\)). The user can provide further information \((r)\) to a question or take the initiative to ask a follow-up question \((q)\): \(R_i \in \{q, r\}\).

Figure 1 illustrates the architecture of a ConvSearch system, which consists of three components: (1) a dialogue management system (DMS), (2) online action modules, and (3) offline knowledge-base modules. DMS coordinates the conversation between the system and the user. The two most essential features of DMS are query understanding and response generation. For example, DMS might involve in a process for specifying the user’s intent by generating a clarifying question. The query understanding module analyzes the initial retrieval results and the user query to predict if it is necessary to ask a clarifying question. The response generation module determines how to form a question and executes it when needed.

Online action modules operate on the given system action by DMS, which consists of three modules: information retrieval system, question-answering, and recommender system. ConvSearch systems allow queries in a more natural format, such as a complete question sentence, which makes the question-answering feature a crucial module in developing ConvSearch systems. Furthermore, the ConvSearch system should also work as a recommender system to let users explore available data items in the database and possible search options. We conceive an architecture of ConvSearch systems which comprises two important system action modules and the information retrieval system.

### 4.1 Dialogue Management System

Dialogue Management Systems (DMS) engage in conversational interactions between a user and the system, primarily concerned with understanding a user’s intent and generating an appropriate system action. Typically, DMS consists of two modules: dialogue state tracking and dialogue policy.

With the IR-oriented DMS, the retrieval task is tracked in the form of the belief state to represent the user’s information needs throughout the conversation. Methods utilizing structured states, such as frame-based slot filling approaches [30], are widely adopted in task-oriented dialogue systems. An inevitable limitation of this approach is the pre-defined slot-value pair structure, which is considered unsuitable for open-domain conversation search systems [16].

**Dialogue State Tracking** (DST) is the process of identifying entities across a multi-turn conversation. In ConvSearch, this process consequently poses a conversational query reformulation (CQR) problem such as coreference resolution and query expansion [75]. Researchers attempted to resolve the ambiguity issues in a local query by adding keywords from its session (including answers to any queries) or prior knowledge [71, 103]. Recently, leveraging neural networks for predicting slot values has been a general approach to the DST problems [5, 33, 60].

**Dialogue policy** (DP) is a function that decides how the system responds to the user’s previous utterances. Rather strict belief state representation is not suitable for the ConvSearch systems. Hence, DP is expected to be a more versatile and flexible learning system with a high priority because the ConvSearch systems integrate the components in IR/QA/RecSys/KB/NLP.

In general, DP should select the most effective action to minimize uncertainty, which is the ultimate goal of the ConvSearch systems. More specifically, Penha et al. [67] listed fifteen sub-goals (e.g., query disambiguation from IR, intent/domain prediction from DMS, MRC from NLP, etc.) and categorized them into two groups: information-need elucidation and information presentation, which correspond to our last two stages in the information-seeking pipeline (i.e., elaboration and presentation). The action set has no standardized scheme. For example, depending on the search goals, we can suggest actions for each stage as follows: [clarify, elicit, interpret] for elaboration and [provide results, recommend, summarize, explain] for presentation.

**Query Performance Prediction** (QPP) is one of the IR evaluation methods actively studied over the past two decades. QPP estimates the effectiveness of the retrieval results without using human relevance judgments. QPP has become more critical in the context of ConvSearch because it can be a significant factor in dialogue policy learning [77]. Pal and Ganguly [66] proposed a method that involves shifting a window of terms through a conversation and score the segments using QPP. Recent studies focus on using neural approaches such as using a pre-trained language model for QPP [4].

#### 4.1.1 Conversational Query Understanding (CQU)

A query understanding module is expected to transform textual data (i.e., query) into semantic representations by inferring the user’s intent using various NLP techniques, including deep learning methods. When it becomes Conversational Query Understanding (CQU), we can formulate this problem as context-aware query reformulation. Essentially, the goal is to construct a conversational query that precisely encodes the user’s information needs. Researchers have made significant progress on related tasks, including user intent prediction and query disambiguation.

The challenges in CQU originated in the unclarity of user utterances due to anaphora and ellipsis. Hence, the query reformulation process, such as coreference resolution and query expansion, is a necessary feature regardless of the type and modality of the search. We consider these problems critical sub-tasks of dialogue management systems. Recently proposed methods commonly leverage neural network models to accomplish these tasks. Yu et al. [105] proposed a visual-aware pronoun coreference resolution method between a dialogue and an image. Lin et al. [50] enhanced a neural query rewriting module using term importance estimation for conversational passage retrieval tasks.

#### 4.1.2 System Response Generation

Response Generator elicits outputs by converting structured data into natural language utterances, representing NLG task engines. The output should contain various linguistic components, such as function words and punctuations. Gatt and Krahmer [26] categorized the linguistic realization into
three different approaches: (1) human-crafted templates (2) human-crafted grammar-based systems, and (3) corpus-based statistical approaches.

Mikolov et al. [58] introduced the recurrent neural network approach for training a language model. A similar approach was advanced by Wen et al. [97, 98] with an LSTM-based structure to NLG. Subsequently, according to the dialogue policy, the challenges in NLG (e.g., information omitting and duplicating problems in surface realization) are addressed by gating the input vector of LSTM-based structure [89, 98]. Recently, Dudy et al. [22] highlighted the relevance as in IR to be a crucial measure for designing text-generating tasks.

### 4.2 Knowledge Base

Designing and constructing a knowledge base (KB) is crucial in developing a ConvSearch system. KB represents a knowledge model of the search system, which may integrate different elements such as words, concepts, entities, and their predicates. Typically, KBs are modeled employing taxonomy or ontology, structured in relational databases or knowledge graphs (KG). Recent research studies explore the efficacy of using KBs embedded in a semantic vector space using neural networks. The following highlights some of the use cases of various knowledge models.

The most dominant use case of KB in the information-seeking process is question-answering. KBQA (QA over KB) is to elicit answers to given questions using the semi-structured relationships between entities, and this field has been studied actively [24]. In essence, KB in IR exploits the semantic regularities among data elements (e.g., concepts, documents), and researchers have recently leveraged deep learning to learn the regularities. For example, Lukovnikov et al. [53] trained an NN model to rank subject-predicate pairs to retrieve relevant facts for a given question. Shen et al. [81] tackled the coreference resolution problem in semantic parsing, utilizing a large-scale knowledge base such as DBpedia [47] and Wikidata [92].

KB involves in every aspect of ConvSearch. Agarwal et al. [1] incorporate a multi-modal knowledge base in generating system responses related to product-specific attributes, whereby a response can be a mixture of textual and visual information. Gaur et al. [27] utilize a knowledge graph to generate information-seeking questions. Clearly, KB can improve the NLP [95] and recommendation [110] modules in conversational search systems.

### 4.3 Question-Answering

The properties of QA and IR overlap with information-seeking perspectives. A subtle difference between them is that IR considers a set of relevant documents to answer the user’s question. In contrast, QA provides more specific information to a question, such as factoids, a list of entities, definitions, relationships, etc. Recently, QA captured more attention due to the increased interest in conversational aspects of the information-seeking process. There are two types of approaches in QA: Machine Reading Comprehension (MRC) [12, 79] and Knowledge Graph-based Question-Answering (KGQA) [6]. ConvQA follows the same approaches. In this paper, we highlight two aspects of QA in IR. Firstly, the research efforts to address the challenges in ConvQA, and secondly, Biomedical QA (BQA) and its applications.

In ConvQA, the user’s questions are not in a full-fledged interrogative sentence. Inherently, search systems are demanded to resolve more complex linguistic problems of conversations [25]. Query rewriting is an approach to replenish the partial view of a user’s information need by identifying repeated or omitted mentions and replacing them with more specific words [90]. Christmann et al. [14] proposed a method to better understand users’ incomplete questions in terms of fact-centric entities and predicates over a knowledge graph, automatically inferring missing or ambiguous pieces in a conversation. Gao et al. [25] proposed a generative model that maps incomplete utterances to logical forms, which can be executed on a knowledge base in searching for answers. Qu et al. [72] focused more on the positional information of conversation history.

---

**Figure 1: The Architecture of Conversational Search Systems**

![Diagram showing the architecture of conversational search systems: Dialogue Management (DM) including Dialogue state tracker and Dialogue policy, Conversation, Query Understanding, Online Action Modules with Document Retrieval, Question Answering, Recommender System, Off-line Knowledge-base Control with Reading Comprehension, User Modeling, KB Construction, and Summarization. Queries/feedback flow to System responses, connecting databases, user profiles, system logs, and knowledge bases connected to online and offline action modules and control systems.](image-url)
They proposed an MRC model with the history attention mechanism to demonstrate the importance of positional information of utterances in conversation modeling.

Our underlying hypothesis is that we can improve information retrieval quality by judicious use of QA systems to find items containing clues to answer user questions and queries. It is imperative to automatically construct and maintain domain-specific knowledge bases along with a corpus and exploit them in ConvSearch.

The QA mechanism with PLMs (Pre-trained Language Model) typically involves fine-tuning a pre-trained language model [49] on a specific domain or task, such as answering questions related to a particular topic or industry. The fine-tuning process involves training the PLM on a dataset of questions and corresponding answers [23], which enables the model to learn how to generate accurate and relevant responses to similar questions. During the conversational search process, the PLM receives the user’s question as input, and generates a response based on its training on the relevant dataset. The response can be further refined using additional information, such as the user’s context, search history, or preferences, to generate a more personalized and relevant answer.

4.4 Recommender Systems

Recommender Systems (RecSys) and IR have similar goals; both return a list of relevant items. However, the goals are slightly different in that IR attempts to find the best item for the given query. In contrast, recommender systems allow exploration of the search space to find the available items from different perspectives. Like other information-seeking fields, RecSys started employing conversational approaches (ConvRecSys) that enable interactions to refine the understanding of users’ preferences [35].

The most crucial task in RecSys is user modeling regardless of the methods: collaborative filtering or content-based filtering. Learning user preferences or degree of expertise should be a desired attribute of the conversational search systems. With user modeling, a system can provide tailored search results and recommendations. Indeed, this personalization can raise privacy and opacity concerns [94] since users do not have direct access to the information collected by the system. Nevertheless, the techniques for ConvRecSys are gaining more attention due to the shared tasks with IR and the benefits of providing recommendations in the information-seeking process.

Recommendations can improve search experiences in various aspects: (1) The system can select high-quality items based on other measures beyond their content-based features. (2) Recommendations also provide intuitive explanations that help humans understand why certain items are recommended [107]. (3) Ultimately, recommendations provide multi-dimensional views of the targeted information. For example, scholarly document search engines (e.g., Semantic Scholar) recommend highly related citations to a research topic of interest. We can recommend items such as citation [36], publication venue [56], authors to follow, and even queries. A query suggestion technique like “People Also Ask” recommends relevant alternative queries that lead to a different set of search results.

4.5 Generative Language Model

The Transformers architectures, which rely on the self-attention mechanism [91] capturing the context of terms or characters in a sequence, have been exploited in predictive models for downstream language tasks. Numerous large language models (LLM) are continuously proposed, which are commonly using the transformer architecture with typically more than billions of parameters. BERT (Bidirectional Encoder Representations from Transformers) [39], trained on next-sentence prediction and masked-language modeling, has become a ubiquitous baseline model for NLP tasks. GPT3.5 (Generative Pre-trained Transformer) and ChatGPT [8] have achieved state-of-the-art performance on a range of NLP tasks, including text completion, question answering, and text generation (trained specifically for generating human-like responses to chat-based conversations; hence, the models are called generative language models, GLM). The success of ChatGPT is attributed to their training process characterized by reinforcement learning from human feedback (RLHF) [13, 84], prompt tuning [48], and chain-of-thoughts [96].

LLMs like GPT-3 can actively be engaged in the information-seeking process, and they are showing great potential — but also raising concerns. Microsoft, a prominent investigator in the ChatGPT project, has already integrated a ChatGPT-style feature into Bing; this conversational interface generates summarized search results in colloquial language expected to answer what users are looking for. The use of LLMs in search engines is promising because of the capability of understanding more complex queries and generating answers to the needs, although it has some limitations: (1) LLMs can generate biased or offensive text due to training data bias. (2) LLMs may not have domain-specific expertise. (3) LLMs cannot verify the accuracy or reliability of information sources, leading to faithfulness and factuality issues [54]. It is important to note that careful language model training and additional fact-check techniques are necessary to address or mitigate the abovementioned limitations. For example, the technologies for linking valid and trustworthy URLs to relevant mentions in the LM-generated passage have become an important future research topic by which we can ensure that LLMs do not present any misleading or false information to users.

4.6 Evaluation

Although the evaluation component is not shown in Figure 1, the study of evaluation methodology in ConvSearch is vital and still needs to be explored to a full extent. ConvSearch differs from traditional IR systems in that it involves multi-turn interactions (e.g., sessions) using more likely well-formed natural language questions. Traditional evaluation methods (based on the Cranfield paradigm) focus on computing scores to a single query that essentially indicate how many relevant items are found and are positioned higher in a ranked list of results (e.g., precision, recall, and their variants).

There exists a line of work investigating the evaluation methods for interactive IR. Earlier, Carterette et al. [11] proposed dynamic test collections to systematically evaluate interactivity using click information and time spent reading documents. Jiang et al. [37] developed evaluation methods for voice-activated intelligent assistants in terms of user utility in speech recognition and intent
Zhang et al. [108] generalized IR evaluation metrics as search simulation using the notions of interaction reward and cost; for example, precision can be defined as the ratio of interaction reward and cost. Lipani et al. [51] developed an offline evaluation framework of ConvSearch based on the idea of “subtopics” that models novelty and diversity in search and recommendation. Specifically, they defined an evaluation metric called Expected Conversation Satisfaction (ECS) which estimates conversation satisfaction over many simulated dialogues. Various attempts have been made to assess ConvSearch systems in multiple dimensions, including user intent prediction [70], next question prediction [104], user satisfaction prediction [86], sub-goals prediction as appeared in the TREC Conversational Assistant Track [21].

5.3 Biomedical Question-Answering and Conversational Agents

Regarding biomedical information-seeking problems, the most actively studied fields are question-answering (BQA) and conversational agents (healthcare chatbots). One of the early BQA systems is AskHERMES [10], which allows physicians to write a question in natural language and navigate possible answer sentences to meet their information needs. BQA is an essential task for clinical decision support and personal health information seeking, which has been extensively studied. Interested readers can refer to Jin et al. [38]’s survey, highlighting recent developments and approaches to BQA.

Another popular adaptation is using conversational agents (i.e., chatbots) to support sociopsychological health interventions [34]. Kretzschmar et al. [45] discussed the strength and limitations of using chatbots in mental health support. Kowatsch et al. [44] demonstrated the potential of chatbots as a social mediator between healthcare professionals, patients, and family members. Physicians perceive conversational agents as a practical and innovative tool when seeking information in general; However, physicians emphasize the significance of obtaining specific and trustworthy answers [43]. Developing systems with the properties of interpretability and privacy-preserving mechanisms is a crucial topic in biomedical ConvSearch [2, 9].

5.4 Personalization

Information seekers with different backgrounds and degrees of domain knowledge may express different information needs with the same query. Search engines should comprehend the underlying search intents by understanding the user’s preference in context.
Personalization is a process of dynamically changing the information access and content, interface, and search functionality to tailor search outcomes for a user’s information need. In a broader scope, personalization in healthcare includes personalized medicine [83] and services like education and therapy [29].

There exist many concerns regarding personalization in the information-seeking scenario. As pointed out by Kocaballi et al. [42], most ConvSearch systems with a personalization feature are implemented without theoretical or evidence-based support, which can cause patient safety, privacy, and fairness issues. Personalization in ConvSearch is also prone to amplifying bias [28], such as filtering results by information sources and reducing the diversity, which is a potential violation of the usability principles [63].

5.5 Trustworthiness and Privacy Concerns

Recently, more attention has been directed to trustworthiness and privacy issues in deep learning technologies, such as the knowledge transfer mechanism behind the pre-trained language models (PLM). For example, word embeddings obtained from a publicly available PLM (e.g., BERT) can easily produce unwanted outcomes due to the unclear definition of similarity in the feature vector space. As Noh and Kavuluru [65] pointed out, PLMs may not be able to distinguish words in the same context, which can be a critical issue in natural language understanding/generation in biomedical applications. For example, words in antonym relationships (e.g., high/low blood pressure, opioid and nonopioid) or disease-symptom relationships (e.g., pharyngitis and sore throats) may have similar semantic representations.

To address this problem, a rapidly growing body of work focuses on analyzing the knowledge captured by PLM. A typical approach is probing via prompt engineering, which utilizes standardized biomedical vocabularies and ontologies. Commonly used benchmarks for probing biomedical knowledge include BioLAMA [88] and MedLAMA [57]. Another focus is the privacy leakage issue with PLM. Nakamura et al. [61] examine the privacy risk of language models pretrained with a document collection containing sensitive personal information. Using deep learning techniques is inevitable in ConvSearch systems, requiring more research to examine the potential pitfalls of transferring knowledge from other information sources.

6 CONCLUSION

Our survey summarizes current research on conversational search (ConvSearch). We have built our work on major components in ConvSearch and focused on the operation of dialogue and conversational interactions. Further, we have collaborated our work in the biomedical and healthcare domains to establish an ideology of social technology. The application of ConvSearch in the bio-field would bring a drastic change in the social world.

The emerging field of ConvSearch is adopting ideas from already established areas such as QA, RecSys, DS, and NLP, and the degree of integration is coarse at this stage (c.f., Figure 2). For an instance, current ConvSearch research is heavily skewed towards QA; it should be supported by diverse set of language understanding techniques such as reasoning over knowledge derived from an information source utilizing knowledge bases. Consequently, policy learning acting as a ‘control tower’ takes a vital role in the future development of ConvSearch systems. More researchers should focus on improving the integrated system of various methodologies from different fields.
[87] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (2018). https://doi.org/10.48550/arXiv.1806.03277

[88] Mujirem Sung, Jinhyuk Lee, Sean S. Yi, Ming Jieon, Sungdong Kim, and Jaewoo Kang. 2021. Can Language Models be Biomedical Knowledge Bases? ArXiv abs/2109.07154 (2021).

[89] Van-Kha Nh Tran and Le-Minh Nguyen. [n. d.] Semantic Refinement GRU-based Neural Language Generation for Spoken Dialogue Systems. Kôiti Hasida [n. d.], 63.

[90] Svitlana Vakulenko, S. Longpre, Zhucheng Tu, and R. Anantha. 2021. Question Critiquing for Recommender Systems. In Proceedings of the 31st International Conference on Neural Information Processing Systems. 6000–6010.

[91] Marilyn Walker and Steve Whittaker. 1990. Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation. In Proceedings of the 28th annual meeting on Association for Computational Linguistics. 70–78.

[92] Dongsheng Wang, Qiuchi Li, Lucas Chaves Lima, Jakob Grue Simonsen, and Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. ArXiv abs/2109.07154 (2021).

[93] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji rong Wen, and Jingsong Yu. 2020. Explainable Recommendation: A Survey and New Perspectives. Found. Trends Inf. Retr. 14 (2020), 1–101. https://doi.org/10.1561/1500000066

[94] Jie Zou, Yifan Chen, and E. Kanoulas. 2020. Towards Question-based Recom- mendation as Search Simulation: A General Formal Framework for IR Evaluation. Proceedings of the 14th International ACM SIGIR Conference on Research & Development in Information Retrieval (2021). https://doi.org/10.1145/3397271.3401180

[95] Xintong Yu, Hongming Zhang, Yangqiu Song, Tian Song, and Changzhui Zhang. 2019. What You See is What You Get: Visual Pronoun Coreference Resolution in Conversations. In Conference on Empirical Methods in Natural Language Processing. 2080–2087. https://doi.org/10.1109/BIBM52615.2021.9669650

[96] Xiaoxue Zhao, Weinan Zhang, and Jun Wang. 2013. Interactive Collaborative Filtering. Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 235–242.

[97] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. Found. Trends Inf. Retr. 14 (2020), 1–101. https://doi.org/10.1561/1500000066

[98] Chang Zhou, Hang Chen, and Kun Zhou. 2021. Can Language Models be Biomedical Knowledge Bases? ArXiv abs/2109.07154 (2021).

[99] Jheng-Hong Yang, Sheng-Chieh Lin, Chuan-Ju Wang, Jimmy J. Lin, and Ming-Feng Tsai. 2019. Query and Answer Expansion from Conversation History. In Text Retrieval Conference.

[100] Honghan Wu, Giulia Toti, Katherine I Morley, Zina M. Ibrahim, Amos A. Fararin, Richard G. Jackson, Ismail Emre Karloglu, Asha Agrawal, Clive Stringer, Darren Gale, Genevieve Gorrell, Angus Roberts, Matthew T. M. Broadbent, Robert J Stewart, and Richard J. B. Dobson. 2018. SemiEHR: A General-purpose Semantic Search System to Surface Semantic Data from Clinical Notes for Tailored Care, Trial Recruitment, and Clinical Research. "Journal of the American Medical Informatics Association : JAMIA 25 (2018), 530 – 537. https://doi.org/10.1101/235622

[101] Jiadun Wu, Ruonan Zhang, Tieqiang Gong, Yang Lin, Chunbao Wang, and Chen Li. 2021. BioE: Biomedical Information Extraction with Multi-head Attention Enhanced Graph Convolutional Network. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2080–2087. https://doi.org/10.1109/ BIBM52615.2021.9669650

[102] Jianqiong Xiao and Xiaqing Zhou. 2022. Attention-based Generative Adversarial Hashing for Cross-modal Retrieval. In 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Vol. 10. 1885–1889. https://doi.org/10.1109/ITAIC54216.2022.9836826

[103] Zheng-Hong Yang, Sheng-Chieh Lin, Chuan-Ju Wang, Jimmy J. Lin, and Ming-Feng Tsai. 2019. Query and Answer Expansion from Conversation History. In Text Retrieval Conference.

[104] Lu Yang, Hamed Zamani, Yongfeng Zhang, J. Guo, and W. Bruce Croft. 2017. Neural Matching Models for Question Retrieval and Next Question Prediction in Conversation. ArXiv abs/1707.05409 (2017).

[105] Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changzhui Zhang. 2019. What You See is What You Get: Visual Pronoun Coreference Resolution in Conversations. In Conference on Empirical Methods in Natural Language Processing.

[106] Hamed Zamani, Johanne R. Trippe, Jeffrey Dalton, and Filip Radlinski. 2022. Conversational Information Seeking. ArXiv abs/2201.08808 (2022).

[107] Lisa Yang, Hamed Zamani, Yongfeng Zhang, J. Guo, and W. Bruce Croft. 2017. Explainable Recommendation: A Survey and New Perspectives. Found. Trends Inf. Retr. 14 (2020), 1–101. https://doi.org/10.1561/1500000066

[108] Yinan Zhang, Xueqing Liu, and ChengXiang Zhai. 2017. Information Retrieval Knowledge Bases and Word Embeddings. Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 637–646.

[109] Xiaomeng Zhang and Xueqing Liu. 2016. Explainable Recommendation: A Survey and New Perspectives. Found. Trends Inf. Retr. 14 (2020), 1–101. https://doi.org/10.1561/1500000066

[110] Jianqiong Xiao and Xiaoqing Zhou. 2022. Attention-based Generative Adversarial Hashing for Cross-modal Retrieval. In 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Vol. 10. 1885–1889. https://doi.org/10.1109/ITAIC54216.2022.9836826

[111] Jie Zou, Yifan Chen, and E. Kanoulas. 2020. Towards Question-based Recom- mendation as Search Simulation: A General Formal Framework for IR Evaluation. Proceedings of the 14th ACM Conference on Recommender Systems. 137–145.

[112] Kang. 2021. Can Language Models be Biomedical Knowledge Bases? ArXiv abs/2201.11903 (2022).

[113] Tsung-Hsien Wen, Milica Gasic, Dongho Kim, Nikola Mrksic, Pei Hao Su, David Vandyke, and Steve Young. 2015. Stochastic Language Generation in Dialogue using Recursive Neural Networks with Convolutional Sentence Reranking. ArXiv abs/1508.01755 (2015). https://doi.org/10.48550/arXiv.1508.01755

[114] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei Hao Su, David Vandyke, and Steve Young. 2015. Stochastically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. arXiv preprint arXiv:1508.01745 (2015).

[115] Gu Wu, Kai Luo, Scott Sanner, and Harold Soh. 2019. Deep Language-based Critiquing for Recommender Systems. In Proceedings of the 13th ACM Conference on Recommender Systems. 137–145.

[116] Feng Tsai. 2019. Query and Answer Expansion from Conversation History. In Text Retrieval Conference.

[117] Xueming Sun and Yi Zhang. 2018. Conversational Recommender System. The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (2018). https://doi.org/10.48550/arXiv.1806.03277

[118] Mujirem Sung, Jinhyuk Lee, Sean S. Yi, Ming Jieon, Sungdong Kim, and Jaewoo Kang. 2021. Can Language Models be Biomedical Knowledge Bases? ArXiv abs/2109.07154 (2021).

[119] Van-Kha Nh Tran and Le-Minh Nguyen. [n. d.] Semantic Refinement GRU-based Neural Language Generation for Spoken Dialogue Systems. Kôiti Hasida [n. d.], 63.

[120] Svitlana Vakulenko, S. Longpre, Zhucheng Tu, and R. Anantha. 2021. Question Critiquing for Recommender Systems. In Proceedings of the 31st International Conference on Neural Information Processing Systems. 6000–6010.

[121] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanzhang Zhou, Ji rong Wen, and Jingsong Yu. 2020. Improving Conversational Recommender Systems via Knowledge Graph based Semantic Fusion. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2020). https://doi.org/10.48550/arXiv.2007.04032

[122] Jie Zou, Yifan Chen, and E. Kanoulas. 2020. Towards Question-based Recom- mendator Systems. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2020). https://doi.org/10.1145/3397271.3401180