How Am I Doing?: Evaluating Conversational Search Systems Offline

ALDO LIPANI, University College London, UK
BEN CARTERETTE, Spotify, USA
EMINE YILMAZ, University College London & Amazon, UK

As conversational agents like Siri and Alexa gain in popularity and use, conversation is becoming a more and more important mode of interaction for search. Conversational search shares some features with traditional search, but differs in some important respects: conversational search systems are less likely to return ranked lists of results (a SERP), more likely to involve iterated interactions, and more likely to feature longer, well-formed user queries in the form of natural language questions. Because of these differences, traditional methods for search evaluation (such as the Cranfield paradigm) do not translate easily to conversational search. In this work, we propose a framework for offline evaluation of conversational search, which includes a methodology for creating test collections with relevance judgments, an evaluation measure based on a user interaction model, and an approach to collecting user interaction data to train the model. The framework is based on the idea of "subtopics", often used to model novelty and diversity in search and recommendation, and the user model is similar to the geometric browsing model introduced by RBP and used in ERR. As far as we know, this is the first work to combine these ideas into a comprehensive framework for offline evaluation of conversational search.

Additional Key Words and Phrases: information retrieval, conversational search, evaluation, test collections

1 INTRODUCTION

Conversation is increasingly becoming an important mode of interaction with search systems. As use of handheld and in-car mobile devices and in-home "smart speakers" grows, people are utilizing voice as a mode of interaction more and more. And since search remains one of the most common ways people find and access information, search via voice interfaces is more important than ever.

Thus search engines increasingly need to be built for "dialogues" with users. A full ranked list of results—a SERP—is not likely to be useful in such interactions; systems should instead provide a single high-precision answer. Since a user may only get that one answer, it is likely that their next query will be dependent on what that answer is, whether it is to directly follow up with another query, to clarify what they wanted, to move to a different aspect of their need, or to stop the search altogether. In other words, the dialogue the user has with the system is heavily influenced by the system itself.

This is difficult to model in evaluation, particularly in offline evaluations that are meant to be reproducible. In typical offline evaluations, each question in a dialogue is evaluated independently. This, however, does not capture the ability the system has to influence the direction of the conversation. It may fail to identify when the user is following up on a system response. It leaves open whether the system could have better helped the user by providing different information earlier in the conversation. In short, it makes it difficult to optimize or evaluate the system over complete dialogues.

These problems have been raised in the information retrieval literature before. In particular, the TREC Session Track attempted to build test collections for session evaluation and optimization by including short logs of user interactions with search systems [10]. But in the Session track test collections, the previous session is fixed, and only the final query is given as a candidate for retrieval. Thus it is more a relevance feedback task than a session task—it does not solve the problems listed in the previous paragraph. More recently, the TREC Conversational Assistance Track (CAsT) provided

Authors' addresses: Aldo Lipani, aldo.lipani@acm.org, University College London, London, UK; Ben Carterette, Spotify, New York, New York, USA, carteret@acm.org; Emine Yilmaz, emine.yilmaz@ucl.ac.uk, University College London & Amazon, London, UK.

the user side of recorded "dialogues" with a conversational search system [18]. A candidate system would retrieve answers for each of the fixed user inputs provided. While the previous session is not fixed, the user inputs cannot adapt to varying system responses—the user dialogue remains static no matter what the system does.

In this paper we introduce a framework for reproducible offline evaluation of conversational search. The framework includes a novel approach to building test collections as well as a new evaluation metric based on a user model. Our approach is based on the idea of *subtopics*, typically used to evaluate search for novelty and diversity to determine how much redundant information a search system is returning and the breadth of information the user is exposed to. We essentially abstract queries and answers to a subtopic representation, then model the progression of the user and system through the dialogue by modeling subtopic-to-subtopic transitions. We show empirically that our offline framework correlates with online user satisfaction.

The rest of this paper is structured as follows: in Section 2 we discuss related work on similar evaluation problems and conversational search. Section 3 provides a detailed overview of our framework for reproducible evaluation. In Section 4, we describe a specific user model and metric for evaluating conversational search, and in Section 5 we describe the test collection we have assembled. In Section 6 we analyze the results, the test collection, and the metric. We conclude in Section 7.

2 RELATED WORK

A significant amount of research has been devoted to the development of conversational systems [9, 15, 33]. Most research on conversational systems has focused on devising user interfaces for conversational systems [4, 11], using knowledge graphs for question answering in a conversational dialogue [23], building neural models for developing conversational systems [42], incorporating context in response generation [13] and asking clarifying questions [3].

Though a lot of progress has been made regarding the development of conversational systems, until recently, relatively little work had been done in evaluating the quality of conversational systems. Hence, most researchers have been using automatic evaluation metrics such as the BLEU score from the machine translation domain, or the ROUGE from text summarization domain [31, 35]. While these metrics have the advantage to not require explicit human annotations, they were shown to not correlate with actual user satisfaction [26].

Until a few years ago, there was a lack of high quality conversational datasets [37], which was a major challenge towards the development of evaluation metrics for measuring the quality of conversational systems. During the last few years more effort has been devoted to creating such datasets and making them publicly available. Availability of datasets such as MS Marco [30] and Alexa Prize [2] were a significant step forward in the development and evaluation of conversational systems. These datasets were further followed by some other domain specific conversational datasets [32, 44]. Trippas et al. [40] defined a method for creating conversational datasets (which could serve as a guide in constructing such datasets) and used that for building one such dataset.

More research has been devoted to the design of new methodologies for evaluating conversational systems during the last few years. There has been some work on devising benchmark datasets and evaluation metrics for evaluating the natural language understanding component of conversational systems [8, 28]. Focusing on non-goal oriented dialogues, such as the setup for Alexa Prize, Venkatesh et al. [41] proposed several metrics to evaluate user satisfaction in context of a conversational system. Choi et al. [14] proposed a new metric that can be used to predict user satisfaction using implicit feedback signals from the users. Guo et al. [22] evaluate the quality of a conversation based on topical coverage and depth of the conversation. The Conversational Assistance Track in TREC 2019 [18] focused on devising test collections and evaluation methods for evaluating the quality of conversational systems.

Some dimensions of conversational search evaluation have been identified as similar to those of evaluation of interactive information retrieval systems [6]. There has been significant more work in evaluation of interactive information retrieval systems (which have more recently been studied in context of dynamic search) compared to the evaluation of conversational search systems [39]. Dynamic Domain Track aimed devising evaluation methodologies for evaluating dynamic search [43].

Most currently used conversational evaluation metrics including the ones used by the Conversational Assistance Track and the Dynamic Domain Track are based on computing traditional information retrieval metrics on the conversation session. Below we provide a summary of evaluation metrics for a traditional search scenario.

While there has been relatively little work in evaluating the quality of conversational search systems, significant amount of work has been devoted to devising evaluation metrics for evaluation in context of traditional search and recommender systems. However, the analysis of some commonly used offline evaluation metrics used for this purpose show little correlation with actual user satisfaction in context of recommender systems [7, 19] and moderate to negligible correlation in context of search [1].

Offline evaluation metrics based on actual user models have the potential to be more correlated with actual user satisfaction as they are aiming at directly modeling the actual users, where parameters of these models can be learned from user logs [12, 29]. Moffat et al. [27] also argued for having metrics that are attuned to the length of the ranked list, to better align with users who may abandon search early.

One commonly used metric that is based on an explicit user model is Rank-Biased Precision (RBP) [29], which models users' persistence in examining the next retrieved document in a search result. The assumptions made by this metric are that, users examine documents from the top and in order, and that the examination of each document is solely dependent on the willingness of users in doing it, their persistence.

$$\text{RBP}(q) = (1 \quad p) \sum_{n=1}^{N} p^{n \ 1} j(r_n, q),$$

where q is a query, r is the list of documents returned by the search system when querying it with q, r_n is the document retrieved at position n, N is the number of retrieved results, j is the relevance function, which returns 1 if r_n is relevant to q and 0 otherwise, and p 2 $[0, 1]$ is the persistence parameter.

Another commonly used metric based on modeling user behavior is ERR [12]. ERR is based on a similar user model to RBP, but assumes that the probability of user stopping at each rank depends on the relevance of the document observed.

Queries in search could be ambiguous; even the same query could mean different things to different users. Hence, evaluation measures that capture the diverse interests from different users are needed, if the goal is to evaluate the satisfaction of a random user using the search engine. Various evaluation metrics for diversity and novelty based information retrieval have been developed [5, 16]. Some previous work [5, 36] did an analysis of several diversity metrics and proposed new diversity evaluation metrics, which are based on an adaptation of the RBP user model to diversity evaluation.

Most current evaluation metrics used for conversational search are based on session based evaluation metrics, which have been investigated in context of the Session Track [10]. Session based metrics have been widely studied in the literature [10, 38]. Kanoulas et al. [24] proposed two families of measures: a model-free family that makes no assumption about the user behavior over a session, and a model-based family with a simple model of user interactions over the session. Most such session based metrics are adaptions of traditional information retrieval metrics to search sessions.

Metrics used by the Conversational Assistance Track and the Dynamic Domain Track are also based on variants of session based evaluation metrics.

One of the most commonly used session based metrics for evaluating the quality of conversational systems is the fraction of correct answers in the session (i.e., precision of responses in the session) [41]. Lipani et al. [25] extended the RBP user model towards modeling user behavior over the entire search session and proposed the session RBP (sRBP) metric, which could be used for evaluating the quality of conversational search systems. In addition to modeling the probability of a user persisting in (i.e., not abandoning) the search task, the sRBP metric also models the trade-off between re-querying the system and examining a new document down the search result via a new parameter named *balancer*.

$$\text{sRBP}(s) = (1 \quad p) \sum_{m=1}^{} \frac{p \quad bp}{1 \quad bp}^{m \;\; 1} \sum_{n=1}^{} (bp)^{n \;\; 1} j(r_{m,n}, s_m),$$

where $s = [q_1, \ldots, q_m]$ is a session, i.e., a series of queries, $r_{m,\cdot}$ is the search result returned when querying with q_m, $r_{m,n}$ is the document retrieved at position n for q_m. M is the length of the session, that is number of queries submitted in s, and b 2 $[0, 1]$ is the balancer parameter.

One of the primary problems associated with using all the aforementioned metrics, including the session based metrics for evaluating conversational search systems is that all these metrics require that user sessions are known in advance. However, the availability of such data requires that a system has already been deployed and is in use. This is not always feasible. A new system or an academic research system may have no significant user base, and even if it did, one would first want to have an offline evaluation to ensure that the system is of reasonable quality to deploy. Furthermore, within a session, queries issued by a user may depend on the relevance of previous responses by the system; hence, the session itself would be system dependent, and a different system shown to the same user could lead to a completely different session. This means that these metrics cannot be reliably used to compare the quality of multiple systems.

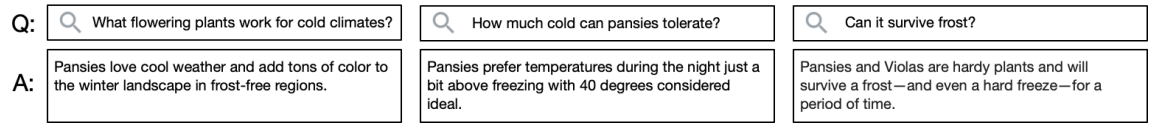
There has been some recent work on simulating dialogues and computing metrics on top of these simulated dialogues [20, 21]. However, this work would require using an agent to simulate the conversations and then using a metric on top of these simulated conversations – a process where metric computation is completely separated from dialogue generation. Furthermore, simulations used in this work are not based on a realistic user model, which is of critical importance when the goal is to devise an evaluation metric correlated with user satisfaction.

Hence, evaluation of conversational search systems without the need for having access to actual search sessions is still an open problem, which was also discussed in a recent Dagstuhl seminar [6], where some of complexities of devising such metrics have been identified.

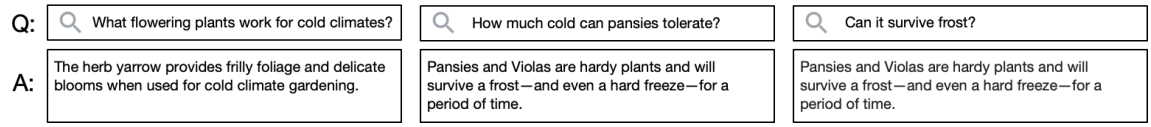
In this paper, we propose a novel offline evaluation framework, which does not rely on having access to real user sessions. Our evaluation metric is based on a user model, which has been validated in context of search [25]. Hence, our experimental results suggest that our offline proposed evaluation metric is highly aligned with actual user satisfaction, in contrast to most information retrieval evaluation metrics that are not based on realistic user models [1, 7, 19].

3 MODELING CONVERSATIONAL SEARCH

In this section we describe an abstract framework for modeling conversational search for reproducible offline evaluation. We contrast against previous approaches, which are based on assessing the relevance of answers to questions independently of one another, as described above.



(a) A dialogue between a user and a candidate conversational search system on the topic of cold flowering plants.



(b) A dialogue between a user and a different candidate conversational search system on the same topic.

Fig. 1. Two dialogues on the same topic with different candidate conversation search systems. The first proceeds in a natural way, with the user's questions answered by the system, which in turn informs the user's next question. The second is less natural, with the second question not clearly following from the answer to the first, and the third question coming despite having just been answered. Both candidate systems are retrieving equally relevant answers.

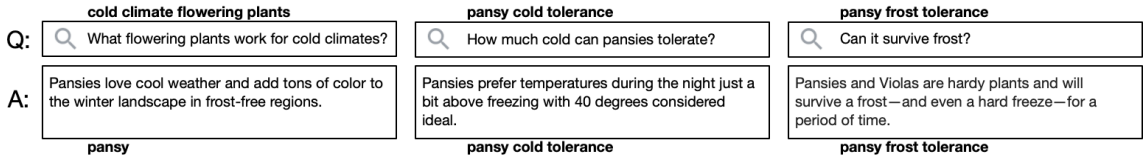
A fully interactive, dialogue-based conversational search system is difficult to evaluate offline in a holistic way. Each turn in the dialogue may be influenced by previous interactions between user and system. Results generated by candidate systems that have not been tested with real users will not necessarily be able to capture these influences. An offline evaluation using fixed, recorded dialogues and independent relevance judgments will almost certainly mis-represent the system's effectiveness in a dialogue for this reason.

Consider the example conversational search dialogues shown in Figure 1. The sequence of user questions is the same in each, but the system responses are very different. Both dialogues start out with the same question: the user asks about flowering plants robust to cold climates. The first system (Fig. 1a) responds with a sentence about pansies. This leads the user to ask how much cold pansies can tolerate; they receive a relevant answer which also motivates the third question, about whether pansies can survive frost. The dialogue makes sense to us: the answers to the questions are relevant; user questions proceed in a natural sequence, motivated by the systems responses to the previous question.

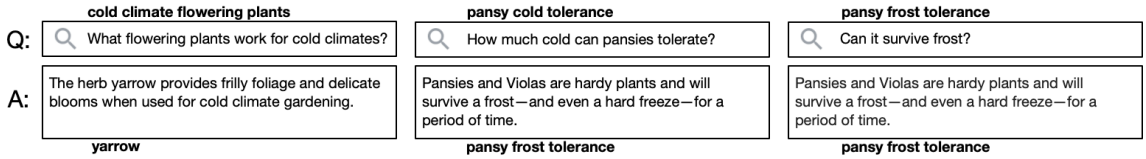
The second system (Fig. 1b) responds to the same first question with a sentence about yarrows, which is relevant to the question. But because the dialogue is static, there is no follow-up question. The user responds by asking about pansies, which is not motivated by anything the system has done. This time, the answer is specifically about the ability of pansies to survive frost—and then the follow-up question is about whether pansies can survive frost. This dialogue makes much less sense: while the answers to the questions are relevant, the questions themselves seem to proceed without much logic given the system responses.

In other words, the two systems shown in this example would perform equally well if judged on the relevance of their answers, but to our eyes, one fares much better than the other. Note that we are not claiming that the second dialogue would never happen. We claim only that it produces an evaluation result that is less satisfying to us than the first: the dialogue being evaluated does not appear to be representative of real dialogues, and therefore we suspect that the outcome of the evaluation is biased in some way.

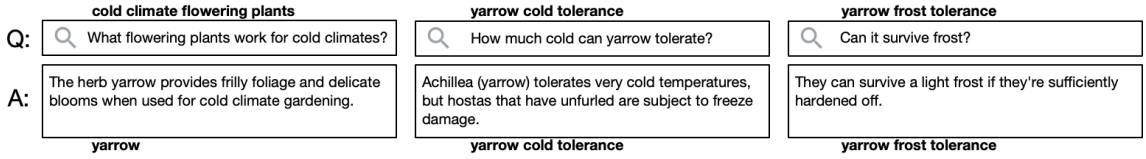
One possible solution to this problem is to evaluate over many different dialogues from users with the same information need. But where do these dialogues come from? Unless we are able to deploy a variety of candidate systems to a large user base, it is unlikely we will be able to obtain them. Can we instead simulate these dialogues? Can we use system responses to generate follow-up questions, and thus do a better job of testing candidate systems over responsive



(a) A dialogue identical to Fig. 1a, but including the subtopic representation of each question and answer. The subtopics follow an intuitive progression.



(b) A dialogue identical to Fig. 1b, but including the subtopic representation of each question and answer, clearly showing that the second question does not follow from the first answer, and the third question is on a subtopic that has already been answered.



(c) A new dialogue with the same candidate system as that in Fig. 2b, in which user questions proceed more naturally from the system responses, and the candidate system's ability to provide relevant results is less clear.

Fig. 2. Three dialogues on the same topic with different candidate conversation search systems, augmented with abstract subtopic representations of questions and answers. The third example suggests that the second candidate system is not as successful at answering questions when they proceed more naturally from each interaction. In our simulation, all three sequences of interactions could occur, but the second would be significantly less likely than the other two.

dialogues? This sounds like a very difficult problem, potentially involving natural language understanding of system responses and natural language generation of user queries.

We propose a simpler approach: we abstract queries and answers to a higher-level representation in a restricted domain to make the simulation manageable. Our representation is based on the idea of *subtopics*, similar to those used in diversity and novelty evaluation [17]. Given a sample user information need, we define a set of subtopics that cover the space of possible things users may learn by interacting with the system. Any given user question is mapped to one of these pre-defined subtopics. Similarly, each system response is mapped to a subtopic. Our evaluation is based on modeling the transitions from the subtopic represented by a system's response to the subtopic represented by possible follow-up questions, and the relevance of the system's responses to those questions.

Figure 2 demonstrates this using the same dialogues as in Figure 1, as well as one new dialogue. Question and answer subtopics are shown above/below respectively. Here we can identify the transitions in the second example (Fig. 2b) that may be less likely: a user following up an answer about the subtopic *yarrow* to a question about the subtopic *pansy*; a user following up an answer about the subtopic *frost tolerance* with a question about *frost tolerance*. Again, this is not to say that the dialogue is "wrong"; only that this dialogue is less likely to occur than one in which the user follows up an answer about *yarrow* with a question about *yarrow*, and does not ask something that has just been answered, and that the framework we propose in this work can capture these differences in likelihood and therefore provide a more robust

evaluation. Fig. 2c demonstrates this with a sequence of questions that are more likely, and that also retrieve responses of lower quality. The answer to the second question is somewhat relevant, but hardly satisfying. It makes sense that a user would follow up by asking if yarrow can survive frost. And the answer to that question is good *if* the user can trust that the system's "they" has the same referent as the "it" in the question. This example shows that when tested using a more natural sequence of questions, the candidate system reveals its performance to be less satisfactory.

Since we do not want to generate natural language questions, our simulation will require that we have a set of possible user questions to sample from, with each question associated with a subtopic. This requirement seems to bring us back to the problem of obtaining many different dialogues for the same information need. But in fact it is significantly lighter than that: since subtopics are treated independently of one another, we can manually develop or crowdsource questions; they do not need to occur in a dialogue in order to be useful in the evaluation. We also need to have a model of transitions from one subtopic to another. This is a heavier requirement, but still lighter than using natural language techniques. We can design tools specifically to obtain these transitions through crowdsourcing.

Given the ideas outlined above, a full dataset for offline evaluation of conversational search would consist of the following:

(1) a sample of user information needs or *topics*, high-level descriptions of things users want to accomplish using a conversational search system;
(2) for each topic, a pre-defined set of subtopics that cover aspects of the need and that influence "turns" in the conversation;
(3) transition probabilities between subtopics, capturing the likelihood that a user goes from one subtopic to another during the course of their interaction;
(4) user queries that model the subtopics;
(5) a corpus of items (documents, passages, answers, etc.) that can be retrieved;
(6) relevance judgments between these items and the subtopics.

An offline evaluation would use the subtopics, queries, transition probabilities, and judgments to simulate a user and system together progressing through a dialogue. It would proceed as follows: an "evaluation agent" is set up to interface with a conversational search system to be evaluated. This evaluation agent may submit any query provided in the test collection and receive an answer from the system being tested. Based on the relevance of this answer, it uses the transition probabilities to sample the next subtopic for which to pose a query. It continues in this way, using an abandonment model to determine when done, at which point the relevance of the answers received along the way are combined into one possible evaluation result. The process is repeated for the same topic over many trials, to simulate many different possible dialogues, and the resulting evaluation scores indicate the effectiveness of the conversational search system for that topic. Over the sample of topics in the collection, we can understand the variation in its effectiveness.

We can now state our research questions. They are:

RQ1 Does a simulation-based offline evaluation accurately capture user satisfaction?
RQ2 Is our metric based on simulations a better fit to user behavior than other metrics for conversational search evaluation?
RQ3 Can our framework detect differences in effectiveness in conversational search systems?

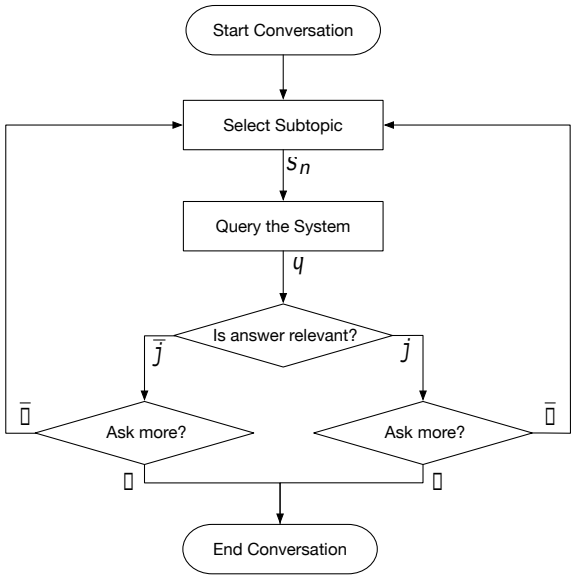


Fig. 3. Flow-chart of the proposed user model.

Before we address those, the next two sections present implementations of the ideas above. First, in Section 4 we describe in detail the user model and metric that we would like to use to evaluate conversational search systems. Then Section 5 describes a specific dataset and a user study performed to gather the data.

4 USER MODEL AND METRICS

In the following we will define two components of a user model for conversational search interaction: a component modeling the user persistence in performing the task, and a component modeling the gathering of information that the user is trying to achieve through the dialogue. The former component is inspired by metrics like RBP for search evaluation, as described in Section 2. The latter is the formalization of the subtopic transition model mentioned above. The two combine into a metric for expected conversational satisfaction.

Given a topic with a set of subtopics S a user wants to learn about, we define a conversation c 2 C as a list of system user interactions with the system. Each interaction consists of a query q and an answer a:

$$c = [(q_1, a_1), \ldots, (q_m, a_m)],$$

where each q can be abstracted to a subtopic s 2 S (q 2 Q_s) and each pair (q_i, a_i) has a relevance judgment j 2 J .

4.1 User Persistence

Figure 3 depicts the user model in a flow chart. When users have an information need, they begin a "dialogue" with a conversational system by issuing a query. Based on the relevance of the result, they decide whether to continue querying or not; moreover, each turn in the dialogue is modeled as the user trying to find information about a particular subtopic. Thus their persistence in querying the system is dependent on what is observed by the previous query. We

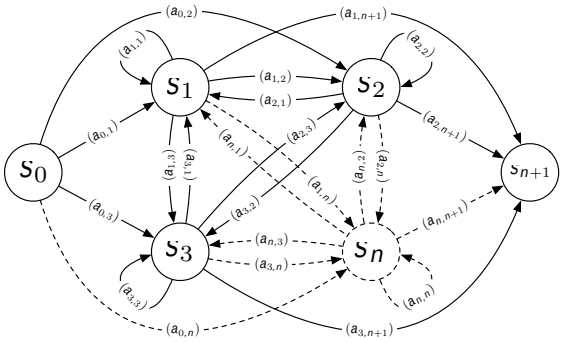


Fig. 4. Graphical model of subtopic transitions during information gathering. Note that s_0 is a "dummy" *start* state, and s_{n+1} is an *end*.

model this using the following recursive definition:

$$p(Q_1 = q) = 1$$
$$p(Q_m) = p(L_{m-1} | Q_{m-1}, J_{m-1}) p(Q_{m-1}).$$

This recursive definition uses three random variables, Q, L, and J, where $Q = \{q, \overline{q}\}$ indicates the act of querying or not, $L = \{\ell, \overline{\ell}\}$ the act of leaving or not, and $J = \{j, \overline{j}\}$ indicates the relevance or not of the system reply. The first equation models the probability of starting a conversation when the user has an information need. The second equation models the probability of continuing the conversation with the system, which is modeled as dependent to the previous interaction with the system. Note that the second equation does not specify the outcomes of the random variables in order to consider all possible combinations of the outcomes thereof.

We assume that the probability of continuing the conversation given that the user has not previously queried the system is equal to 0:

$$p(L_{m-1} = \ell | Q_{m-1} = \overline{q}) = 0.$$

For the sake of clarity, we introduce parameters α^+ and α^- to substitute with the probability of continuing the conversation given that the user has previously queried the system and the previous returned result was relevant:

$$p(L_{m-1} = \ell | Q_{m-1} = q, J_{m-1} = j) = \alpha^+,$$

and when the previous returned result was not relevant:

$$p(L_{m-1} = \ell | Q_{m-1} = q, J_{m-1} = \overline{j}) = \alpha^- .$$

Both probabilities will be estimated from user logs; Section 5.2 provides more detail. Moreover, when these two αs are equal, the model is equivalent to that of RBP and sRBP, with α functioning as the persistence parameter p.

4.2 User Information Gathering

The user's task is to gather information about a topic by interacting with the conversational system. To describe the user interaction we use the probabilistic graphical model shown in Figure 4. Given a topic, we define the set of subtopics the user wants to satisfy as follows:

$$S = \{s_1, \ldots, s_n\}.$$

We model this by treating each subtopic as a state. To these states we add a *start* state and an *end* state, indicating the initial (s_0 = *start*) and final states (s_{n+1} = *end*) of the conversation. We define the probability of transitioning to any state given that we have started the conversation as:

$$p(S_m = s_1 | S_{m-1} = s_0) = a_{0,1},\ p(S_m = s_2 | S_{m-1} = s_0) = a_{0,2},\ \ldots$$
$$\ldots,\ p(S_m = s_n | S_{m-1} = s_0) = a_{0,n},\ p(S_m = s_{n+1} | S_{m-1} = s_0) = 0.$$

The final probability guarantees that at least one search interaction needs to be performed.

To define the probability of going from subtopic i to any other subtopic, including the end state, we use two approaches. The first, called *relevance independent* (RI), assumes that these probabilities are independent of the relevance of the system's answers:

$$p(S_m = s_1 | S_{m-1} = s_i) = a_{i,1},\ \ldots,\ p(S_m = s_n | S_{m-1} = s_i) = a_{i,n},\ p(S_m = s_{n+1} | S_{m-1} = s_i) = a_{i,n+1},$$

where $i \in \{1, \ldots, n\}$.

Our more advanced *relevance dependent* (RD) representation assumes that these probabilities depend on the relevance of the system's answers and estimates these probabilities conditioned also on relevance:

$$p(S_m = s_{i_1} | S_{m-1} = s_{i_2}, J_{m-1} = j),\ p(S_m = s_{i_1} | S_{m-1} = s_{i_2}, J_{m-1} = \bar{j}),$$

where $i_1 \in \{1, \ldots, n+1\}$ and $i_2 \in \{0, \ldots, n\}$.

We indicate the act of sampling a state (subtopic) using these estimations as $s \sim S_{j,s'}$, where j represents the relevance of the previously retrieved document and s' is the previous subtopic to which the previously submitted query belongs to. This last relationship is formalized by the set Q_s, which indicates the set of queries associated to the subtopic s. The act of sampling a query is indicated as $q \sim Q_s$. The relevance of a subtopic to an answer is obtained using a qrels file and it is indicated as $r \sim J_{s,a}$. This modeling gives us the opportunity to also capture a more noisy concept of relevance, where user factors like agreement, disturbance, and distractions are modeled by sampling using the probability that a is relevant to s.

4.3 Evaluating a Conversation

Based on the user model defined above, we now define our proposed evaluation metric Expected Conversation Satisfaction (ECS), which is an expectation over many dialogues simulated offline using the two models above. Algorithm 1 shows how we estimate Conversation Satisfaction with a single (simulated) dialogue, given a system and a series of estimates. Over many trials of this algorithm, we obtain ECS. Later in the paper we will describe how to estimate the probabilities needed to compute this metric.

Finally, to ensure that the metric is in the range of 0 to 1, we normalize the metric by dividing it by the metric value for an ideal conversation in which every reply is correct. We call this nECS, and define this metric as:

$$\text{nECS}(c) = \frac{\text{ECS}(c)}{\text{IECS}(c)}.$$

This is similar to the way many information retrieval metrics such as RBP and ERR are normalized.

Algorithm 1: Computation of ECS

Input: α^{+}, α , S, Q, J , system()
Output: score

score 0
$p(Q = q)$ 1
relevant false
subtopic start
subtopic $S_{\text{relevant, subtopic}}$
while subtopic 6 *end* **do**
  query Q_{subtopic}
  answer system(query)
  relevant $J_{\text{subtopic, answer}}$
  if relevant **then**
    score score + $p(Q = q)$
    $p(Q = q)$ $\alpha^{+} p(Q = q)$
  else
    $p(Q = q)$ α $p(Q = q)$
  end
  subtopic $S_{\text{relevant, subtopic}}$
end

4.4 ECS in Practice

In order to compute ECS in practice, one would need to identify (annotate) the possible subtopics given a topic and compute the transition probabilities between the different subtopics.

If usage logs of a real conversational system are available, the metric can be computed with respect to one fixed conversation from the log, and the expected value of the metric could be obtained by averaging across all such conversations that fall under the same topic. Given a real (non-simulated) conversation, it can be shown that ECS can be computed as:

$$\text{ECS}(c) = \sum_{m=1}^{|c|} j(c_m) \prod_{m'=1}^{m-1} (\alpha^{+} j(c_{m'}) + \alpha \ (1 \quad j(c_{m'}))),$$

where j returns the relevance of the system answer to the user query provided at step m.

Note that this is not generally an option for offline evaluation of new systems, as we expect them to retrieve answers that have not previously been seen in user dialogues, in which case the metric with the simulated dialogues (as described in Algorithm 1) needs to be used. Similarly, the simulated dialogues also need to be used when there is no access to the usage logs of a conversational system.

However, such logs are not necessarily always available in practice. In such cases, test collections need to be constructed in order to estimate the parameters of the model and compute the value of the metric. In the next section we describe a procedure that can be used to construct such a test collection and show how the parameters of Algorithm 1 can be estimated using such a test collection.

5 DATA COLLECTION

In this section we describe the work we did to operationalize the framework presented in Section 3 and collect data to fit the user model presented in Section 4. We created a dataset based on SQuAD [34] for question answering. SQuAD consists of topics defined from manually-chosen Wikipedia pages. Each selected page is broken down by paragraph, and for each paragraph there are several associated questions that can be answered by that paragraph.

SQuAD is designed for evaluating one-off question answering, not conversations or dialogues. To simplify the use of this data, we decided to focus on systems that respond to questions with full paragraphs. This is because the paragraphs are straightforward to use as a unit of retrieval; we elaborate in Section 5.1 below. For future work we will look at subdividing paragraphs into smaller units for assessing and retrieval within this framework.

An example paragraph from the SQuAD topic "Harvard University" is as follows:

> Established originally by the Massachusetts legislature and soon thereafter named for John Harvard (its first benefactor), Harvard is the United States' oldest institution of higher learning, and the Harvard Corporation (formally, the President and Fellows of Harvard College) is its first chartered corporation. Although never formally affiliated with any denomination, the early College primarily trained Congregationalist and Unitarian clergy. Its curriculum and student body were gradually secularized during the 18th century, and by the 19th century Harvard had emerged as the central cultural establishment among Boston elites. Following the American Civil War, President Charles W. Eliot's long tenure (1869–1909) transformed the college and affiliated professional schools into a modern research university; Harvard was a founding member of the Association of American Universities in 1900. James Bryant Conant led the university through the Great Depression and World War II and began to reform the curriculum and liberalize admissions after the war. The undergraduate college became coeducational after its 1977 merger with Radcliffe College.

Each of the following questions is provided as part of SQuAD and can be answered by the paragraph above:

(1) What individual is the school named after?
(2) When did the undergraduate program become coeducational?
(3) What was the name of the leader through the Great Depression and World War II?
(4) What organization did Harvard found in 1900?
(5) What president of the university transformed it into a modern research university?

SQuAD also provides questions that cannot be answered by the text. We chose to discard these for this study.

In order to use this data in our framework, we first need to define subtopics, then associate questions and paragraphs with our subtopics. We selected 11 SQuAD topics to use in our study. They are listed in Table 1. For each topic, we defined a set of subtopics by manually examining the SQuAD questions and the original Wikipedia page. We attempted to develop a set of subtopics that were largely mutually exclusive with one another and that covered most of the desired information reflected by the provided questions. Subtopics are represented as short keyphrases with a longer text description explaining exactly what should and should not be considered relevant to the subtopic. Each topic has between four and nine subtopics. Table 2 shows some example subtopics and relevant questions from the SQuAD data.

We then manually judged each question in the SQuAD dataset for that topic relevant to one of the defined subtopics. Each question could be relevant to at most one subtopic. Questions that were judged not relevant to any subtopic were marked nonrelevant and excluded from the study. On average, the ratio of subtopic-relevant questions to topic-relevant questions is 0.11, that is, each subtopic represents about 11% of the topic's questions.

topic	subtopics	questions	paragraphs
harvard university	5	259	29
black death	5	108	23
intergovernmental panel on climate change	6	99	24
private school	12	113	26
geology	6	116	25
economic inequality	6	291	44
immune system	6	214	49
oxygen	9	239	43
normans	4	95	40
amazon rainforest	5	181	21
european union law	11	231	40

Table 1. Eleven topics selected from the SQuAD data, with the number of questions and paragraphs contained in SQuAD as well as the number of subtopics we manually developed for each topic.

topic	subtopic	example question
harvard university	harvard facts	How many individual libraries make up the main school library?
	harvard alumni	What famous conductor went to Harvard?
	harvard finances	How much more land does the school own in Allston than Cambridge?
	harvard academics	How many academic units make up the school?
	harvard history	In what year did Harvard President Joseph Willard die?
economic inequality	historical inequality	During what time period did income inequality decrease in the United States?
	economic theory	What does the marginal value added by an economic actor determine?
	economists	What organization is John Schmitt and Ben Zipperer members of?
	current state of inequality	In U.S. states, what happens to the life expectancy in less economically equal ones?
	causes of inequality	Why are there more poor people in the United States and Europe than China?
	solutions to the problem	Who works to get workers higher compensation?

Table 2. For two of the topics we selected, our subtopics and, for each one, an example question from the SQuAD data. Each of these questions has an accepted answer that can be extracted from a paragraph in the SQuAD data for the topic.

Based on these question-subtopic relevance judgments, paragraph relevance could then be automatically assessed by mapping the relevance of the questions associated with that paragraph. A paragraph could be relevant to zero or more subtopics, depending on the questions that it answered.

After completing this process, we have five of the six components required by the framework:

(1) topics: subject-based Wikipedia articles that have been used in the SQuAD dataset;
(2) subtopics: manually developed based on the SQuAD questions and Wikipedia pages;
(3) user queries: questions from the SQuAD dataset that have been judged relevant to the selected subtopics;
(4) retrievable items: paragraphs from the topical Wikipedia pages;
(5) relevance judgments: obtained from the question-subtopic relevance judgments.

This is enough to evaluate the relevance of paragraphs retrieved in response to the provided user questions. However, this does not give much more than a standard question-answering system. To evaluate a conversation, we need more: we need a model of how a user progresses through the conversation based on the responses they are provided with. The work in Section 4 describes such a model; we now turn to collecting data to fit it.

5.1 Crowdsourcing Study

In this section we describe a crowdsourcing study to gather user queries and data to fit the user model. We designed a prototype search system for the SQuAD-derived dataset described above. Users (Mechanical Turk workers), upon accepting the work, were shown instructions to ask questions on a topic provided to them. They were given an interface to input questions. The system responded to questions with a paragraph. The user was asked whether the paragraph is relevant to their question, and to what subtopic their question related. They could end their search at any time, at which point they were asked to indicate their satisfaction with the session.

Note that this is not meant to reflect a "real" search scenario. Users in a real search setting would not be asked to select a subtopic to represent their question. They would likely not be asked about the relevance of each response. Furthermore, we imposed a strong restriction on the set of candidates that could be retrieved for each question: the system would only select paragraphs from a small, manually-selected set relevant to the subtopic the user specified. Clearly this information would not be available to a real search engine. The reason for these decisions is that our goal is not to evaluate our system with users, but to collect user data for our models described in Section 4.

Here we would like to discuss two important decisions regarding the retrieval of paragraphs in response to user questions. First, we recognize that full paragraphs (like the one exemplified above) would not typically be a retrieval unit in a real conversational search system. The reason we chose to use full paragraphs regardless is that paragraphs often touch on several different subtopics or aspects of the topic. Our subtopic-based framework thus makes it possible to extract more insight into how system responses affect user questions than if we had used shorter units of retrieval such as sentences or passages that only answer the question posed. It is straightforward to use our framework to evaluate a system that is retrieving shorter passages than a full paragraph; it is primarily a matter of assessing the relevance of retrieval units to the defined subtopics.

Second, our system only retrieves paragraphs from a small, manually-curated set that are known to be relevant to the query (by the question-subtopic judgments and their mapping to paragraphs). One reason for this is that many of the paragraphs are difficult to grasp without the context of other paragraphs in the full document—for example, a paragraph entirely focused on Radcliffe College has little meaning to a user that does not already know about the 1977 merger of Harvard University and Radcliffe College. Thus we specifically selected paragraphs that start new sections or that can be easily understood without additional context. Furthermore, the paragraphs are selected to be "exemplary" for the subtopic, so that if the paragraph is relevant to the subtopic, it is easy to understand why and to see why other paragraphs (that require more context to understand) would be relevant to the same subtopic. We chose the example paragraph above as being a straightforward introduction to the topic, being relevant to the subtopics "harvard history" and "harvard academics", and being exemplary for the former. One hypothesis raised by this paragraph is that users may next query about academics, since the paragraph alludes to curriculum.

Since there are other candidate paragraphs relevant to one or the other of these example subtopics ("harvard history" and "harvard academics"), we need a way for the system to select one to respond when given a user query. We use a simple language-modeling approach, where paragraphs are modeled as a multinomial distribution of terms in the vocabulary. Here the vocabulary is restricted to the terms used in the topic itself (not the full corpus). As a form of smoothing, each paragraph is expanded using the terms in the subtopic label, so for example the paragraph above would be expanded with repetitions of the terms "harvard", "history", and "academics". Additional smoothing is done using Dirichlet priors based on the prevalence of terms in all of the paragraphs relevant to the subtopic, plus all terms

in the topic. When a user enters a query and its subtopic, the relevant paragraphs are scored with this language model, and the top-scoring paragraph is selected for retrieval.

Our users were Amazon MTurk workers. The HIT they had to complete involved completing a dialogue for one topic. They were free to do as many HITs as they wished. In all, we collected 207 dialogues from 220 unique workers. The average length of a dialogue was 5.43 turns. Users marked 816 out of 1123 responses relevant, and indicated satisfaction with 72% percent of their sessions.

Since we specifically restricted retrieval to relevant paragraphs, why were nearly 30% marked nonrelevant? The explanation boils down to disagreements between users and ourselves about the meaning of the subtopic labels as well as the relevance of paragraphs to those labels. We trust that the users know best, so for the remainder of the study we use the user-assessed judgments rather than the subtopic-question judgments.

5.2 Subtopic Transitions

The final component of the framework is the transition probabilities between subtopics in a topic. We compute these from the user data collected as described above. We use a simple Bayesian prior and updating approach. Since a relevance-independent (RI) transition probability $p(S_m = s_j | S_{m-1} = s_i)$ is multinomial, we initially assume a Dirichlet prior with equal-valued parameters $a_{0,i}, \ldots, a_{j,i}, \ldots, a_{n+1,i}$, resulting in a uniform posterior. Each time we observe a transition from s_i to s_j in a user dialogue, we simply increment the prior parameter $a_{j,i}$ by one, causing the posterior probability to increase.

For the relevance-dependent (RD) model, we use a very similar approach. The only difference is that we additionally condition each multinomial distribution on the user-assessed relevance of the answer received at turn m.

6 RESULTS AND ANALYSIS

In this section we use the experiment data from above to analyze our ability to perform reliable, reproducible offline evaluation of conversational search. As we wrote in Section 3, our research questions are:

RQ1 Does a simulation-based offline evaluation accurately capture user satisfaction?
RQ2 Is our ECS metric a better fit to user behavior than other metrics for conversational search evaluation?
RQ3 Can our framework detect differences in effectiveness in conversational search systems?

The data we use to investigate these questions is as follows:

(1) Users' self-reported satisfaction in their search, as described in Section 5.1.
(2) Three evaluation metrics computed for the recorded user dialogues (*not* simulated):
  (a) Precision (P), the proportion of correct answers in the dialogue;
  (b) Rank-biased precision (RBP), a geometric-weighted version of precision;
  (c) Expected conversation satisfaction (ECS), the measure we propose in Section 4.
(3) The same three evaluation metrics averaged over N simulated dialogues generated using Alg. 1.
(4) A candidate conversational search system based on language modeling.

We will answer the research questions by showing that:

RQ1 The metrics computed over simulated dialogues (item 3) all correlate well with user satisfaction (item 1).
RQ2 ECS correlates better with user satisfaction and fits user querying behavior better than the other metrics, with both non-simulated (item 2) and simulated (item 3) data.
RQ3 As system (item 4) effectiveness decreases in a controlled way, ECS decreases in an expected pattern.

	Sim.	Parameters	ln(TSE)	ln(TAE)	KLD
P	RI		-2.1201	0.5277	2.1544
	RD		-2.0980	0.5572	2.3984
RBP	RI	$\alpha = 0.79$	-5.2268	-1.3926	0.0734
	RD	$\alpha = 0.79$	-5.2268	-1.3926	0.0734
ECS	RI	$\alpha^+ = 0.82, \alpha = 0.70$	-5.2617	-1.4062	0.0718
	RD	$\alpha^+ = 0.85, \alpha = 0.64$	**-5.2774**	**-1.4174**	**0.0706**

Table 3. Model parameters and errors.

	Sim.	Parameters	τ	ρ	r
P	RI		0.3963	0.4184	0.4659
	RD		0.6606y	0.8200y	**0.8577**y
RBP	RI	$\alpha = 0.79$	0.3963	0.4184	0.4660
	RD	$\alpha = 0.79$	0.6606y	0.8200y	0.8389y
ECS	RI	$\alpha^+ = 0.82, \alpha = 0.70$	0.3963	0.4184	0.4515
	RD	$\alpha^+ = 0.85, \alpha = 0.64$	**0.6972**y	**0.8383**y	0.8492y

Table 4. Correlations (Kendall's τ, Spearman's ρ, and Pearson's r) between self-reported user satisfaction and metric values (with two different dialogue simulation models over 100,000 trials per topic).

6.1 Results

We first investigate correlation between conversational evaluation metrics and self-reported user satisfaction. The primary results are based on simulating dialogues using Algorithm 1 with subtopics sampled according to one of two approaches described in Section 4: *relevance independent* (RI) transitions between subtopics, and *relevance dependent* (RD) transitions. For each topic, we simulate 100,000 dialogues with the candidate system and compute the three metrics for each, averaging over dialogues to obtain an expected value for the topic. We estimate expected satisfaction with a topic by averaging the binary user-reported satisfaction responses. We then compute correlations between one of the metrics and the expected satisfaction values.

In order to compute RBP and ECS, we first need to optimize the parameters of the metrics. To set the parameters we performed a simple grid search to minimize total square error (TSE) between $p(Q_m)$ as estimated by their user models and the actual probability of a user reaching the mth turn in a dialogue. In Table 3 we show the model hyperparameters obtained by optimizing the fit of the model. In the table we also report the best achieved TSE (as well as corresponding total absolute error (TAE) and KL-divergence) between these two quantities for simulations based on all three metrics, as well as the two types of transitions (RI and RD). As it can be seen, the two variants of ECS tend to achieve lower modeling error compared to the other metrics.

Once the model parameters are computed, we can now compute the correlations between the different metrics and user satisfaction labels obtained from our participants. Table 4 shows results for three different correlation coefficients: Kendall's τ rank correlation, based on ranking topics by expected satisfaction across all our participants and counting the number of pairwise swaps needed to reach the ranking by the metric; Spearman's ρ rank correlation, a linear correlation on rank positions; and Pearson's r linear correlation between the numeric values themselves. All three correlation coefficients range from 1 to 1, with 0 indicating a random relationship between the two rankings. A metric

	Parameters	ln(TSE)	ln(TAE)	KLD
P		-2.7129	-0.1005	0.5317
RBP	$\alpha = 0.80$	-5.3199	-1.5106	0.0496
ECS	$\alpha^+ = 1.00, \alpha = 0.53$	**-7.5696**	**-2.6574**	**0.0054**

Table 5. Model parameters and errors for the non-simulated case.

	Parameters	τ	ρ	r
P		0.6972y	0.8383y	**0.9178**y
RBP	$p = 0.80$	**0.7340**y	0.8611y	0.8952y
ECS	$\alpha^+ = 1.00, \alpha = 0.53$	**0.7340**y	**0.8702**y	0.9088y

Table 6. Correlations between self-reported user satisfaction and metrics computed on real sessions.

that consistently scores higher on topics for which users self-report higher satisfaction, and vice-versa, will have a higher correlation.

The maximum reported correlations in this table are very strong (and statistically significant) correlations, showing that indeed our simulation-based framework can accurately capture user satisfaction, supporting a positive answer to **RQ1**. Note that the simulation using relevance-dependent transition probabilities correlates far better than the simulation using relevance-independent transition probabilities. This suggests that it is the case that users adjust their questions in response to system answers, and an evaluation that fails to model this fails to model user satisfaction well.

Table 3 and Table 4 together demonstrate the strength of ECS as a metric: it achieves lower modeling error and better correlations than the other two metrics (apart from the higher linear correlation that precision achieves). This supports a positive answer to **RQ2**, that ECS is a better fit to user behavior than other metrics. The differences are small but consistent, suggesting a real effect, though we must note that sample sizes are too small to detect statistical significances in the differences of the closest results. We leave tests with larger datasets for future work.

To investigate correlations and model errors in more depth, consider Tables 5 and 6. These tables are similar to Tables 3 and 4, but take as input a single real user dialogue—they are based on no simulation. They are thus not reflective of offline evaluation scenarios for which real dialogues do not exist. But they could be thought of as a sort of ceiling (or floor) for the correlation (or model error, respectively). Our RD simulation achieves correlations very close to those reported in Table 6, suggesting that not only is it a good model of user satisfaction, it is approaching the best possible performance for any model based on the same assumptions. Furthermore, the errors are substantially lower than with simulated data. These tables reinforce **RQ1**, supporting the idea that simulation is an acceptable substitute for real user dialogues, as well as **RQ2**, in that ECS fits user behavior better than either P or RBP on real user dialogues.

Finally, to answer **RQ3**, we simulated "real" retrieval systems by degrading the one used by our users. These systems were progressively more likely to return irrelevant answers, by adding additional paragraphs to the candidate sets for each subtopic. Recall that the system users used would only retrieve responses relevant to the subtopic. The same system degraded to by 10% could potentially retrieve an additional 10% of the full corpus of paragraphs, introducing much more possibility for error in retrieval results. These additional answers would be both irrelevant to users, and also redirect the dialogue in unexpected ways, thus testing both the ability of the metric to measure degraded performance as well as the ability of the simulation to respond to such degradation. Figure 5 shows the result. As noise (irrelevant responses) increases, effectiveness drops precipitously. This supports a positive answer to **RQ3**.

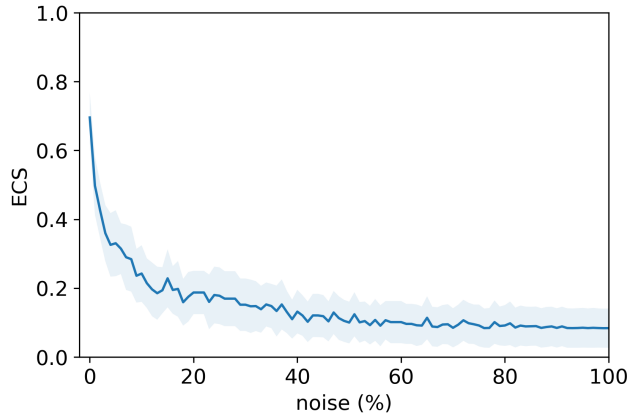


Fig. 5. ECS computed across all topics (over 10,000 trials per topic) varying the system noise, from 0%, where only the relevant answers to each subtopic can be retrieved, to 100% where any answer in the dataset can be retrieved.

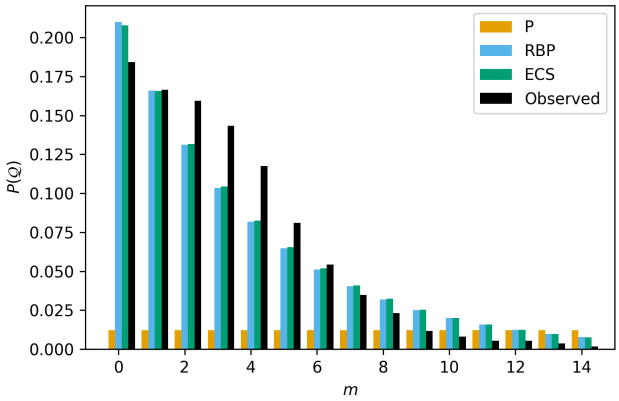


Fig. 6. Comparison of estimated $P(Q_m) = m$ in relevant-dependent simulations (over 100,000 trials per topic) vs. observed user behavior.

p(of querying the **same subtopic** at step m j the **answer was relevant** at step m 1)	0.121
p(of querying the **same subtopic** at step m j the **answer was nonrelevant** at step m 1)	0.340
p(of querying the **another subtopic** at step m j the **answer was relevant** at step m 1))	0.879
p(of querying the **another subtopic** at step m j the **answer was nonrelevant** at step m 1)	0.660

Table 7. Marginal probabilities computed on the observed conversations.

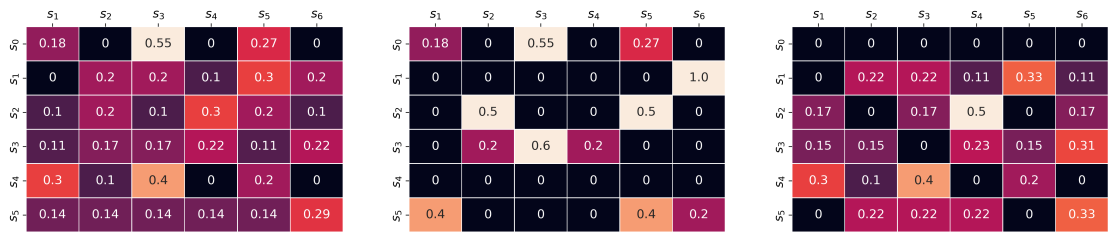


Fig. 7. Transition probability tables for the 'Harvard University' topic. This topic has 5 sub-topics. s_0 is the initial state and s_6 is the end state. On the left we have the RI case, on the center and the right we have the RD case: the first is when the answer is not relevant the second when the answer is relevant.

6.2 Additional Analysis

6.2.1 Modeling errors. Table 3 above summarized model errors. Figure 6 illustrates the errors in a more granular way, showing how well the user model fits the observed user data, specifically in terms of the probability of a user reaching turn m in a dialogue. Note that RBP and ECS are about equal; which is also reflected in Table 3.

6.2.2 Conditional transition probabilities. Table 7 reports some marginal conditional probabilities from our user experiment. In particular, users are more likely to switch to a different subtopic if they have just seen a response relevant to their current subtopic than if they have not. As well, users are more likely to ask about the same subtopic if the answer is not relevant than if it is. This shows that the system's answers *do* influence user behavior, as we argued in Section 3.

Figure 7 shows examples of empirical subtopic transition probabilities for the "Harvard University" topic with five subtopics, plus the start and end states s_0 and s_6. Each row contains the probability of transitioning from the state indicated by the row label to the state indicated by the column label. The first matrix is used in the RI case, while the second and third matrices are used in the RD case; the second is conditional on non-relevance while the third is conditional on relevance.

We make some observations based on these figures. The first is fairly uniform: when transitions are not based on relevance of responses, the simulation produces no strong tendency to move in any particular way through the subtopic graph. The second is quite sparse: when users are provided with answers that they do not find relevant (recall that we asked users to indicate relevance as well, and despite the system retrieving from a subset of relevant paragraphs, users could disagree) there are typically only a few options they take. In some cases (s_2, s_3, and s_5), they are likely to issue another query on the same subtopic. In one case (s_1) they give up immediately. The chance of switching subtopics for this topic is relatively low, as we saw in aggregate in Table 7. The third is interesting in that the diagonal is all zeros: users never follow up a relevant answer on one subtopic with another question on the same subtopic. This demonstrates part of our original motivation for the work, that users questions are dependent on system responses.

7 CONCLUSION

We have introduced a novel approach for offline reproducible evaluation of conversational search systems. Our approach models user queries and system responses with subtopics, an idea from novelty and diversity search evaluation. We

propose a metric based on simulating users transitioning between subtopics in the course of a dialogue. Our simulation-based methodology correlates strongly with user satisfaction, and our metric correlates better than others.

Our approach has limitations. The label "conversational search" could be applied to a wide variety of problems and search scenarios, and like any class of search problems, it is unlikely there is any one-size-fits-all solution to evaluation for all possible settings. Ours is ideal for settings with relatively complex information needs that cannot be answered in a single turn, but that do have factual answers; for which the desired information can be represented by a finite set of discrete subtopics; for which information returned for one query may influence future queries; and when the information returned is relatively long-form (sentence or paragraph length).

The proposed approach considers only conversational systems where the main initiative is provided by the user. In fact, the notion of persistence is only modeled from the user perspective, that is the only one who decides when the interaction should stop. In mixed-initiative conversational systems, where the initiative is also taken by the system, the system could also decide when to stop the interaction. Hence, a possible extension of this approach could be the introduction of a system's persistence similar to the concept of user's persistence. This would be in line with the notion of pro-activity of conversational search systems as suggested by Trippas et al. [40].

We cannot infer from the collected data if a user found a paragraph relevant to the rest of the subtopics. This is because in the crowdsourcing experiment we only asked users to indicate if a document is relevant to the submitted query to which a subtopic is associated. Therefore we can only be certain about the relevance of the retrieved paragraph to the queried subtopic. This limitation, although makes the crowdsourcing task more realistic, does not allow us to make stronger assumptions about to which subtopics the user has already been exposed in previous interactions with the system.

Nevertheless, an offline evaluation framework that accurately captures user satisfaction, that can address the problem of dialogues with new systems taking turns that are not seen in online systems, that is fully reproducible, and that is relatively straightforward to implement will be an invaluable tool for developers of conversational search systems. As our immediate next step, we intend to implement our framework to train and test real conversational search systems.

ACKNOWLEDGMENTS

This project was funded by the EPSRC Fellowship titled "Task Based Information Retrieval", grant reference number EP/P024289/1.

REFERENCES

[1] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The Relationship between IR Effectiveness Measures and User Satisfaction. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands) *(SIGIR '07)*. Association for Computing Machinery, New York, NY, USA, 773–774. https://doi.org/10.1145/1277741.1277902

[2] AlexaPrize. 2020. The Alexa Prize the socialbot challenge. https://developer.amazon.com/alexaprize/

[3] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. 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* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 475–484. https://doi.org/10.1145/3331184.3331265

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[5] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 625–634. https://doi.org/10.1145/3209978.3210024

[6] Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational Search (Dagstuhl Seminar 19461). *Dagstuhl Reports* 9, 11 (2020), 34–83. https://doi.org/10.4230/DagRep.9.11.34

[7] Joeran Beel, Marcel Genzmehr, Stefan Langer, Andreas Nürnberger, and Bela Gipp. 2013. A Comparative Analysis of Offline and Online Evaluations and Discussion of Research Paper Recommender System Evaluation. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation* (Hong Kong, China) *(RepSys '13)*. Association for Computing Machinery, New York, NY, USA, 7–14. https://doi.org/10.1145/2532508.2532511

[8] Daniel Braun, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. 2017. Evaluating Natural Language Understanding Services for Conversational Question Answering Systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Saarbrücken, Germany, 174–185. https://doi.org/10.18653/v1/W17-5522

[9] H.C. Bunt. 1981. Conversational principles in question-answer dialogues. In *Zur Theorie der Frage: Kolloquium, 1978, Bad Homburg: Vortraege / hrsg. von D. Krallman und G. Stickel (Forschungsberichte des Instituts fuer Deutsche Sprache Mannheim)*. Tübingen, 119–142.

[10] Ben Carterette, Evangelos Kanoulas, Mark M. Hall, and Paul D. Clough. 2014. Overview of the TREC 2014 Session Track. In *TREC*.

[11] Justine Cassell. 2001. Embodied Conversational Agents: Representation and Intelligence in User Interfaces. *AI Magazine* 22, 4 (Dec. 2001), 67. https://doi.org/10.1609/aimag.v22i4.1593

[12] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 621–630. https://doi.org/10.1145/1645953.1646033

[13] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *"Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing"*. Association for Computational Linguistics, Brussels, Belgium, 2174–2184. https://doi.org/10.18653/v1/D18-1241

[14] Jason Ingyu Choi, Ali Ahmadvand, and Eugene Agichtein. 2019. Offline and Online Satisfaction Prediction in Open-Domain Conversational Systems. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 1281–1290. https://doi.org/10.1145/3357384.3358047

[15] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 815–824. https://doi.org/10.1145/2939672.2939746

[16] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore) *(SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 659–666. https://doi.org/10.1145/1390334.1390446

[17] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Ellen M. Voorhees. 2011. Overview of the TREC 2011 Web Track. In *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, USA, November 15-18, 2011*. http://trec.nist.gov/pubs/trec20/papers/WEB.OVERVIEW.pdf

[18] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The Conversational Assistance Track Overview. In *TREC*.

[19] GG Gebremeskel and AP de Vries. 2016. Recommender Systems Evaluations: Offline, Online, Time and A/A Test. In *CLEF 2016: Working Notes of CLEF 2016-Conference and Labs of the Evaluation forum Évora, Portugal, 5-8 September, 2016*. [Sl]: CEUR, 642–656.

[20] David Griol, Javier Carbó, and José M. Molina. 2013. An Automatic Dialog Simulation Technique To Develop and Evaluate Interactive Conversational Agents. *Appl. Artif. Intell.* 27, 9 (Oct. 2013), 759–780.

[21] David Griol, Javier Carbó, and José M. Molina. 2013. A Statistical Simulation Technique to Develop and Evaluate Conversational Agents. 26, 4 (2013), 355–371.

[22] Fenfei Guo, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram. 2017. Topic-based Evaluation for Conversational Bots. *NIPS Conversational AI Workshop*.

[23] Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. 2015. Learning Knowledge Graphs for Question Answering through Conversational Dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, 851–861. https://doi.org/10.3115/v1/N15-1086

[24] Evangelos Kanoulas, Ben Carterette, Paul D. Clough, and Mark Sanderson. 2011. Evaluating Multi-Query Sessions. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 1053–1062. https://doi.org/10.1145/2009916.2010056

[25] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2019. From a User Model for Query Sessions to Session Rank Biased Precision (sRBP). In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval* (Santa Clara, CA, USA) *(ICTIR '19)*. Association for Computing Machinery, New York, NY, USA, 109–116. https://doi.org/10.1145/3341981.3344216

[26] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2122–2132. https://doi.org/10.18653/v1/D16-1230

[27] Fei Liu, Alistair Moffat, Timothy Baldwin, and Xiuzhen Zhang. 2016. Quit While Ahead: Evaluating Truncated Rankings *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 953–956. https://doi.org/10.1145/2911451.2914737

[28] Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking Natural Language Understanding Services for building Conversational Agents. In *10th International Workshop on Spoken Dialogue Systems Technology 2019 (IWSDS '19)*. https://iwsds2019.unikore.it/

[29] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1, Article 2 (Dec. 2008), 27 pages. https://doi.org/10.1145/1416950.1416952

[30] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *CoCo@NIPS*, Vol. abs/1611.09268. http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[31] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[32] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Stockholm, Sweden, 353–360. https://doi.org/10.18653/v1/W19-5941

[33] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 117–126. https://doi.org/10.1145/3020165.3020183

[34] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2383–2392. https://doi.org/10.18653/v1/D16-1264

[35] Ehud Reiter. 2018. A Structured Review of the Validity of Bleu. *Comput. Linguist.* 44, 3 (Sept. 2018), 393–401. https://doi.org/10.1162/coli_a_00322

[36] Tetsuya Sakai and Zhaohao Zeng. 2019. Which Diversity Evaluation Measures Are "Good"?. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 595–604. https://doi.org/10.1145/3331184.3331215

[37] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A Survey of Available Corpora For Building Data-Driven Dialogue Systems: The Journal Version. *Dialogue & Discourse* 9, 1 (2018), 1–49.

[38] Zhiwen Tang and Grace Hui Yang. 2017. Investigating per Topic Upper Bound for Session Search Evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (Amsterdam, The Netherlands) *(ICTIR '17)*. Association for Computing Machinery, New York, NY, USA, 185–192. https://doi.org/10.1145/3121050.3121069

[39] Zhiwen Tang and Grace Hui Yang. 2019. Dynamic Search–Optimizing the Game of Information Seeking. *arXiv preprint arXiv:1909.12425* (2019).

[40] Johanne R. Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a model for spoken conversational search. *Information Processing & Management* 57, 2 (2020), 102162. https://doi.org/10.1016/j.ipm.2019.102162

[41] Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. 2017. On Evaluating and Comparing Conversational Agents. In *NIPS Conversational AI Workshop*.

[42] Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. In *ICML Deep Learning Workshop*. http://arxiv.org/pdf/1506.05869v3.pdf

[43] Grace Hui Yang and Ian Soboroff. 2016. TREC 2016 Dynamic Domain Track Overview. In *TREC*.

[44] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A Dataset for Document Grounded Conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 708–713.