Conversational AI from an Information Retrieval Perspective: Remaining Challenges and a Case for User Simulation

Krisztian Balog

University of Stavanger, Norway

Abstract

Conversational AI is an emerging field of computer science that engages multiple research communities, from information retrieval to natural language processing to dialogue systems. Within this vast space, we focus on conversational information access, a problem that is uniquely suited to be addressed by the information retrieval community. We argue that despite the significant research activity in this area, progress is mostly limited to component-level improvements. There remains a disconnect between current efforts and truly *conversational* information access systems. Apart from the inherently challenging nature of the problem, the lack of progress, in large part, can be attributed to the shortage of appropriate evaluation methodology and resources. This paper highlights challenges that render both offline and online evaluation methodologies unsuitable for this problem, and discusses the use of user simulation as a viable solution.

Keywords

Conversational information access, conversational AI, user simulation

1. Introduction

Conversational AI may be seen as the holy grail of computer science: building machines that are capable of interacting with people in a human-like way. With rapid advances in AI technology, there are reasons to believe that such an ambition is within reach [1]. Conversational AI is a vast and complex problem, which requires a combination of methods, tools, and techniques from multiple fields of computer science, including but not limited to artificial intelligence (AI), natural language processing (NLP), machine learning (ML), dialogue systems (DS), recommender systems (RecSys), human-computer interaction (HCI), and not the least information retrieval (IR). Each of these fields may have its own particular interpretation of what conversational AI should entail and a specific focus on certain research challenges that are involved. For example, in spoken dialogue systems the main motif is to be able to *talk* to machines, i.e., on developing speech-based human-computer interfaces [2], and thus automatic speech recognition is a central component. Many other communities, on the other hand, assume a chat-based interface and voice is not among the supported modalities. At the same time, there are many shared aspects, including handling the semantics involved in the dialogue process, generating contextually appropriate responses, and developing effective end-to-end (neural) architectures, which engage multiple research communities. In this paper, we focus on the problem of *conversational information access* (CIA), one that the IR community is uniquely suited to address.

Conversational search or *conversational information seeking* has already been identified in 2012 as a research direction of strategic importance in IR [3], and its significance has been re-iterated in 2018 [4]. There, the problem focus has been defined to include complex user goals that require multi-step information seeking, exploratory information gathering, and multi-step task completion and recommendation, as well as dialog settings with variable communication channels. Our analysis of recent works, however, leads us to the observation that current efforts do not seem to be fully aligned with the directions set out there. In terms of end-to-end tasks, there are two main threads of work: conversational QA and conversational recommendations. Currently, these are treated as two separate types of systems, with different goals, architectures, and evaluation criteria. Instead, for a more effective *assistance* of users, the two should be seamlessly integrated in CIA systems, thereby moving from a siloed to a more unified view. Additionally, the multi-modality of interactions needs to be more fully embraced, in order to more *actively* support effective interaction [5]. On the component level, most proposed techniques are not truly *conversational* in the sense that they are applicable to any interactive IR system (e.g., modern web search engines). A critical blocker to progress, on both the end-to-end and component levels, is the shortage of appropriate evaluation methodology and resources.

DESIRES 2021 – 2nd International Conference on Design of Experimental Search & Information REtrieval Systems, September 15–18, 2021, Padua, Italy

krisztian.balog@uis.no (K. Balog)
https://krisztianbalog.com/ (K. Balog)
0000-0003-2762-721X (K. Balog)

Table 1
Categorization of conversational AI systems, based on [1, 6].

Task-oriented	Social chat	Interactive QA
Aim to assist users to solve **a specific task** (as efficiently as possible)	Aim to **carry on an extended conversation** ("chit-chat") with the goal of mimicking human-human interactions	Aim to **provide concise, direct answers** to user queries
Dialogues follow a **clearly designed structure** (flow) that is developed for a particular task in a closed domain	Developed for **unstructured, open domain** conversations	Dialogues are **unstructured**, but commonly follow a question-answer pattern; mostly **open domain** (dictated by the underlying data)
Well-defined measure of performance that is explicitly related to **task completion**	Objective is to be **human-like**, i.e., able to talk about different topics (breadth and depth) in an engaging and coherent manner	Evaluated with respect to the **correctness of answers** (on the turn level)

In summary, the contributions of this paper are threefold.

- We argue for a broader interpretation of conversational information access, one that embraces multiple user goals (mixing task-oriented and QA elements) and multi-modal interactions (Sect 2).
- We provide a synthesis of progress on conversational information access and identify open challenges around methods and evaluation (Sect. 3).
- We argue for (a more extensive) use of simulation as a viable evaluation paradigm for conversational information access and describe a simulator architecture (Sect. 4).

2. Defining Conversational Information Access

This section defines conversational information access and places it in the broader context of conversational AI.

2.1. Conversational AI: The big picture

Conversational AI[1] is casually used to denote a broad range of systems that are capable of (some degree of) natural language understanding and responding in a way that mimics human dialogue. A conversational AI system may thus be considered successful if it offers an experience that is indistinguishable from what could have been delivered by a human. These systems often focus on a particular type of conversational support, naturally lending themselves to categorization.

[1]In this paper, the terms *conversational AI*, *conversational agent*, and *dialogue system* are used interchangeably. We, however, avoid using the term *chatbot*, which has a different meaning in industrial and academic contexts; in the former case it refers to a task-oriented system, while in the latter it means a non-task-oriented system [7].

2.1.1. Traditional 2-way Categorization

Traditionally, conversational agents are categorized as being *goal-driven* (or *task-oriented*) or *non-goal-driven* (also known as *chatbots*) [8, 9, 7]. *Goal-driven systems* aim to assist users to complete some specific task. Dialogues are constrained to a specific domain and characterized by having a designated structure, designed for particular tasks within that domain. The main success criteria for the conversational agent is its ability to help the user solve their task as efficiently as possible. Typical examples include travel planning and appointment scheduling.

Non-goal-driven systems, on the other hand, aim to carry on an extended conversation ("chit-chat") with the goal of mimicking unstructured human-human interactions. The main purpose of these systems is usually entertainment or providing an "AI companion" [10]. Therefore, the objective for these systems is to be able to talk about different topics in an engaging and cohesive manner.

2.1.2. Contemporary 3-way Categorization

Most recently, the traditional categorization has been extended with a third category, *interactive question answering* (QA) [1, 6], in recognition of the fact that it fits neither in task-oriented nor in social chat, but deserves a separate category on its own right. Interactive QA systems are designed to provide answers to specific questions. They are not characterized by a rigid dialogue flow, although they typically follow a question-answer pattern. Apart from some notable recent examples [11], the human-like conversation aspect for QA systems is much less pronounced than for the other two types of systems, and evaluation is restricted to answer correctness.

Table 1 summarizes the characteristics of the three categories of conversational AI systems. Given their unique goals and objectives, each of these problem categories is addressed by a distinctive system architecture [1, 6].

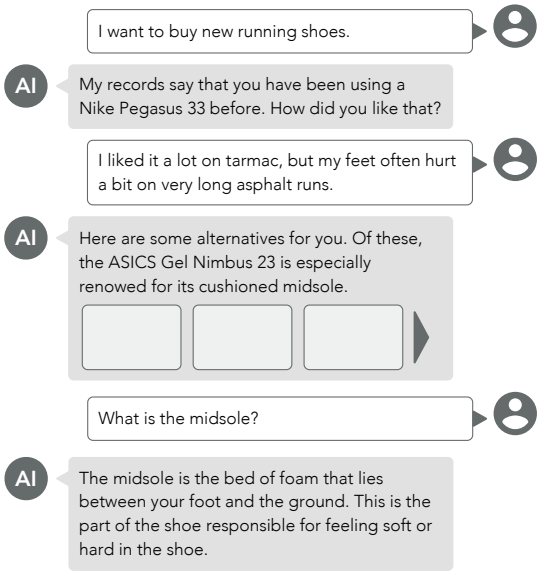

Figure 1: Envisioned dialogue with a CIA system.

2.2. Conversational Information Access

Building on [4], we use the term *conversational information access* (CIA) to define a subset of conversational AI systems that specifically aim at a task-oriented sequence of exchanges to support multiple user goals, including search, recommendation and exploratory information gathering, that require multi-step interactions over possibly multiple modalities. Further, these systems are expected to learn user preferences, personalize responses accordingly, and be capable of taking initiative.

Consider the conversation shown in Fig. 1, illustrating some of the above requirements. It is primarily a task-oriented dialogue (the user wanting to buy new running shoes), which requires an exploration of the item space. Assuming a chat-based interface, this can be done most effectively by combining multiple modalities; not just text, but also a carousel for cycling through items, in this example. Up until the second user utterance, it is a strictly task-oriented sequence of exchanges (cf. the task-oriented category in Table 1). But, then, the third user utterance breaks out of the task flow and switches to "QA mode" (cf. interactive QA in Table 1).

2.2.1. From Siloed to Unified View

One key realization the above example is meant to illustrate is that conversational information access cuts across the task-oriented and interactive QA categories. This blending makes CIA suited to *assist* users meaningfully with their needs. Conversely, existing work—and, importantly, evaluation initiatives—in IR almost exclusively focus on a question-answering paradigm (see Sect. 3). This does not allow for *interaction with sets of items*—one of the main properties that makes a search system conversational [12].

It has been shown that the "siloed" view, represented by the three categories in Table 1, in practice does not align well with users' information needs and behavior [13]. Gao et al. [1] acknowledge the need for a "top-level bot" that would act as a broker and switch between different user goals. Most commercial assistants are hybrid systems, with different degrees of support for switching. There is, however, little published research on it. In summary, there is need for a more holistic view where multiple user goals are supported.[2]

2.2.2. Multi-modality

Another key point highlighted by the example in Fig. 1 is the need for embracing multi-modality. Text-only responses are motivated by an audio-only channel, without a screen [14]. However, more often than not a chat-base interface is available, which allows for a richer set of input controls and navigational components. These, in turn, would enable CIA systems to more actively support effective interaction [5]. We note that the need for multi-modality has been recognized independently by other scholars as well [15].

3. Progress to Date and Remaining Challenges

In this section, we reflect on progress achieved so far, organized around methods and evaluation, and identify remaining challenges.

3.1. Methods

In our discussion, we distinguish between end-to-end conversational tasks and specific component-level subtasks.

3.1.1. End-to-end tasks

There are two main tasks that have received attention: conversational QA [16, 1, 17] and conversational recommendations [18, 19, 20]. What distinguishes conversational QA from traditional single-turn QA is the need for contextual understanding. Hence, much of the research revolves around modeling conversation history [16, 17, 21]. However, in terms of evaluation, the

[2]We note that this problem is not specific to IR. However, conversational information access is a good starting point that the IR community is uniquely suited to address. Lessons and finding could then be generalized to broader applications.

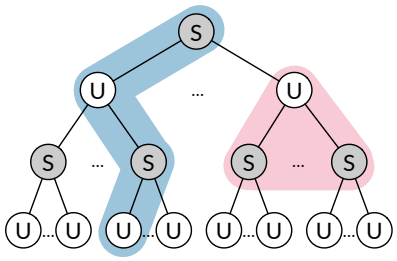

Figure 2: The space of possible dialogue states increases exponentially with the number of turns between system (S) and user (U). Evaluation is currently limited to either a single path (blue area) or a single turn (red area).

problem is simplified to a single-turn passage retrieval task, where the relevance of system response at a given turn does not consider the responses given by the system at earlier turns [22, 17]. It is only conversational recommender systems where the multi-turn nature of conversations is more fully embraced [18, 19, 20].

3.1.2. Component-level sub-tasks

Recently, progress has been made on specific subtasks for CIA, including response retrieval [23] and generation [24], query resolution [25, 26], asking clarifying questions [27] or suggestion questions [28], predicting user intent [29], and preference elicitation [18, 30]. Each of these studies makes the point that the conversational setting calls for a different set of approaches. However, most of these subtasks are applicable in any interactive IR context, adhering to the stance that search is inherently a conversational experience: it is a dialogue between a human and a search engine [31]. From this perspective, there has been substantial progress, and especially on the *mixed initiative* aspect, e.g., question clarifications and suggestions [27, 28]. Alternatively, one may take a more critical stance and ask: What separates conversational information access from any other interactive IR system (most prominently: search engines)? According to Croft [5], the key distinguishing factor is that a conversational system is more *active* partner in the interaction. From that regard, there is surprisingly little work, with only a handful of notable exceptions [32, 11].

3.2. Evaluation

3.2.1. Offline Evaluation

Traditionally, system-oriented evaluation in IR has been performed using *offline* test collections, following the Cranfield paradigm [33]. This rigorous methodology ensures the repeatability and reproducibility of experiments, and has been instrumental to progress in the field. To date, work on CIA still employs offline evaluation [22, 27], but this has severe limitations. First, reusability requires that the system is limited in selecting the best response, in answer to a user utterance, from a restricted set of possible candidates (i.e., some predefined corpus of responses). Second, it is limited in scope to a single conversation turn and does not consider dialogue history that led to that particular user utterance (cf. red area in Fig. 2). An alternative is to let human evaluators assess an entire conversation, once it has taken place [6]. However, this is a single path (see blue area in Fig. 2), without considering the other choices the user could have taken during the course of the dialogue. Moreover, it is expensive, time-consuming, and does not scale. Most importantly, it would not yield a reusable test collection. In summary, offline test collections have their merits, but their use is limited to the purpose of evaluating specific components, in isolation. Further, the choice of evaluation metrics is an open challenge [34].

3.2.2. Online Evaluation

Online evaluation involves fielding an IR system to real users, and observing how they interact with the system *in situ*, in their natural task environments [35]. This requires a live service as a *research platform*. Currently, this possibility is only available to researchers working at major service providers that develop conversational assistants (Google, Microsoft, Apple, Amazon). Even there, experimentation with live users is severely limited due to scalability, quality, and ethical concerns. Of these companies, only Amazon has decided to open up its platform for academic research, by organizing the Alexa Prize Challenge [36]. It represents a unique opportunity for academics to perform research with a live system used by millions of users, and provides university teams with real user conversational data at scale. While this effort points in the right direction, it is inherently limited in that it addresses social conversations ("chit-chat"), with the target goal of conversing coherently and engagingly with humans on popular topics such as sports, politics, or technology for 20 minutes. This is a non-goal-driven task, which is rather different from goal-driven CIA. Currently, there is no publicly available research platform for CIA. *Living labs* represents a novel evaluation paradigm for IR [37], which allows researchers to evaluate their methods with real users of live search services. This methodology has been successfully employed at world wide benchmarking campaigns [38, 39]. It, however, needs to be extended to a conversational setting, which brings about methodological and practical challenges.

3.2.3. User Simulation

With a long history in the field of spoken dialogue systems, *user simulation* is seen as a critical tool for automatic dialogue management design [40]. The idea is to train a *user model* that is "capable of producing responses that a real user might have given in a certain dialog situation" [40]. This is in line with our goals, but there are two crucial differences. First, the primary purpose of user simulation in DS is to generate a synthetic training data at scale, which in turn can be used to learn dialogue strategies (typically, using reinforcement learning). Assessment of the quality of simulated dialogues and user simulation methods, however, is an open issue [41]. Second, dialogue systems, as well as recent work on conversational recommender systems [18], are focused on supporting the user with a single goal that can be fulfilled by eliciting preferences on a set of attributes. CIA systems, on the other hand, need to deal with complex search and recommendation scenarios. This requires a more holistic user model.

3.3. Summary and Remaining Challenges

3.3.1. Understanding User Needs and Behavior

Current characterizations of information seeking behavior for CIA are limited either in the set of actions considered [42] or in sequences of conversational turns [43]. To cater for the functionality defined by Radlinski and Craswell [12] and further expanded by us in Sect 2.2, one would need user and interaction models capable of representing (1) multi-modal interactions (speech, text, pointing&clicking), (2) users' ability to change their state of knowledge (learn and forget) and (3) users' ability to learn how a system works and what its limits are (and change their expectations and behavior accordingly).

3.3.2. Truly Conversational Methods

Conversational recommendations and QA have been studied as end-to-end tasks. However, as we argued in Sect. 2.2, in practice these two are not clearly delineated applications, but rather different "modes" that should be seamlessly integrated with a CIA system. There has been significant progress on various components, which are indispensable building blocks. Integrating these into a *unified system* that supports multiple user goals remains an open challenge [1]. Further open questions in this space include (1) deciding when and what type of *initiative* a system should take, and (2) determining the best *modality* based on task and context.

3.3.3. Evaluation

There is a need to go beyond turn-based evaluation to multi-turn-based and eventually end-to-end evaluation. To be able to perform end-to-end evaluation of CIA systems, additional methodologies need to be considered, including online evaluation and simulated users. For online evaluation, the living labs paradigm represents an alternative, but it requires agreement on a canonical architecture in order to be able to open up individual components for experimentation. Further, it requires an existing service with live users, which is currently lacking. It should be noted that the need for such an open research platform has been identified and a plan for the academic search domain has recently been outlined [44].

As for simulation, most existing approaches are meant to advance reinforcement learning techniques in a strictly goal-oriented setting. This is different from our purpose of evaluation. The simulation techniques that are currently used for evaluation lack the desired conversational complexity.

4. A Case for Simulation

This section presents a proposal for robust large-scale automatic evaluation of CIA systems via user simulation.

4.1. Methodology

Our main hypothesis is that it is possible to simulate human behavior with regard to interacting with CIA systems. To validate this hypothesis, we need to show that simulated users behave indistinguishable from real humans, in the context of a specific conversational application and with respect to specific evaluation measures.

Formally, let S_1 and S_2 denote two CIA systems, which differ in some component(s). Both systems are assumed to be operated by a set U of users from some user population. Let us assume that there is a statistically significant difference observed in their relative performance, according to some evaluation measure M, such that $M(S_1, U) < M(S_2, U)$. Simulation is considered successful, if by engaging a set U^* of simulated users, we observe the same relative system differences as with real users, i.e., $M(S_1, U^*) < M(S_2, U^*)$. Further, this observation should generalize across systems S and evaluation measures M.

The above formulation ensures that the behavior of simulated users aligns with those of real users. Notice that to be able to perform this validation, an operational CIA system is needed; we discuss the practical aspects of setting up such an experimental platform below, in Sect. 4.4. For the human evaluation part, i.e., measuring $M(S, U)$, two distinct approaches may be employed: (1) asking users themselves inside the CIA system to

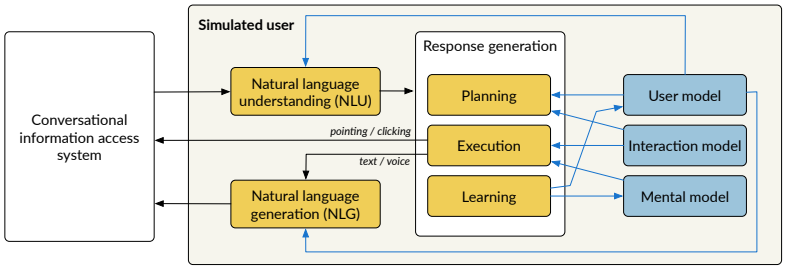

Figure 3: Conceptual architecture of the user simulator.

give feedback on either the entire conversation or on specific system utterances, and (2) sampling interesting/meaningful branches from conversation logs, which will be annotated by external human labelers (e.g., crowd workers).

Once a user simulator is created and validated against real users, it may be used evaluating a given CIA system. It is important to note that, in principle, a given user simulator instance should be used only once, the same way that an offline test collection should only be used once—to avoid overfitting systems against a particular test suite.

4.2. Requirements

We identify a realistic user simulator with the ability of capturing:

- **(R1)** Personal interests and preferences, and the changes of preferences over time;
- **(R2)** Persona (personality, educational and socio-economical background, etc.);
- **(R3)** Multi-modality of interactions (speech, text, pointing&clicking, etc.);
- **(R4)** The user's ability to change their state of knowledge (learn and forget);
- **(R5)** The user's ability to learn how a system works and what its limits are, and change their expectations and behavior accordingly.

We note that not all these requirements are critical for an initial simulator and some may be highly ambitious. Nevertheless, we shall discuss our conceptual architecture with reference to these requirements.

4.3. Architecture

Figure 3 shows the conceptual architecture of a user simulator addressing the stated requirements. We discuss its main components below, and provide specific starting points for each of them.

User, interaction, and mental models provide the foundation for simulation behavior.

- **User model**. To represent all personal information related to a given user, including persona (**R2**), preferences (**R1**), and knowledge (**R4**), *personal knowledge graphs* (PKGs) [45] may be used. The reason for using a PKG is to ensure the consistency of the preferences that are revealed by the simulated user, as it is done in [46]. To fully address **R1** and **R4**, PKGs will need to be extended along two dimensions: (1) include concepts, in addition to entities, to represent the user's knowledge on specific topics, with further distinction to be made between entities/concepts the user heard about vs. has in-depth knowledge on; (2) capture the temporal scope, to be able to distinguish between short- vs. long-term preferences and fresh vs. diminishing knowledge.
- **Interaction model**. To characterize the CIA process between humans and systems for a given application, the key actions and decisions that manifest in dialogues need to be abstracted out. A starting point for a taxonomy of user/system actions is provided in [47]. This taxonomy may be revised and extended to multi-modal interactions (**R3**) based on conversations collected in laboratory user studies with an "idealized" CIA system using the Wizard-of-Oz approach [48] and from interaction data from actual CIA systems.
- **Mental model**. To capture how a particular user thinks about a given CIA system (**R5**), mental models need to be developed. The thinking aloud method is commonly used for such purposes in usability testing, psychology, and social sciences [49]. There is work in HCI on identifying and analyzing experiences and barriers qualitatively [50, 51, 52, 53]. A main difference from those studies is that the goal here is to build a *quantifiable* mental model that represents the user's expectations and perceived capabilities of a CIA system.

Next, we describe the components responsible for interacting with CIA systems.

- **Natural language understanding**. Obtaining a structured representation from a system utterance is analogous to NLU in dialogue systems and involves *domain classification*, *intent determination*, and *slot filling* [7]. These tasks are effectively tackled by neural architectures [54, 55, 1]. These approaches, however, are created for conversational *systems* and assume "perfect" world knowledge, based on some underlying knowledge repository. For user simulation, they need to be adapted to consider personal knowledge. For example, the user may or may not be able to guess the corresponding type or category of an entity/concept that is mentioned for the first time, depending on their knowledge of the given domain.
- **Response generation**. Determining how a simulated user should respond to a system utterance is modeled in three stages: planning, execution, and learning. In the *planning* stage, a structured representation of an information need (what to ask the system) or user response (how to respond to if prompted by the system) is generated. This is informed by the user model, in terms of interests and preferences, as well as the interaction model, to help interpret what the system is asking in terms of a task-specific dialog flow. In the *execution* stage, the simulator decides on the course of execution, based on the user's mental model of the given system's capabilities (e.g., it will not attempt to navigate a list using voice, but rather click, if voice navigation did not function in the past as expected). Based on how the system responds to a given user utterance, the *learner* module can make updates to the user model (whether the user learned something new about a given topic) and also to the mental model of the system (how successful it was in understanding/executing what was requested). Response generation can be framed within the well-established agenda-based simulation approach [56].
- **Natural language generation**. Finally, a structured intent representation (what to say to the system) needs to turned into a natural language utterance (how to say it). The exact articulation is influenced by the persona and knowledge level of the simulated user. A possible starting point is to generate templated responses and then apply transfer learning for text [57, 58, 59]. Later, more end-to-end approaches may also be devised, eliminating the need for manual template generation. It should be noted that not all requests get passed through NLG, as the executor may decide to use a different modality.

Each simulated user requires *instantiating* the user and mental models. (The interaction model is application-specific and is shared by all simulated users.) To make the *user model* realistic, it should be anchored in actual user profiles (while maintaining *k-anonymity*). For that, a generative model may be used, with parameters learned on publicly available corpora, e.g., item ratings for recommendation scenarios [46] and discussion fora for information seeking tasks. The *mental model* may be initialized using a small set of pre-trained skill profiles, created as part of laboratory user studies.

From a system architecture perspective, the user simulator in many regards resembles a CIA system, comprising of natural language understanding, dialog management, and natural language generation components. One major difference is that CIA systems may be assumed (in fact, expected) to have "perfect world knowledge," only limited by the availability of data. Conversely, user simulation also needs to consider the user's knowledge level in language understanding and generation. Another major difference is that while a CIA system is modeled after a single person, each simulated user has a unique persona. This requires each of the components to be parametrizable with respect to personal characteristics. Further, the choice of dialogue actions is affected by the user's mental model of the system (i.e., what the system is perceived to be able to understand and execute).

4.4. Operationalization

Note that simulation capability is application specific. That is, different simulators would need to be trained for item recommendation, interactive QA, and, ultimately, for scenarios that cater for multiple user goals. To ensure that the behavior of simulated users aligns with that of human users, an operational CIA system with actual users would also be needed for each application. Setting up such applications should be seen as a community effort. Indeed, discussions in this direction have already begun and one specific proposal for a CIA system supports scholarly activities has been outlined in [44]. There are a number of challenges involved in building a CIA system that can serve as such a living lab. One is that it would have insufficient traffic for meaningful online evaluation (an issue that has indeed been encountered in the past [38]). To remedy that additional users may be recruited, e.g., by involving students as part of their course work or hiring workers on crowdsourcing platforms (i.e., increasing traffic volume). Another potential difficulty is that building a sufficiently performant CIA system for the application at hand turns out to be too challenging (thereby making the online service unattractive to users). While this is not easily solvable on the system front, it is possible to manage users' expectations. Indeed, one of the key ideas behind operating in the academic domain in [44] is to build a tool by researchers to researchers, and embrace its imperfection.

Simulation approaches are evaluated by comparing them against real users on a given live research platform. In practice this means that a small portion of the usage data collected from humans (i.e., first few weeks of the live evaluation period) is disclosed and can be used for training the simulators, while the remaining data is used for evaluating them. The set of systems participating in the live evaluation (referred to as *experimental systems*) are also evaluated using the different simulators. Ultimately, the question we seek to answer is whether we can observe the same relative ranking of experimental systems with real users (based on the live experiment) as with simulated ones—being able to answer this question positively would mean that the simulator is sufficiently *realistic*.

5. Conclusions and Future Directions

In this paper, we have considered conversational AI from an IR perspective, and focused in particular on the problem of conversational information access, with the goal to identify open challenges that the IR community is uniquely suited to address.

One critical area concerns the understanding of users' information needs and their information seeking behavior—one of fundamental research directions in IR from the very beginning [60]. Currently, there is a lack of understanding of what would be desirable conversational experiences for information access scenarios that combine multiple user goals. Consequently, there are no suitable models of user behavior that could serve as foundations for unified architectures that can support such behavior.

Another aspect that represents a major open challenge is evaluation. *Measurement* is an area where IR has an unparalleled history [61, 62, 63, 64, 33, 35]. Building on the rich tradition and experience of community benchmarking campaigns such as TREC [62] and CLEF [63], our community is in a unique position to take a lead on the development of novel evaluation paradigms and methodologies. This paper has outlined a specific plan for such an effort along user simulation.

References

[1] J. Gao, M. Galley, L. Li, Neural approaches to conversational AI, Found. Trends Inf. Retr. 13 (2019) 127–298.

[2] M. F. McTear, Spoken dialogue technology: Enabling the conversational user interface, ACM Comput. Surv. 34 (2002) 90–169.

[3] J. Allan, B. Croft, A. Moffat, M. Sanderson, Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne, SIGIR Forum 46 (2012) 2–32.

[4] J. S. Culpepper, F. Diaz, M. D. Smucker, Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in Lorne (SWIRL 2018), SIGIR Forum 52 (2018) 34–90.

[5] W. B. Croft, The importance of interaction for information retrieval, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '19, 2019, pp. 1–2.

[6] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, M. Cieliebak, Survey on evaluation methods for dialogue systems, Artificial Intelligence Review (2020) 1573–7462.

[7] D. Jurafsky, J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 3nd Edition draft, Prentice Hall, Pearson Education International, 2019.

[8] H. Chen, X. Liu, D. Yin, J. Tang, A survey on dialogue systems: Recent advances and new frontiers, SIGKDD Explor. Newsl. 19 (2017) 25–35.

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

[10] L. Zhou, J. Gao, D. Li, H.-Y. Shum, The design and implementation of XiaoIce, an empathetic social chatbot, Comput. Linguist. 46 (2020) 53–93.

[11] I. Szpektor, D. Cohen, G. Elidan, M. Fink, A. Hassidim, O. Keller, S. Kulkarni, E. Ofek, S. Pudinsky, A. Revach, S. Salant, Y. Matias, Dynamic composition for conversational domain exploration, in: Proceedings of The Web Conference 2020, WWW '20, 2020, pp. 872–883.

[12] F. Radlinski, N. Craswell, A theoretical framework for conversational search, in: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR '17, 2017, pp. 117–126.

[13] Z. Yan, N. Duan, P. Chen, M. Zhou, J. Zhou, Z. Li, Building task-oriented dialogue systems for online shopping, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI '17, 2017, pp. 4618–4625.

[14] J. R. Trippas, Spoken Conversational Search: Audio-only Interactive Information Retrieval, Ph.D. thesis, RMIT University, 2019.

[15] Y. Deldjoo, J. R. Trippas, H. Zamani, Towards multimodal conversational information seeking, in: Proceedings of the 43th International ACM SIGIR Con-

ference on Research and Development in Information Retrieval, SIGIR '21, 2021, pp. 1577–1587.
[16] C. Qu, L. Yang, M. Qiu, W. B. Croft, Y. Zhang, M. Iyyer, Bert with history answer embedding for conversational question answering, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '19, 2019, pp. 1133–1136.
[17] S. Reddy, D. Chen, C. D. Manning, CoQA: A conversational question answering challenge, Transactions of the Association for Computational Linguistics 7 (2019) 249–266.
[18] Y. Zhang, X. Chen, Q. Ai, L. Yang, W. B. Croft, Towards conversational search and recommendation: System ask, user respond, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, 2018, pp. 177–186.
[19] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, 2020. arXiv:2004.00646.
[20] C. Gao, W. Lei, X. He, M. de Rijke, T. Chua, Advances and challenges in conversational recommender systems: A survey, 2021. arXiv:2101.09459.
[21] C. Zhu, M. Zeng, X. Huang, SDNet: Contextualized attention-based deep network for conversational question answering, 2018. arXiv:1812.03593.
[22] J. Dalton, C. Xiong, J. Callan, TREC CAsT 2019: The Conversational Assistance Track overview, 2020. arXiv:2003.13624.
[23] L. Yang, M. Qiu, C. Qu, J. Guo, Y. Zhang, W. B. Croft, J. Huang, H. Chen, Response ranking with deep matching networks and external knowledge in information-seeking conversation systems, in: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '18, 2018, pp. 245–254.
[24] Y. Song, C.-T. Li, J.-Y. Nie, M. Zhang, D. Zhao, R. Yan, An ensemble of retrieval-based and generation-based human-computer conversation systems, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI '18, 2018, pp. 4382–4388.
[25] N. Voskarides, D. Li, P. Ren, E. Kanoulas, M. de Rijke, Query resolution for conversational search with limited supervision, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, 2020, pp. 921–930.
[26] S. Vakulenko, N. Voskarides, Z. Tu, S. Longpre, A comparison of question rewriting methods for conversational passage retrieval, in: Proceedings of the 43rd European Conference on IR Research, ECIR '21, 2021, pp. 418–424.
[27] M. Aliannejadi, H. Zamani, F. Crestani, W. B. Croft, Asking clarifying questions in open-domain information-seeking conversations, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '19, 2019, pp. 475–484.
[28] C. Rosset, C. Xiong, X. Song, D. Campos, N. Craswell, S. Tiwary, P. Bennett, Leading conversational search by suggesting useful questions, in: Proceedings of The Web Conference 2020, WWW '20, 2020, pp. 1160–1170.
[29] C. Qu, L. Yang, W. B. Croft, Y. Zhang, J. R. Trippas, M. Qiu, User intent prediction in information-seeking conversations, in: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR '19, 2019, pp. 25–33.
[30] K. Christakopoulou, F. Radlinski, K. Hofmann, Towards conversational recommender systems, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, 2016, pp. 815–824.
[31] P. Ren, Z. Chen, Z. Ren, E. Kanoulas, C. Monz, M. de Rijke, Conversations with search engines, 2020. arXiv:2004.14162.
[32] S. Zhang, Z. Dai, K. Balog, J. Callan, 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, 2020, pp. 1537–1540.
[33] M. Sanderson, Test collection based evaluation of information retrieval systems, Found. Trends Inf. Retr. 4 (2010) 247–375.
[34] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, J. Pineau, 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, EMNLP '16, 2016, pp. 2122–2132.
[35] K. Hofmann, L. Li, F. Radlinski, Online evaluation for information retrieval, Found. Trends Inf. Retr. 10 (2016) 1–117.
[36] A. Ram, 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, A. Pettigrue, Conversational AI: The science behind the Alexa Prize, 2018. arXiv:1801.03604.
[37] A. Schuth, K. Balog, Living labs for online evaluation: From theory to practice, in: Proceedings of the 38th European conference on Advances in Information Retrieval, ECIR '16, 2016, pp. 893–896.
[38] R. Jagerman, K. Balog, M. D. Rijke, Opensearch: Lessons learned from an online evaluation cam-

paign, J. Data and Information Quality 10 (2018).
[39] F. Hopfgartner, K. Balog, A. Lommatzsch, L. Kelly, B. Kille, A. Schuth, M. Larson, Continuous evaluation of large-scale information access systems: A case for living labs, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF, Springer, 2019.
[40] J. Schatzmann, K. Weilhammer, M. Stuttle, S. Young, A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies, Knowl. Eng. Rev. 21 (2006) 97–126.
[41] O. Pietquin, H. Hastie, A survey on metrics for the evaluation of user simulations, Knowl. Eng. Rev. 28 (2013).
[42] S. Vakulenko, K. Revoredo, C. Di Ciccio, M. de Rijke, QRFA: A data-driven model of information-seeking dialogues, in: Proceedings of the 41st European Conference on IR Research, ECIR '19, 2019, pp. 541–557.
[43] J. R. Trippas, D. Spina, L. Cavedon, M. Sanderson, 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, 2017, pp. 325–328.
[44] K. Balog, L. Flekova, M. Hagen, R. Jones, M. Potthast, F. Radlinski, M. Sanderson, S. Vakulenko, H. Zamani, Common conversational community prototype: Scholarly conversational assistant, 2020. arXiv:2001.06910.
[45] K. Balog, T. Kenter, Personal knowledge graphs: A research agenda, in: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR '19, 2019, pp. 217–220.
[46] S. Zhang, K. Balog, Evaluating conversational recommender systems via user simulation, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20, 2020, pp. 1512–1520.
[47] L. Azzopardi, M. Dubiel, M. Halvey, J. Dalton, Conceptualizing agent-human interactions during the conversational search process, in: Proceedings of the 2nd International Workshop on Conversational Approaches to Information Retrieval, CAIR '18, 2018.
[48] J. F. Kelley, An iterative design methodology for user-friendly natural language office information applications, ACM Trans. Inf. Syst. 2 (1984) 26–41.
[49] C. Lewis, J. Rieman, Task-centered User Interface Design: A Practical Introduction, University of Colorado, Boulder, Department of Computer Science, 1993.
[50] J. Edlund, J. Gustafson, M. Heldner, A. Hjalmarsson, Towards human-like spoken dialogue systems, Speech Commun. 50 (2008) 630–645.
[51] B. R. Cowan, N. Pantidi, D. Coyle, K. Morrissey, P. Clarke, S. Al-Shehri, D. Earley, N. Bandeira, "What can i help you with?": Infrequent users' experiences of intelligent personal assistants, in: Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI '17, 2017.
[52] B. R. Cowan, H. P. Branigan, H. Begum, L. McKenna, É. Székely, They know as much as we do: Knowledge estimation and partner modelling of artificial partners, in: Proceedings of the 39th Annual Meeting of the Cognitive Science Society, CogSci '17, 2017.
[53] A. Sciuto, A. Saini, J. Forlizzi, J. I. Hong, "Hey Alexa, what's up?": A mixed-methods studies of in-home conversational agent usage, in: Proceedings of the 2018 Designing Interactive Systems Conference, DIS '18, 2018, pp. 857–868.
[54] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, D. Yu, et al., Using recurrent neural networks for slot filling in spoken language understanding, IEEE/ACM Trans. Audio, Speech and Lang. Proc. 23 (2015) 530–539.
[55] B. Liu, I. Lane, Attention-based recurrent neural network models for joint intent detection and slot filling, in: Interspeech 2016, 2016, pp. 685–689.
[56] J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, S. Young, Agenda-based user simulation for bootstrapping a POMDP dialogue system, in: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers, 2007, pp. 149–152.
[57] Z. Yang, Z. Hu, C. Dyer, E. P. Xing, T. Berg-Kirkpatrick, Unsupervised text style transfer using language models as discriminators, in: Advances in Neural Information Processing Systems 31, NIPS '18, 2018, pp. 7287–7298.
[58] Z. Fu, X. Tan, N. Peng, D. Zhao, R. Yan, Style transfer in text: Exploration and evaluation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
[59] T. Shen, T. Lei, R. Barzilay, T. Jaakkola, Style transfer from non-parallel text by cross-alignment, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS '17, 2017, pp. 6833–6844.
[60] T. Wilson, Information needs and uses: Fifty years of progress, Fifty Years of Information Progress: A Journal of Documentation Review 28 (1994) 15–51.
[61] D. Ellis, The dilemma of measurement in information retrieval research, J. Am. Soc. Inf. Sci. 47 (1996) 23–36.
[62] E. M. Voorhees, D. K. Harman, TREC: Experiment

and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing), The MIT Press, 2005.

[63] N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF, volume 41 of *The Information Retrieval Series*, Springer, 2019.

[64] D. Kelly, Methods for evaluating interactive information retrieval systems with users, Found. Trends Inf. Retr. 3 (2009) 1–224.