A Theoretical Framework for Conversational Search

Filip Radlinski*
filiprad@microsoft.com
Microsoft, United Kingdom

Nick Craswell
nickcr@microsoft.com
Microsoft, United States

ABSTRACT

This paper studies conversational approaches to information retrieval, presenting a theory and model of information interaction in a chat setting. In particular, we consider the question of what properties would be desirable for a conversational information retrieval system so that the system can allow users to answer a variety of information needs in a natural and efficient manner. We study past work on human conversations, and propose a small set of properties that taken together could measure the extent to which a system is conversational. Following this, we present a theoretical model of a conversational system that implements the properties. We describe how this system could be implemented, making the action space of an conversational search agent explicit. Our analysis of this model shows that while theoretical, the model could be practically implemented to satisfy the desirable properties presented. In doing so, we show that the properties are also feasible.

Keywords: Conversational Search, Chatbot, Personal Agent

1. INTRODUCTION

Recent progress in Machine Learning has brought tremendous improvements in natural language dialogs between humans and conversational agents. This has led to a plethora of commercial conversational agents (also called *chat bots* or simply *bots*) that are able to answer user requests from ordering pizza to suggesting holiday destinations. Such systems are *conversational* in that they assist users using a dialog interaction, be it in written or spoken form, usually with a rich human-like vocabulary.

To build an information retrieval system with a conversational user interface, it is useful to define a computational model that describes the process of conversational search. The model should allow the user to make a natural language request, akin to a traditional information retrieval query. It should allow the system to propose search results, but also ask the user for clarification if necessary. It should allow the user to give feedback on the system's results and suggestions, including negative feedback. Over time, the process should allow the system to build a cumulative picture of the user's information need based on their query statements and other relevance feedback.

We observe that conversational search is in keeping with trends in the design of computing devices and interfaces [15]. Modern devices with small or no screen may provide responses via small on-screen cards and speech synthesis, so succinct conversational responses are appropriate. With speech recognition accuracy also improving due to progress in machine learning, the popularity of speech-based search input is also growing[1]. Such a growth in natural language dialog between users and search systems may even lead to the dominant interaction model of one-shot keyword queries being displaced with conversational systems.

To build a computational model for conversational search it is important to define which steps are allowable during a conversation: The types of statements that can be made by the system and by the user. The system must build a model of the user's information need over the course of the conversation, such that he or she does not need to repeat important aspects of the information need. Cumulative clarifications should tend to move the process closer to success. To make the conversation more flexible and natural, ideally most conversational steps that humans would take should be interpretable by the system and also potentially generated by the system. For example, in a real conversation about restaurants a person might ask "*Do you like sushi? I went to a great place yesterday*" or perhaps ask key questions such as "*Are you looking for somewhere fancy?*" In certain contexts the question may not directly appear to be about food, such as "*How much time do you have?*" People take into account what they know about the other person from past conversations and other aspects of context, and even ask for direct feedback "*What did you think of the restaurant that I suggested to you last week?*"

Many conversational search tasks are similar: People offer a reference point, or a key choice, to elicit the information they consider most important for separating the sorts of places that they might recommend. In the restaurant domain, people very rarely enumerate the types of cuisines or ask for a specific limit on the number of miles you are willing to travel. The same applies in other domains, such as when searching personal photo collections – a particular photo may serve as a reference point from which the target may for instance be earlier/later, in a different location, but with the same people.

In the field of spoken dialog systems, approaches already exist allowing conversational slot filling of a structured query within a schema (e.g. [44]). This allows users to book a ticket for a certain concert on a certain night, or set a certain reminder message to

*Now at Google UK.

[1]http://www.prnewswire.com/news-releases/teens-use-voice-search-most-even-in-bathroom-googles-mobile-voice-study-finds-279106351.html

CHIIR '17, March 07 - 11, 2017, Oslo, Norway

ISBN 978-1-4503-4677-1/17/03…$15.00

DOI: http://dx.doi.org/10.1145/3020165.3020183

appear on their mobile device at a certain time. By contrast, in other conversational search scenarios there may not be a fixed schema or the underlying data may not be structured as a database. In those cases rather than slot filling it may be beneficial to use a more free-form conversation that nevertheless builds its understanding over the user's needs over multiple rounds of conversation and may provide responses as well as ask clarifying questions.

We hypothesize that two aspects of conversations are particularly pertinent to search settings. First, users often do not know how to describe their information need – be it for a recommendation, or information regarding a new topic. Part of the role of the conversation is to elicit the actual need from the user by helping them formulate it clearly [24]. Second, for many tasks particularly suited to multi-turn conversational interactions, a *set* of results interact to produce a single item response which satisfies the need. For instance, when selecting a product to purchase, it is often driven by a preference among available options [13]. On the other hand, in some settings the solution is by its very nature a set, as in when deciding upon a holiday destination which requires travel, accommodation and dining requirements to be satisfied [21].

This paper presents a formalism that allows for richer sharing of initiative, and longer-term adaptation and personalization. Our goal is to capture the desirable properties of conversation specifically from an *information retrieval* perspective. Intuitively, the setting is that a conversational agent (assistant) has been asked satisfy an information need. At each point in time, the agent can perform one of a fixed set of actions, to which the user responds, with a back-and-forth. Mixed initiative and memory are key parts of this model. Our key contributions are

1. We suggest a formal definition of a conversation from an information retrieval perspective, showing why each property is desirable.
2. We propose a theoretical model of conversations that allows agents to satisfy the formal properties, demonstrating that the definition is also practical.

2. RELATED WORK

Human conversations have been studied for decades, and conversational research can be understood in relation to a number of existing research areas. We present an overview of the space to place our work in context. We start with conversations as applied to information retrieval, then discuss the broader literature regarding the essence of conversations.

2.1 Search and Recommender Systems

In traditional ad hoc search, the system allows the user to provide a natural language request (query) describing results that they want. A minimal response from the system is a ranked list of results (akin to "try looking at these") and a search box that remains available (the system allowing the user to "tell me what you want"). Systems may also correct the user's query ("did you mean?"), suggest other queries ("try these related searches") or provide faceted browsing in the results ("refine by") [43]. In that sense the user is having a conversation with a system that is providing a variety of responses, and many results at once are bundled into a single search engine results page.

During such a conversation, even if the user's search task does not change their understanding of the task, their query vocabulary may change and they may apply a variety of search strategies [2]. In some systems the retrieval response is based only on the user's most recent query, but other systems can take into account past queries and other context [4, 22]. In that sense modern contextual information retrieval systems already allow some co-active development, where both the system and the human user develop their understanding over time. However, real conversations may have mixed initiative [37], where control of the conversation passes from one side to the other via assertions, commands, questions and prompts. For instance, Dredze et al. [14] showed how in the context of email search an agent may propose pertinent ways to select subsets of the result set by adding key-value pairs to the query.

Already over two decades ago, Belkin et al. [3] considered conversational information retrieval by characterizing information-seeking strategies. They proposed scripts that can be followed by a system for different types of retrieval tasks, using case-based reasoning to select next steps and offer users choices. This differs from our work, as we assume a simpler conversational interface (such as a chat) where users enter text in response to agent actions also consisting of simple statements. The users may or may not respond directly to the system's requests. Also, we model the retrieval problem as one where the system reasons about items that can be retrieved, rather than over the space of possible user intents.

Conversational agents for more advanced multi-turn tasks have been proposed continuously since then, for instance recent work to isolate and resolve technical issues typically handled by a help desk [42]. This differs from our work, as we address the task of information retrieval rather than guiding a process by which a problem may be resolved. While a related informational goal could be to identify an instruction document, our goal is a characterization of a more general class of conversations.

In a spoken conversation or on a device with a small screen, it also becomes important for the search system to chose one response or a small number of responses, rather than bundling a large number of results and suggestions into a results page. For example, if the user's query was ambiguous it may be optimal to show search results for just one intent and query suggestions for another [19]. If we consider the query suggestion to be a clarifying question, then showing such a suggestion prominently allows for a greater reward later by incurring an initially costly question. The idea of reinforcement learning, to optimally plan for a delayed reward rather than greedily always choosing the maximum immediate reward, is also explored under the card model of information retrieval [45]. These are important steps towards a mixed initiative conversational search system, although still with traditional system responses such as results and query suggestions.

Methods for conversational recommendation have also been proposed. Recently, Christakopoulou et al. [9] studied whether to ask absolute or relative questions, comparing the utility of each for learning about users. They also asked questions contextually, based on what is already known. Much earlier, Linden et al. [21] proposed a conversational travel agent that allows the user to find an optimal or near-optimal trip by presenting the user with examples that characterize the solution space and allowing the user to express and modify their criteria. A key method for expressing such updates is critiquing, which gives feedback on facets of importance – such as airline, price or departure time of a flight – with respect to the options already presented. In general a critique can be directed at a particular attribute of a particular item, for example "like this one but cheaper" [23]. Another form of critiquing is at the item level, dating to the Rocchio relevance feedback algorithm where users of a search system may annotate results as relevant or not to refine the search query [29]. A critique differs in that it explains how a result could be modified to improve its utility to the user.

We note that in human conversation critiquing also happens, but it is not limited to a pre-defined set of facets. An ideal conversational information retrieval system might allow free form critiquing of suggested results in natural language. To enable free-form queries and critiques, the information retrieval system could build its models based on the language modeling approach to information retrieval. However, more advanced forms of reasoning may be required, particularly when the user answers a question or refers to other parts of the conversation, suggesting the use of more sophisticated natural language technology. Despite this, today most end-to-end conversational systems based on deep neural networks lack the ability to explicitly focus on a search task, rather giving generic contextual responses (for example, [31]). Memory networks have proved very effective at complex question answering scenarios, able to provide correct answers given complex pieces of information and potentially a large knowledge base [33]. However, they are unable to request clarification of the task at hand when the solution is uncertain.

2.2 Spoken Dialog Systems

Spoken dialog system research enables a flexible conversation to take place including corrections and clarifications, usually in a closed domain such as setting a reminder or booking tickets (for instance, [44]). In an early system, Paek and Horvitz modeled spoken conversations using a Bayesian network that tracks confidence from the level of the audio signal obtained from the user through to predicting the user's goal with appropriate back-off depending on detected failure modes [26]. More recently, Chen et al. [8] have studied how a system can estimate the user's intent within a particular conversation step. In a simpler task, the Dialog State Tracking Challenge has pushed forward the ability of systems to fill known slots for the task of bus travel planning (e.g. [16]). Yet such a slot-driven approach differs from human recommendation where it is rarely important to fill all slots [9].

Co-reference resolution can successfully track references to entities across spoken queries[2], yet back references to preferences expressed in a search scenario have not been explored to the best of our knowledge. Further, these systems do not involve mixed initiative, with the system simply keeping up with the human. Even in closed domain dialog systems, additional work is needed to make the turn-taking behavior of the system more flexible and efficient [27]. In a more open domain, Jiang et al. recently studied the most popular commercial personal agent systems capable of multi-turn task solving. They identified a set of actions that agents tend to perform, albeit at a high level [17]. Such an agent performs a mix of slot filling and information tasks, although in many cases for an information retrieval task it resorts to a traditional search engine results page.

2.3 The Human Perspective

Finally we turn to work studying conversations as performed by people. Perhaps among the most famous attempts to replicate conversations, Eliza was one of the first chat bots, replying to user statements consistently with how a therapist may engage a patient [39]. The algorithm rephrased statements made by the patient, reformulating them as questions back to the patient. More recently, deep learning systems have attempted to build contextualized chit chat systems, for instance as a Twitter bot that responds to context [31]. We consider what roles conversations per se appear to play as part of information exchange.

Conversations as Revealment

One significant role of conversation from an information retrieval perspective is to allow the two parties to reach an understanding as to what is required by the user, and what the answerer knows. Before Web search became prevalent, as much information retrieval occurred in libraries, it was noted that the role of librarians was to help users to express their information needs. In particular, [24] studied how librarians assist in this task. The author found that the method of the librarians wasn't as important as that the conversation was happening. This suggests that automated conversational systems may also be effective even if using very different techniques.

Initiative and Engaging Behavior

A number of authors have studied how a "virtual human" should behave [6, 38]. For instance Traum et al. describe desirable aspects of a system conversing with humans, such as being real-time and incremental as utterances are formed over time [36]. Similarly there has been extensive work on multi-modal systems, expressing emotion and so forth. These aspects are beyond the scope of our work, as we restrict ourselves to chat type interfaces. Additionally, our focus is to consider conditions on *what* needs to be possible to be said rather than *how* the information should be conveyed.

One of the key aspects of human conversations is initiative. A number of authors have considered what constitutes initiative in dialog systems [1, 10, 25]. Of key interest to us is *mixed initiative*: At different times in the conversation, the human or the agent may take initiative. We use a generic definition of mixed initiative compared to past work, defining it as both the human and the system having initiative at different points in time. For instance, the agent may take initiative to clarify or elicit information from the user whenever appropriate, while allowing the user to drive the conversation at other times.

Trust and Moral Character

A final important concept in agents emulating human behavior is one of *moral character* [18]. Any agent taking part in a conversation conveys a personality, and inherently builds a relationship with the user (for instance, trust with regards to what happens to information shared by the person with the agent). However, this aspect of conversational behavior is outside the scope of our work and is not a goal of the conversational model.

It is also the case that when provided information (such as advice or recommendations), the source matters to people – it has been established that different sources have different influence on purchase decisions [30]. Effectiveness of a conversational system would likely depend on the system saying *why* it made a specific recommendations [35]. As with moral character, we do not address this question in this work.

3. CONVERSATIONS FOR SEARCH

In this section we consider the properties of conversations, proposing aspects that are applicable to search.

3.1 What is a Conversation?

The Oxford English Dictionary defines a conversation as *a talk, especially an informal one, between two or more people, in which news and ideas are exchanged.* While broad, this provides some guidance in information retrieval settings. In particular, we note that information is *exchanged*, suggesting symmetry where initiative may belong to both sides at different point in the conversation (rather than say a lecture). Hence we postulate that a conversational search system is a mixed-initiative system.

We may also classify conversations by their outcomes. Often, a conversation may be an end in itself. We do not consider this type of conversation here as it does not involve information retrieval. Similarly, conversations may have as a goal to assist a person to follow a known sequence of steps. Once more, this type of conversation falls

[2]http://searchengineland.com/googles-impressive-conversational-search-goes-live-on-chrome-160445

beyond the scope of this work. We focus on conversations that aim to elicit user preferences, and identify target information.

As a third aspect, we postulate that there is an element of memory: The conversation is a unit, and earlier statements can be referenced later in the conversation. Indeed, it should be possible to reference earlier statements in earlier conversations. A first consequence of the ability to index earlier statements is the existence of repair mechanisms, for instance the ability to clarify with *"what I meant is..."* [34, chapter 7]. More importantly in a search setting, memory allows information to be elicited from the user in a piecemeal fashion, maintaining simple steps that can together describe an arbitrarily complex information need. Indeed, it has been shown that loss of context is a common reason for user frustration with conversational systems [20]. It is important to note that memory thus plays two roles:

1. The systems remembers what was previously said by the user or the system to assist in resolving the user's information need.
2. It is possible for the user to explicitly reference past information, for example to indicate what statements are not correct or should be "forgotten".

Finally, the conversation should be adaptive, with neither participant following a script, but rather adapting to the current context. This expands upon common definitions of personalization, while avoiding the challenge of sessions. In particular, a conversational search agent is essentially fulfilling a long-term task, which may otherwise have consisted of many sessions in the traditional search engine sense. The abstraction will prove useful below.

Taken together, these properties lead to the following definition:

DEFINITION 1. *A* **conversational search system** *is a system for retrieving information that permits a mixed-initiative back and forth between a user and agent, where the agent's actions are chosen in response to a model of current user needs within the current conversation, using both short- and long-term knowledge of the user.*

Further, the system has the following five properties, which we term the RRIMS properties:

User **R**evealment	The system helps the user express (potentially discover) their true information need, and possibly also long-term preferences.
System **R**evealment	The system reveals to the user its capabilities and corpus, building the user's expectations of what it can and cannot do.
Mixed **I**nitiative	The system and user both can take initiative as appropriate.
Memory	The user can reference past statements, which implicitly also remain true unless contradicted.
Set Retrieval	The system can reason about the utility of sets of complementary items.

3.2 When should search involve conversation?

The appropriateness of a conversation for a search task is largely driven by the complexity of the task. The simplest search settings, where the user enters a single query and they expect to immediately identify relevant results clearly does not call for a back-and-forth with a search agent.

The next more complex type of tasks require *memoryless refinement*: The user learns the right terms to describe their information need by iterating with a search system. If each step is only informed by the results from the previous iteration, this does not require memory nor agent initiative. In such a setting a more complex model may in fact reduce user utility and does not call for conversational approaches to search.

However, consider these more complex scenarios where a conversation is more likely to be appropriate:

Faceted Elicitation

The user is searching for an item with rich attributes that can be individually specified, but are much simpler to provide piecewise. For instance, it may be possible to describe a difficult to find email such as *I'm looking for an email that contains a link to a research paper that I got from a student who emailed me right after SIGIR last year. I can't remember the student's name, but I had never heard from her before.*

The user is selecting among items based on facets – but cannot be expected to know how to reference these directly as this would involve memorizing a complex query language. As part of the search, the user is identifying aspects that can be used to describe a relevant item. In contrast to memoryless refinement, here the user may need to learn about a facet before returning to the top level of the search process with a tag describing how this facet can be satisfied. For instance, consider a similar case where the user is selecting a vacuum cleaner to purchase. Here, as an aside, the user may need to learn about relevant attributes such as how loud a given number of decibels really is, and then returning to his main task.

Multi-Item Elicitation

The user is searching for a single item supported by a set of nearby items. For instance, a photo which can only be described by the properties of other photos taken earlier such as *the photo Alice took of me right after I took her picture a few months ago*. In this case, the system may need to learn who *Alice* is.

While the search is for an item that has an easy to establish relevance, the user's only known description of this item (i.e. query) depends on other items, which may themselves need to be found. Then, the search system must estimate the relevance of the whole set of items.

Multi-Item Faceted Elicitation

In this setting, the user is searching for a set of items directly. Importantly, not only must the system estimate the utility of each single item, it must combine the utilities of multiple items to reach an assessment of an entire set.

For instance, planning a vacation where the results consist of a hotel, travel arrangements, restaurant plans, places to see, and so forth. During the conversation, the agent elicits users to describe relevant aspects of different destinations, hotels, transport options and attractions. Then, it must elicit information from the user to learn how to combine the utilities of a whole set of items to reach a final decision about a holiday as a package.

It is this last setting in which we hypothesize that conversational approaches to search have the highest usefulness.

Bounding Choices / Building Expectations

Simultaneously, a conversational interface may simplify the problem of need elicitation by providing users with *bounded choices*. It may be easier for a user to clarify their needs given precise choices rather than expecting them to come up with particular terms. Similarly, choices can be bounded by allowing the user to understand complex features available in a search system, as examples of how a need can be presented are given.

For instance users engage with facets for email search much more often when these are suggested contextually, rather than relying on the user to generate the relevant terms [14]. Similarly, expert search users are much more likely to use advanced operators [40], presumably as less expert users are unaware of the options available.

This concept of bounding choices can also be considered as revealment from the side of the system, showing the user examples of the possibilities the system offers.

3.3 Learning, curiosity and serendipity

During any search interaction, the agent may acquire knowledge that is useful to answer the user's current information need, while also building a model to improve personalization in future. For instance, an example in the previous section required the agent to learn to identify Alice in a query. This knowledge would allow the agent to answer future queries that refer to Alice without requiring the label to be provided anew.

It is worth noting that there are also cases where the agent may provide the user with long-term utility at a cost in the current query, perhaps by eliciting information that happens to be related but not directly relevant. For instance, consider a user who searches for a restaurant recommendation and specifies that it should be vegetarian. The agent may clarify if the requirement is simply for the current query, or indicates that the user always requires vegetarian restaurants. Similarly, in a photo search scenario, the agent may elicit a name that can be applied to a face common in candidate photos.

We do not exclude such scenarios from expectations of a conversational agents, although such actions return to issues of trust and moral character, and thus further treatment of them are beyond the scope of the current work.

4. CONVERSATIONAL SEARCH MODEL

For our model, we assume there is a user searching for information, and a system or agent that is assisting the user. Search is performed over a well-defined corpus C, where the user it looking for an item $i \in C$ (which may not be unique) that contains the information needed. Such needed items are said to be of high utility to the user. Within a conversation, the system must estimate the utility of each item, which we write u_i. We note that the agent may have a prior estimate of utility over items $\hat{u}_i$ before the user has specified anything, based on long-term knowledge, although do not further consider how this prior utility is maintained.

4.1 Interaction Approaches

In each back and forth step in a conversation, the system provides some information to the user, and the user responds. Depending on what is provided, and what the response is, we find ourselves in different conversation settings seen in prior work. Existing approaches of which we are aware are summarized in Figure 1.

From the system perspective, our model provides for three basic types of information that the system may provide to the user. We term these *actions* that the system may perform. In increasing specificity, these are *nothing*, *a partial item* and *a complete item*. In particular, an item may be partially described in many ways. In the simplest case, the system may select a specific item feature to focus on, such as the concept of `price` in a product search scenario. Alternatively, the system may provide a suggested value of each field, e.g. `price between $50 and $100`. Finally, the system may present a cluster of items, for instance a grouping of products that are somehow similar; this can be through of as a dynamic field having been created for say "electronic gadgets that make good gifts for a teenager".

Conversely, the user may be expected to provide (equivalently, the system may understand user statements that provide), feedback of different types. The simplest design would be for the user to provide either a binary or ordinal score in response to a question, or a preference given two or more choices. A more sophisticated feedback from the user would be a critique [23, 28] that indicates in what way the item or partial item presented by the system does not represent the user's information need. The most detailed level of feedback a user may provide would be free text. Clearly, the meaning of the user's feedback is only well defined given a specific question.

Considering previous work, we note that each prior system typically falls into a single cell as indicated in the Figure. We now describe each of the labeled cells in turn. This will describe the basis of the richer action space model we propose in this paper.

Null System - Free Text User

This is the starting point for most information retrieval systems such as Web search engines, and often for conversational systems where the user may specify many possible requests (such as commercial intelligent agents including Cortana and Siri). The user is simply presented with a search box into which any query can be entered.

Partial Item System - Pref/Rating User

A user may be presented with partial information about matching items in various ways. The most common approach is for a conversational system to confirm a slot that has been inferred, such as "you are looking for an Italian restaurant, correct?" (see, for example [17]). Some systems may also cluster items, asking for a preference. For instance, it might ask "would you prefer to a fancy restaurant, or an inexpensive one?". A third interaction mode, where a preference is elicited over a set of (feature,value) pairs would for instance "would you prefer a laptop with a 12 inch screen for $1000, or a laptop with a 14 inch screen for $1200". Note that all of these interaction modes – as well as critiques and free text entry discussed below – may also be considered "faceted search".

Partial Item System - Critique User

When the user may provide a richer answer than a simple score or preference, this presents a more powerful information retrieval paradigm. In the simplest case, fielded search provides users with a selection of known fields and users may select or specify ranges for any property they desire. This is common in online shopping scenarios, where often the allowed field values are pre-specified. In other settings a user is presented with specific individual facet values. Some commercial intelligent agents allow users to clarify in this way, rather requiring a simple yes/no. For instance, in response to a prompt "you are looking for an Italian restaurant, correct?", the user may reply "no, I'm looking for an Indian restaurant."

Partial Item System - Free Text User

When a system asks a user to fill in a particular aspect of an information need, this is usually referred to as slot filling. For instance, many recommendation systems work in this way. As an example, systems taking part in the Dialog State Tracking Challenge [41] require users to specify travel details to complete a structured query over a public transit schedule.

Complete Item System - Pref/Rating User

Classic approaches to recommendation often request ratings of items to learn a user model for further recommendations. These may be absolute rating requests ("how much did you enjoy the movie Kill Bill?") or preference requests ("did you enjoy Kill Bill or Pride and

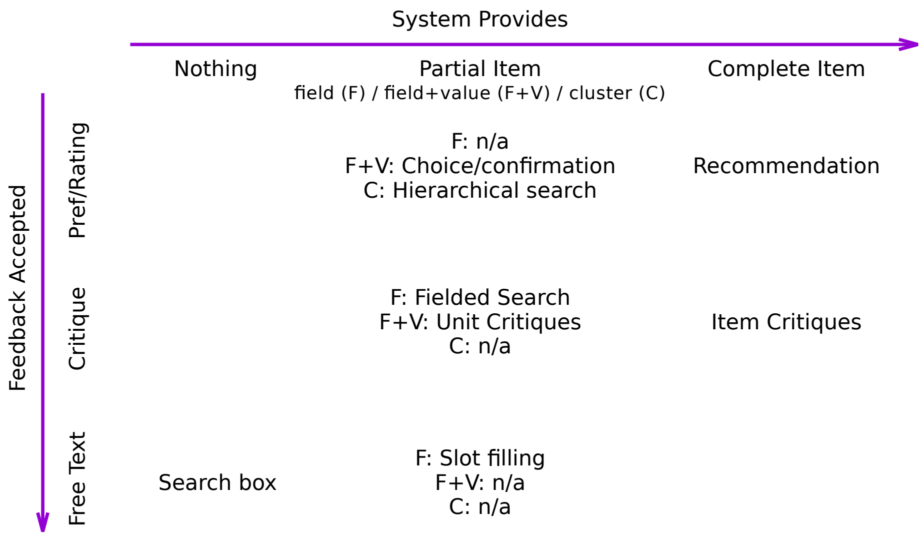

Figure 1: Conversation action space, as matched to previous names from past work. The system may provide three types of feedback, and expect three types of responses in return. In each cell, we describe related work that falls into the appropriate category. We also note that many of the partial item field (F) or field+value (F+V) interaction approaches are often considered variants of faceted search.

Prejudice more?"). For example, Christakopoulou et al. [9] describe such a system in the restaurant domain.

Complete Item System - Critique User

In this case, a system may select a given item, then allow the user to refine their information need anchoring of the properties of the item. For instance, Reilly et al. [28] describe a system where users are presented with an item and possible ways the information need can be refined. Users may select a pre-defined rich critique that allows the system to move closer to the user's goals. Su et al. [32] describe a more sophisticated restaurant recommendation system, where queries are matched to a complete item, which can then be refined or further metadata can be requested.

We also propose an extension to the critique model to allow the agent to learn about the collection directly from user feedback. For example, suppose the agent has been asked to recommend a movie. Given a movie, the user might respond with a critique "that movie is too gory". This can inform the agent about the existence of the concept "gory", which may not have been known to the agent previously. Once given a name and one example, the agent may learn to model it through further interactions with this and other users.

4.2 Interaction Choice

Based on the above variety of existing user/system interaction models, we can now formally model a conversation as a back-and-forth, where the user and agent take turns. For convenience, the conversation always starts with the agent. Each time it is the agent's turn, it will (1) select an action to perform and (2) request for the user to provide a specific type of response. Specifically, the actions available to the agent are:

- $a_\emptyset$ The null action – provide nothing, user is requested for free text describing the information need.
- a_p^1 System provides a single partial item/cluster. User is requested to provide a rating, critique, or free text.
- a_p^{2+} System provides two or a small number of partial items, requesting a preference, critique, or free text.
- a_i^1 System provides a single complete item. User is requested to provide a rating, critique, or free text.
- a_i^{2+} System provides two or a small number of complete items. User is requested to provide a preference, critique, or free text.

The user responses are of the following types:

- r_r A rating of the current item/partial item.
- r_p A preference among the presented items/partial items.
- r_{np} A lack of preference, either $\emptyset$ indicating that none of the option is suitable, or $\star$ indicating that all options are equally suitable.
- r_c A critique of the current item/partial item, indicating how the current item/partial item could be modified to be of higher relevance to the user.
- r_t Unstructured text describing their information need.

4.3 Action Selection

As described at the start of Section 4, the system maintains a distribution over utility values u_i for each item $i \in C$. The goal in an information retrieval setting is to find an item with maximal utility. Thus, a conversational search algorithm must select actions to maximize user satisfaction while tracking expectations over user responses. The motivation behind the above model is that the choice available to the system is simple enough that the utility of each action and response request can be estimated, yet provides the richness necessary for a true conversational retrieval system.

Specific satisfaction metrics that can be optimized are beyond the scope of this work, given the large amount of past work on this topic (see recent work by Kiseleva et al. [20] for an overview). However, we will describe an example algorithm for this selection process in the analysis of our model below.

5. ANALYSIS

We finally present an analysis of the conversational model. In particular we assess four natural analysis questions: Are the criteria suggested for conversations both necessary and sufficient? Are the system action and user response spaces sufficiently rich? How can the system select among the possible actions? How can the system correctly interpret user feedback in this rich environment? We discuss each in turn.

5.1 Are the conversational properties presented both necessary and sufficient?

The goal of the conversational model presented in this work is to allow efficient and effective conversational information retrieval without more complexity than necessary. Thus the first research question we must address considers what kinds of tasks and can be solved with the properties presented. While showing that a wide variety of previously studied information retrieval tasks can be addressed, we argue that each of the properties presented is also necessary.

Example application 1: Basic information retrieval.

We start by considering a standard information retrieval task. We take the first topic from the most recent TREC Web Track [11]:

```
Topic ID: 251
Query:  Identifying spider bites
Description:  Find data on how to identify
spider bites.
```

While the task description is a short query, it suggests the user may need to identify a *specific* spider bite. As the variety of spiders is large, this need would likely be satisfied best in a conversation where the user and the system exchange information to assist the user in narrowing down the candidate set of all possible spiders into the most likely one(s). During this process, the system needs to remember what information the user has already provided, to allow the user to answer individual questions one at a time. It may also be the case that the user needs to revisit or alter answers by referring to past statements. For instance, the user may incorrectly answer one of the questions and not realize until later in the conversation.

Being able to satisfy such needs illustrates the importance of user revealment, as well as memory in the conversational search setting.

Example application 2: Personal information search.

We consider a second information retrieval setting with increasing research attention – search over personal information. This may involve searching for personal emails (e.g. [5]), documents (e.g. [12]) or over personal records that allow a user to investigate for instance where a particular event took place [7]. Such search tasks involve heterogeneous items with rich metadata, where a user may remember a variety of contextual information. An effective conversational retrieval system must aid the user in specifying such information, without requiring the user to remember a complex query language.

In this setting, mixed initiative is particularly valuable, where a system may prompt the user with information that the user may remember. This in turn may lead the user to recall other pertinent information the he or she wishes to provide. Traditional search interfaces for personal information of this nature tend to present rich user interfaces [5, 7, 12]. A mixed initiative system can present users with choices when appropriate to refine the search space, yet allow the user to describe their information need in free text when this is the optimal strategy. Additionally, the value of system revealment and memory is clear.

Example application 3: Product recommendation.

A common information retrieval task is product selection given a general information need. For instance, consider a new parent who must purchase a stroller for the first time. The parent may be unaware of the qualities of such a product without having previously engaged in this task. In an offline recommendation setting, the parent visits a retailer, and an assistant will describe the range of products available, how they differ and elicit the information from the parent as to which features are important, and ultimately guide the parent to suitable choices.

In a conversational search setting, the agent must similarly reveal to the user characteristics of the available search space, knowledge of which features exist in the corpus of items, and assist the user in expressing their information need suitably. Thus this example illustrates particularly the necessity for system and user revealment.

Example application 4: Travel planning.

In some settings, search involves heterogeneous items that give rise to two distinct types of user utility: A given item has a utility to the user, while a set of items is needed to answer the user's information yet has a different type of utility.

One common example is travel planning (see for example [21]), where travel and accommodation are both necessary yet each have their own utilities (cost, convenience, brand, etc), while the combination is the ultimate user's need and has its own utility to the user. For example, even an otherwise perfect inexpensive five star hotel is likely to have low utility for a cost-conscious traveler if it can only be reached by private helicopter on the intended travel date. Other distinct aspect to travel plans – including attractions and restaurants further complicate the retrieval setting, giving rise to distinct item-based and set utility functions.

While in the previous example total cost might be considered a strong indicator of utility for both an item and a set of items, in related settings these utilities may be quite different. For instance, consider an itinerary within a city: The user may have preferences about what types of attractions they prefer to visit, giving rise to an item utility. However for a complete itinerary the user may also value diversity, so as not to spend the entire day visiting only attractions of one type.

Summary.

As we have seen, each of the properties presented – User Revealment, System Revealement, Mixed Initiative, Memory and Set Retrieval – are natural for at least one of the example applications. These applications typify the settings in which information retrieval systems are commonly used, and this in which conversational information retrieval should be possible. As such, we argue that the proposed properties are both necessary, as well as being sufficient to enable many common information retrieval tasks.

5.2 Are the system and user action spaces sufficiently rich?

After presenting the desirable properties for a conversational IR system, we presented a model in Section 4. We now analyze to what extent this model satisfies the desired properties.

User Revealment.

During search, a common strategy to assist users to refine their information needs is to present alternatives within the space of extant items. As such, the action spaces a_p^{2+} and a_i^{2+} provide efficient ways for the user to identify alternative items, as well as dimensions on which the items differ, and ambiguities within the information need described so far.

For instance, when searching for a suitable product in a class where the user is unfamiliar (say, the first time a new parent must purchase a stroller), choices help the user reveal the relevant features to the user.

System Revealment.

Free-form text entry systems are known to have low discoverability (e.g. [40]). By presenting users with confirmations (a_p^1) and requesting a rating (r_r) or critique (r_c) as well as partial item choices a_p^{2+}, the system both demonstrates the ways in which it can partition and refine the search space, as well as common properties of the corpus available.

Set Retrieval.

In modeling partial item presentation as clusters, the model allows for retrieval of sets of items. For instance, taking a travel set retrieval scenario, the system may assist the user in identifying high utility items of various types, then subsequently present sets of complementary items as candidate solutions to the user's information need (for instance, inviting critiques of proposed combinations). We note that as seen in Figure 1, to the best of our knowledge such structured set-based retrieval has not been previously studied in a conversational setting, rather relying on a rich user interface.

Memory.

As previously noted, we postulated that memory plays two distinct roles in a conversational search setting: (1) the system recalling past statements by default, and (2) the ability to reference explicitly to past statements (for instance to indicate that they are no longer correct). The first is addressed in the model implicitly, as a conversation is designed to be a continuous process, thus a continuation of a conversation implicitly requires the continuation of a user's information need and thus all earlier conversational steps.

The second is addressed in the requirement that the model allows the user to always enter free text (r_t). While a weaker constraint on the implementation, this possibility – assuming the user text is interpreted correctly by the system – allows the user to refer to a previous statement to override it specifically. However, the details of how this could be implemented is left as future work.

Mixed Initiative.

In providing the system a number of alternative interaction modes, from a basic free-text ($a_\emptyset$) to structured and preference-based (a_p, a_i), the system is designed to choose the right level of initiative for an information task. In allowing the user to always return unstructured text (r_t), the user can at any time take the initiative from the system.

5.3 Could a system optimally select the next action from the search space?

Given the choices of actions, the system must have a model allowing it to decide which action it is to take at any given point in time. It must (1) select an action in a given context, and (2) interpret the user's response given the previous conversational history. We focus on the action selection process here, arguing that a system could reasonably implement the model presented while leaving the practical details as an open challenge for future work.

In our model, the system maintains a distribution over utility values u_i for each item $i \in C$. As the goal in an information retrieval setting is to find an item with maximal utility, the action must be selected so as to maximize how efficiently this is achieved. While a number of algorithms may be optimal with different implementations of the model, here we present one possible implementation to demonstrate feasibility. Other approaches may be more efficient or lead to better user experiences. One of the goals of this paper is to inspire such approaches, hence we leave them as future work. Rather, we argue here that our model is suitable for describing a conversational information retrieval system.

For each action the system may take, if the user response comes from a known distribution, we can infer the update to the utility of each item. Specifically, suppose that for each item the system has an estimate of utility u_i, as well as an estimate of the uncertainty of the utility σ_i. If the system were to take some action a, and observe response r, the system can predict the update of the item utility and uncertainty (for instance, [9] propose a specific virtual-update multi-armed bandit algorithm for this purpose). Summing over all items and possible user responses, the utility of a system action can be computed as the expected reduction in uncertainty about which items have highest utility. This is more difficult in cases where the distribution over user responses is unknown, such as where the user is requested to provide free text. Here, the utility must be estimated based on prior observations of the system when such an action was requested. A deployed conversational system may estimate the expected utility gain of such open-ended requests, and use this for selecting when to perform such actions.

While the above addresses the system-side utility of an action, a second aspect of question utility is that of the cost a system action incurs upon the user. For instance, a question that has high utility in terms of uncertainty reduction may be difficult for a user to answer. This cost needs to be estimated by the system when selecting actions. A simple approach would be to assume a fixed cost for users to respond to any action. This assumption is commonly used by recommendation system where users are requested to label particular examples as part of the learning process. An alternative would be for a conversational system to observe how long it takes a user to respond to a given type of action, and/or how often the user response is not of the requested type for the given system action.

Finally, in the more complex case of set retrieval, the goal of the user is to find a suitable set of items rather than a single item. In this case, once the system has been able to identify items of high utility, it must learn a combined utility function. The form of this utility function would depend on the particular type of information need being addressed. For instance, in a travel planning scenario there may be complex constraints (e.g. only one hotel is needed at a time, the hotel must be near an airport to which there is a suitable flight, and so forth). While the action space is sufficiently rich to allow the system to propose tradeoffs between combinations, we believe that the details of the utility function learning needs to be addressed in a task-by-task manner.

5.4 How can the system interpret user feedback?

Finally, we consider the question of whether our conversational model allows a system to interpret user feedback effectively.

The model is for the system to take an action a at each turn, requesting the user for a specific response. If the user response is of the expected type, the system will clearly be able to interpret it. However, we have also shown it important that the model allows the user to ignore the system's action (question) and provide alternative feedback, as in a real conversation. In this way, the user would be taking his or her own initiative, for instance if the agent is not providing useful information.

One way to reduce the likelihood of unexpected feedback is to explicitly model common conversational outcomes. For instance, our model specifies one of the possible preference answers is r_{np} – that there is no preference among options being presented (with the uniform label still being positive or negative). This was noted by [9] that such feedback is particularly useful in some settings.

While "unexpected" answers are thus the most problematic, we believe that our model provides the right amount of structure to facilitate interpretability. If the user chooses to provide alternative free-text feedback, interpretation of this feedback is relatively limited given the conversational context. Future success of conversational systems adhering to our general model will hinge on how often users choose to (or need to) revert to other answer types, and how well the system captures such deviations. A system that supports memory is also more robust to errors.

Therefore we claim that our model provides a suitable framework for effectively interpreting user feedback.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have described the characteristics of a conversational information retrieval system. These characteristics are based on a broad overview of previous work on human conversations. What have argued that the properties of conversations described are both necessary and sufficient to allow a rich variety of information retrieval tasks to be naturally performed using a conversational interface. We consider the primary contribution of these properties to provide a framework for design and evaluation of future conversational information retrieval systems. In allowing approaches to be compared, the types of tasks that such systems can address, and the way in which they differ from human-level conversations can be more easily characterized.

In doing so, we also discussed when conversational approaches appear most valuable for information retrieval, illustrating with a number of tasks that appear to be well suited to chat-based search.

Following this presentation, we presented a theoretical model that satisfies the conversational properties. While theoretical, the model provides the framework for a conversational search system that appears practical in implementation. The model is a generalization of many specialized systems that have previously been implemented, and have been shown to be effective by previous authors. However, none of the previous systems satisfy all the proposed properties of a conversational information retrieval system. We view the contribution of this model as a proposed structure that can be employed towards obtaining true conversational information retrieval. Implementing the model proposed is the most important future extension of this work.

It is further worth considering limitations in the characterization of conversational information retrieval. It may be the case that some of the properties presented can be replaced with others that serve similar function but lead to higher user satisfaction through more natural interaction. The properties described reflect previous findings of human conversations, thus it may be the case that automatic conversational system do not need to reflect human level conversation to be widely useful for information retrieval. In particular, we have chosen to represent knowledge about the corpus as a utility function defined over items and sets of items. It is possible that such a utility function may not exist, or may be too complex to model in relatively short interactions. Incorporating prior knowledge about global utility and popularity may allow the conversational properties to be refined.

Acknowledgments

The authors would like to thank Cecily Morrison and Alex Taylor for insightful discussions about the nature of (human) conversations. We also thank Paul Thomas for useful feedback on related work, and the anonymous reviewers for helpful comments.

References

[1] J. Allen, C. I. Guinn, and E. Horvtz. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23, 1999.

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

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

[4] P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui. Modeling the impact of short- and long-term behavior on search personalization. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, 2012.

[5] D. Carmel, G. Halawi, L. Lewin-Eytan, Y. Maarek, and A. Raviv. Rank by time or by relevance?: Revisiting email search. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 283–292, 2015.

[6] J. Cassell. More than just another pretty face: Embodied conversational interface agents. *Communications of the ACM*, 4(43):70–78, 2000.

[7] Y. Chen and G. J. F. Jones. Are episodic context features helpful for refinding tasks? lessons learnt from a case study with lifelogs. In *Proceedings of the Information Interaction in Context Symposium (IIiX)*, 2014.

[8] Y.-N. Chen, D. Hakkani-Tur, and X. He. Zero-shot learning of intent embeddings for expansion by convolutional deep structured semantic models. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2016.

[9] K. Christakopoulou, F. Radlinski, and K. Hofmann. Towards conversational recommender systems. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[10] J. Chu-Carroll and M. K. Brown. An evidential model for tracking initiative in collaborative dialogue interactions. In S. Haller, A. Kobsa, and S. McRoy, editors, *Computational Models of Mixed-Initiative Interaction*, pages 49–87. Springer, 1999.

[11] K. Collins-Thompson, C. Macdonald, P. Bennett, and E. M. Voorhees. TREC 2014 web track overview. *NIST Special Publication 500-308: The Twenty-Third Text REtrieval Conference Proceedings (TREC)*, 2014.

[12] E. Cutrell, S. T. Dumais, and J. Teevan. Searching to eliminate personal information management. *Communications of the ACM*, 49(1):58–64, 2006.

[13] R. Dhar. Consumer preference for a no-choice option. *Journal of Consumer Research*, 24(2):215–231, 1997.

[14] M. Dredze, B. N. Schilit, and P. Norvig. Suggesting email view filters for triage and search. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2009.

[15] M. A. Hearst. 'natural' search user interfaces. *Communications of the ACM*, 54(11):60–67, 2011.

[16] M. Henderson, B. Thomson, and S. J. Young. Deep neural network approach for the dialog state tracking challenge. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, 2013.

[17] J. Jiang, A. H. Awadallah, R. Jones, U. Ozertem, I. Zitouni, R. G. Kulkarni, and O. Z. Khan. Automatic online evaluation of intelligent assistants. In *Proceedings of the International World Wide Web Conference (WWW)*, 2015.

[18] D. G. Johnson and T. M. Powers. Computers as surrogate agents. In J. van den Hoven and J. Weckert, editors, *Information Technology and Moral Philosophy*, chapter 13. Cambridge University Press, 2008.

[19] M. P. Kato and K. Tanaka. To suggest, or not to suggest for queries with diverse intents: Optimizing search result presentation. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 133–142. ACM, 2016.

[20] J. Kiseleva, K. Williams, J. Jiang, A. H. Awadallah, A. C. Crook, I. Zitouni, and T. Anastasakos. Understanding user satisfaction with intelligent assistants. In *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR)*, 2016.

[21] G. Linden, S. Hanks, and N. Lesh. Interactive assessment of user preference models: The automated travel assistant. In *User Modeling*, pages 67–78. Springer, 1997.

[22] N. Matthijs and F. Radlinski. Personalizing web search using long term browsing history. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 25–34, 2011.

[23] K. McCarthy, Y. Salem, and B. Smyth. Experience-based critiquing: Reusing critiquing experiences to improve conversational recommendation. In *Proceedings of the International Conference on Case-Based Reasoning*, pages 480–494. Springer, 2010.

[24] R. Nordlie. "user revealment" – a comparison of initial queries and ensuing question development in online searching and in human reference interactions. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 11–18, 1999.

[25] D. G. Novick and S. Sutton. What is mixed-initiative interaction. In *Procedings of the AAAI Spring Symposium on Computational Models for Mixed Initiative Interaction*, pages 114–116, 1997.

[26] T. Paek and E. Horvitz. Conversation as action under uncertainty. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 455–464, 2000.

[27] A. Raux and M. Eskenazi. Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech Language Processing*, 9(1):1:1–1:23, May 2012.

[28] J. Reilly, K. McCarthy, L. McGinty, and B. Smyth. Incremental critiquing. In *Proceedings of the SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 2004.

[29] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, Englewood, Cliffs, New Jersey, 1971.

[30] S. Senecal and J. Nantel. The influence of online product recommendations on consumers' online choices. *Journal of Retailing*, 80:159–169, 2004.

[31] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT)*, June 2015.

[32] P.-H. Su, M. Gašić, N. Mrkšić, L. Rojas-Barahona, S. Ultes, D. Vandyke, T.-H. Wen, and S. Young. On-line active reward learning for policy optimisation in spoken dialogue systems. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

[33] S. Sukhbaatar, a. szlam, J. Weston, and R. Fergus. End-to-end memory networks. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 2440–2448, 2015.

[34] P. ten Have. *Doing Conversation Analysis: A Practical Guide*. SAGE Publications Ltd, 2007.

[35] N. Tintarev and J. Masthoff. Effective explanations of recommendations: user-centered design. In *Proceedings of the ACM conference on Recommender Systems (RecSys)*, pages 153–156, 2007.

[36] D. Traum, D. DeValut, J. Lee, Z. Wang, and S. Marsella. Incremental dialogue understanding and feedback for multiparty, multimodal conversation. In *Proceedings of Intelligent Virtual Agents (IVA)*, pages 275–288, 2012.

[37] M. Walker and S. Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 70–78, 1990.

[38] Z. Wang, J. Lee, and S. Marsella. Towards more comprehensive listening behavior: Beyond the bobble head. In *Proceedings of Intelligent Virtual Agents (IVA)*, pages 216–227, 2011.

[39] J. Weizenbaum. ELIZA – a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9:36–45, January 1966.

[40] R. W. White and D. Morris. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, 2007.

[41] J. Williams, A. Raux, D. Ramachadran, and A. Black. The dialog state tracking challenge. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, 2013.

[42] K. Yao and G. Zweig. Attention with intention for neural network conversation model. In *NIPS Workshop on Machine Learning for Spoken Language Understanding and Interaction*, 2015.

[43] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 401–408. ACM, 2003.

[44] S. Young, M. Gašić, B. Thomson, and J. D. Williams. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013.

[45] Y. Zhang and C. Zhai. A sequential decision formulation of the interface card model for interactive ir. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 85–94. ACM, 2016.