Analysing Mixed Initiatives and Search Strategies during Conversational Search

Mohammad Aliannejadi
University of Amsterdam
m.aliannejadi@uva.nl

Leif Azzopardi
University of Strathclyde
leif.azzopardi@strath.ac.uk

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

Evangelos Kanoulas
University of Amsterdam
e.kanoulas@uva.nl

Paul Thomas
Microsoft
pathom@microsoft.com

Nick Craswell
Microsoft
nickcr@microsoft.com

ABSTRACT

Information seeking conversations between users and Conversational Search Agents (CSAs) consist of multiple turns of interaction. While users initiate a search session, ideally a CSA should sometimes take the lead in the conversation by obtaining feedback from the user by offering query suggestions or asking for query clarifications i.e. mixed initiative. This creates the potential for more engaging conversational searches, but substantially increases the complexity of modelling and evaluating such scenarios due to the large interaction space coupled with the trade-offs between the costs and benefits of the different interactions. In this paper, we present a model for conversational search – from which we instantiate different observed conversational search strategies, where the agent elicits: (i) Feedback-First, or (ii) Feedback-After. Using 49 TREC WebTrack Topics, we performed an analysis comparing how well these different strategies combine with different mixed initiative approaches: (i) Query Suggestions vs. (ii) Query Clarifications. Our analysis reveals that there is no superior or dominant combination, instead it shows that query clarifications are better when asked first, while query suggestions are better when asked after presenting results. We also show that the best strategy and approach depends on the trade-offs between the relative costs between querying and giving feedback, the performance of the initial query, the number of assessments per query, and the total amount of gain required. While this work highlights the complexities and challenges involved in analyzing CSAs, it provides the foundations for evaluating conversational strategies and conversational search agents in batch/offline settings.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; • **Human-centered computing → HCI design and evaluation methods**.

KEYWORDS

Conversational Search, Mixed Initiatives, Evaluation

ACM Reference Format:
Mohammad Aliannejadi, Leif Azzopardi, Hamed Zamani, Evangelos Kanoulas, Paul Thomas, and Nick Craswell. 2021. Analysing Mixed Initiatives and Search Strategies during Conversational Search. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3459637.3482231

1 INTRODUCTION

Conversational Search (CS) is an emerging area of research that aims to couch the information seeking process within a conversational format [5, 19, 21]. CS differs from the traditional query-response paradigm by providing more agency through improved query understanding and the persistence of the conversational context [61, 65]. The exciting prospect of *Conversational Search Agents* (CSAs) has spurred considerable research into the development of the underlying methods to support such agents. Of particular interest have been methods that facilitate mixed initiative approaches that aim to enhance the agent's understanding of the user's information need through query suggestions (i.e. refinements, expansions, etc.) or through query clarifications (i.e. questions that seek to clarify the query, elicit the user preferences, etc.) [1–3, 37, 50, 68]. This is because mixed initiative interactions are seen as a key property of a conversational search agent [49] which has the potential to increase user engagement and user satisfaction [35]. While various efforts have focused on building the infrastructure to support the inclusion of clarifying questions, and numerous methods proposed to generate or select good questions [3, 17, 25, 39, 52, 68, 69], little work has evaluated or compared the use of such methods within the context of a CS session in a batch/offline setting — largely because the possible state space increases exponentially with interaction, coupled with the lack of a user model for CS. So while there has been considerable effort in the community to engage in single and mixed initiative conversations, little has been done to understand how they impact performance during CS sessions.

In this work, our goal is to provide a user model for conversational search that can be used to evaluate mixed initiative approaches and conversational strategies. While asking query clarifications and offering query suggestions may lead to increases in user satisfaction in certain scenarios [35], it also imposes additional costs on the user; the premise being that the investment in feedback will lead to greater returns later. And so the costs and the expected gains associated with different mixed initiative approaches will determine whether eliciting or giving feedback is worthwhile compared to other actions that could be taken (e.g. re-querying or assessing) [9, 57]. So how should

an agent interact with a user? Should it ask a series of clarifications, and then present results, or present results, and then ask for clarifications? Or, not ask any clarifications? It is very much an open question what conversational search strategy should be employed in order to minimise the conversational cost while maximising the user's gain. And, how different mixed initiative approaches would influence the choice of strategy given the user's interactions (i.e. whether they assess more, give more feedback, or issue more queries). In this paper, we aim to provide insights into these research questions by modelling the CS process and then measuring the costs and benefits of different CS strategies and mixed initiative approaches.

2 BACKGROUND

Over the past few years, an increasing amount of attention has been directed toward developing methods that enable CS and the development of CSA, for example: ranking results given the conversation [22, 66], generating clarifying questions [3, 68, 70], studying system-initiative interactions [62], and presenting results [55]. Less attention, however, has been focused on developing user models for evaluating CS which can be used to analyse CSAs and CS strategies.

One of the first CS systems was proposed by Croft and Thompson [20], called I^{3}R. It acted as an expert intermediary system, communicating with the user during a search session. Since then, other researchers have developed more elaborate approaches. For example, Belkin et al. [14] offered users choices in a search session using case-based reasoning. While, Allen et al. [4] were among the first to study mixed initiative conversations, which they defined as "*a flexible interaction strategy in which each agent (human or computer) contributes what it is best suited at the most appropriate time*". However, since then researchers have mainly focused on single-initiative interaction such as rule-based conversational systems [63] and spoken language understanding approaches [26, 47]. Mixed initiative, though, provides a mechanism for the agent to improve its understanding of the user's information need by obtaining feedback by offering query suggestions (i.e. refinements to the query) or query clarifications (i.e. questions that seek to clarify the query) [3, 50, 68]. As previously mentioned, this idea of mixed initiative and the system taking agency has led to the development of CSAs. Inspired by models and work on conversations and dialogue systems (e.g. COR, etc. [14, 43, 46, 53]), Radlinski and Craswell [49] developed a theoretical framework, that puts forward five key properties that a search system needs to have in order to be "*conversational*". These properties are:

- **User Revealment** where the user discloses to the agent their information needs,
- **Agent Revealment** where the agent reveals what the agent understands, what actions it can perform, and what options are available to the user,
- **Set Retrieval** where the agent needs to be able to work with, manipulate and explain the sets of options/objects which are retrieved given the conversational context,
- **Memory** where the agent tracks and manages the state of the conversation and the user's information need, and,
- **Mixed Initiative** where both the agent and the user can take the initiative and direct the conversation search process.

Azzopardi et al. [9] extended this framework by defining the specific actions associated with these aspects. For example, within mixed initiative, they suggest that agents could seek to provide query suggestions or query clarifications that help to refine the user's information need, or seek to elicit the user's preferences, while users could, conversely, suggest refinements and disclose preferences. Trippas et al. [57] examined how searchers interacted with intermediaries (who used the search engine), engaged actions and observed that searchers generally switched between *query formulation* (user revealment) and result exploration (set retrieval), but also provided relevance feedback and clarifications. Following on from this work, Trippas et al. [59] suggested a more general classification of the different interactions grounded by empirical studies – their high level model delineates between: (i) *discourse level actions*, that enable discourse management, grounding, visibility, and navigation and (ii) *task level actions*, that are specific to the search such as handling queries, search assistance (e.g. clarifying queries), presenting results, and search progression. In another empirical study that analysed a number of CS datasets, Vakulenko et al. [60] found that users issue a query to the agent, and then the agent may respond with a request to clarify/refine the information need, or provide a list of results. The user could then respond by either issuing a new query, responding to the request, providing feedback, or assessing a result. They referred to this as the QFRA model [60]. They observed different patterns of behaviour such as *query-feedback loops*, where several rounds of feedback to clarify/refine their query were observed, before results were assessed (*Feedback-First*), and *assessment-feedback loops*, where the user inspected results and then provided feedback to clarify/refine their query (*Feedback-After*).

Zhang et al. [72] proposed a *System Ask, User Respond* paradigm, which is akin to query-refinement, where after the initial request is made, the agent will ask for refinements/clarifications until it is confident enough to present results. Kaushik et al. [32] presented a system sided workflow model consisting of the steps the agent takes when dealing with a user's request (e.g. handling greetings and error handling). Under their model when a query is entered, it is checked and if a clarification is needed, the user is asked for one clarification, else, the agent retrieves and presents three results. If these are not relevant, or the user wants to see more items, they can request another three results. Alternatively, they can request to view the document (or a summary of). Otherwise, they can issue a new query (or stop). A similar approach is presented in [64] where up to eight rounds of feedback were performed. While Dubiel et al. [23] presented a similar conversational workflow model, however, the agent asks for up to three rounds of feedback to refine the user's request, before presenting two results. More recently, Lipani et al. [38] they proposed a CS model based on exploring subtopics via a query-response paradigm, however, it did not consider feedback. In this work, we aim to explicitly model feedback and explore its impact within the conversational search process.

The various models proposed share a number of commonalities inherent to *Interactive Information Retrieval* (IIR). Interaction consists of a number of turns based around: *Query Formulation*, where the user expresses their query, *Result Exploration*, where the user examines results, and *Query Reformulation*, where the user updates their query [40, 51]. In terms of modelling, simulating, and evaluating the IIR process, the focus has largely been on considering sessions, rather than mixed initiatives. For example, the user browsing model which is at the heart of most IR evaluation metrics, assumes a user

will pose a query, and then examine documents in a top down fashion [18, 45]. For IIR, the model has been extended such that the user decides with some probability of examining the next document, or issuing a new query [13], or examining a fixed number of documents before issuing a new query [7]. Through modelling the IIR process it has been shown trade-offs emerge between querying and assessing [7], where, for example, Azzopardi et al. [10] found that as query cost increased, users submitted fewer queries, and compensated by examining more results. But, to model CS, it is clear that the mixed initiative (feedback turn) needs to be explicitly considered within the user model. However, this will add additional complexities – and invariably introduce new trade-offs because gathering feedback to refine or clarify the query will come at a cost – which may or may not lead to more gain – and so issuing a new query or assessing another result may be more beneficial. In [9, 59], they point out that it is important for CSA to maximise the gain it delivers to the user, while trying to minimise the cost – and thus maximise the rate of gain (following Grice's Maxims of Conversation [24]). With the introduction of mixed initiative approaches and different CS strategies for engaging with a CSA there are many open questions. Is giving feedback (clarifications or suggestions) worth the cost, under what conditions is it beneficial, and, what type of CS strategy (feedback first or after) leads to a higher rate of gain?

3 A MODEL OF CONVERSATIONAL SEARCH

As previously discussed, conceptually conversational search can be seen as a special case of IIR [19] – where the interaction between the user and the agent is based around conversational turns – and the agent is more active in the seeking process through mixed initiative interactions. So far we have not specified the details of the CSA, which could be: (i) a voice only CSA often via a virtual assistant [23, 36, 51, 58, 59]), (ii) a chat based CSA (that is in-situ within a platform like Slack [6] or Telegram [67]), (iii) an augmented search engine interface [16], or (iv) a multi-modal virtual assistant [30, 67]. Given the wide range of CSAs, it is not possible to fully model them all – so we need to make a number of assumptions about the type CSA we will model. Below, we outline what affordances the agent provides, and then what actions a user can take with respect to the search process e.g. we are focusing on the salient search interactions (task level), and not on modelling error handling, chit-chat, etc.(discourse level). Given a CSA, we assume that the agent can initiate or respond as follows:

(i) present results/answers to the user given their query,
(ii) request additional feedback to update the query,
(iii) or some combination of these.

The response (or combined response, and order of) will depend on numerous factors e.g. the modality of the agent, bandwidth available, and agent capabilities. We assume that the user in turn can perform the following actions during the search process:

(i) issue a query,
(ii) provide feedback, or
(iii) assess a result/answer.

Given the affordances of the agent and how the user can interact with the agent, then we can conceptualise the conversational search process as a series of turns consisting of three main turn types: (τ_Q) query turn, (τ_F) feedback turn, and (τ_A) assessment turn.

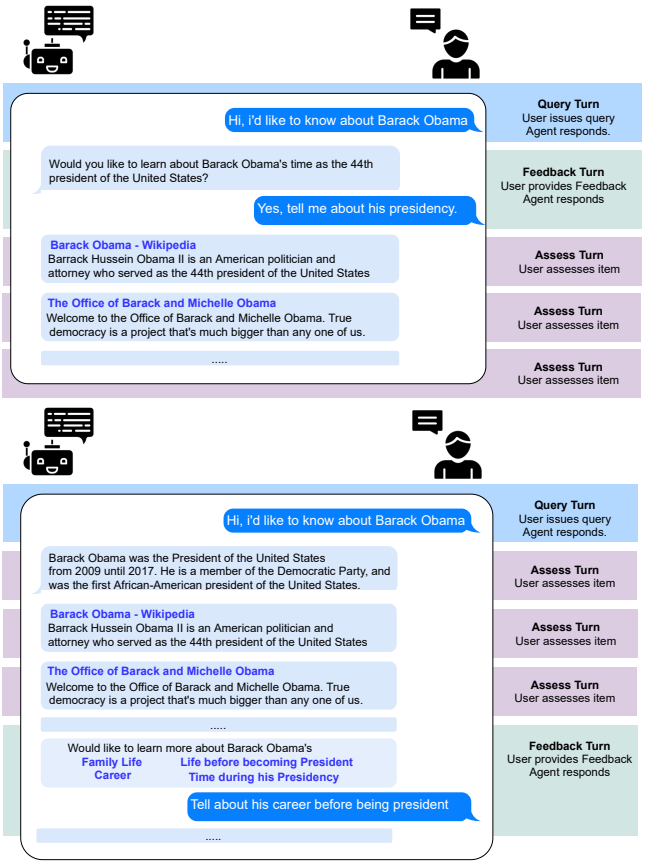


Figure 1: Example chat-based CSA where one agent asks clarifications questions before presenting results (top), while the other agent asks query suggestions after presenting results (bottom). An open question is what strategy (e.g. initiate first or after results) and what type of mixed initiative (e.g. clarifications or refinements) leads to a better CS?

As previously mentioned, how the agent responds will be different depending on the type of agent, its modality, and the current context. For example, (i) a voice only CSA needs to be very sensitive to the limited and serial bandwidth of speech, and so responses are likely to be shorter, (ii) a multi-modal CSA could present a more detailed combined response i.e. a search engine result page that asks a number of requests for feedback (via query suggestions, facets, etc.), provides many results, etc., while (iii) a chat based CSA agent has some restrictions on screen space and bandwidth within the chat window, but has the advantage over voice only CSAs because the conversation is persistent so the user can refer back to options, etc.. For the purposes of this work, we will assume that the CSA is a chat-based CSA like that in Fig. 1 which represents the interfaces explored in [6, 32, 67]. Of note is that depending on the interface and its modality, the cost of different conversational turns will vary – and this will impact on how much gain the user accumulates from their conversation with the agent. As proposed in [9, 58], we also assume that a user wants to maximise the amount of gain they receive from the system, while trying to minimise the cost of the conversation (where the cost could be the total number of turns, or the total time taken to perform those turns). Note that the assumed objective does not necessarily mean that a user prefers a shorter conversation, but

one that yields a higher rate of gain. An open question then is how should a user and CSA work together in order to maximise the user's rate of gain? Should the user give many rounds of feedback, or should they examine some results, and then give feedback to the system? Should the CSA request more feedback, or provide more results?

3.1 Modelling the Conversation Search Process

To model the conversation search process, we draw upon the previous work that has conceptualised conversational and interactive search, with the aim to formalise the key actions/turns described above within a *Markov Decision Process* (MDP) model (as done in [12, 13, 41, 45, 56] for IIR). In these works, MDPs (or variants of) have been used to represent key decisions that users make when searching and interacting with a system and their key actions that they take. For example, the simple *User Browsing Model* (UBM) that underpins most metrics in IR [18], assumes that a user will issue a query, then assess a result item accumulating some gain if it is relevant. The user then decides to either continue and assess the next item with some probability of continuing or stop examining result items [44, 45]. In [41, 56], the UBM was extended to IIR to include additional decision points to model session search. However, because conversational search affords mixed initiative where feedback can be elicited from the user, the process is much more complicated. This additional affordance means that we need to include other decision points to capture these conversational turns – and integrate the feedback process with the browsing process within the larger context of the search session.

Fig. 2 presents our model of the CS process – where we have broken up each of the user choices into binary decisions (denoted by diamonds), and the three actions/turns with circles. We assume that the user starts the search process by issuing a query (τ_Q). Given the query, the agent responds either with a list of results, clarifications, suggestions, or a combination of. Essentially the agent presents the "*search engine result page*" either via a web page or a chat bot via text (as in the Figure 1) or through speech. Given the response, the user may decide to either inspect results or give feedback depending on what options are presented by the agent. If they choose to assess a result (τ_A) they follow the typical UBM, shown in light purple. While if they choose to provide feedback (τ_F) then they follow the User Feedback Model (UFM) shown in light green.

Following the UBM, if a user performs a τ_A turn, then they inspect a result item, where it is assumed that they will accumulate some gain if the result is relevant, and then they need to decide whether to perform another τ_A turn, or not [44, 45]. The decision to continue, would of course, depend on a number of factors such as how much gain has been accumulated, how many items have been examined/assessed, etc. [44]. Once they decide to stop assessing, the user may decide to give feedback to the agent, in order to refine/expand their current query, or not. If not, the user can then decide whether to re-formulate their query, in which case repeating the process. Otherwise, they stop searching. Similarly, in the case, where the user decides to give feedback (τ_F), the can provide feedback to the agent, where it is assumed that the agent will provide an updated response, and then the user needs to decide whether to perform another round of feedback (τ_F), or not. Once they stop giving feedback, the user can decide whether to go back to assessing, or not. And, if not they then need to decide whether to re-query, or stop searching altogether.

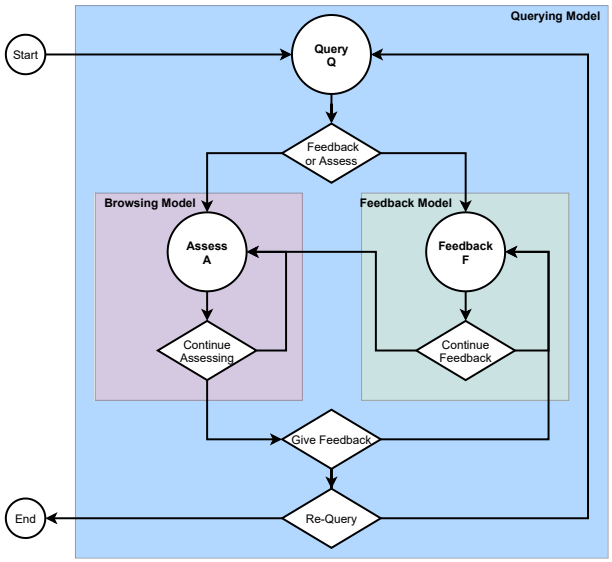


Figure 2: The User Model of Conversational Search which is composed of three sub-components the Querying, Browsing and Feedback Models. Diamonds represent user decision points, while circles represent the action/turn taken.

Fig. 1 presents two example conversations. In the top example, the user issues a query, and the agent responds by asking a query clarification, the user responds and then the agent presents a number of results. In the bottom example, the user issues a query, and the agent responds with a number of results, followed by a request for feedback via query suggestions, to which the user responds, and the conversational continues. However, the space of possible sequences of different turns grows rapidly. And, herein lies the complexity of evaluating conversational search – after t turns, the number of possible conversational sequences, is approximately 3^t for a fully mixed initiative CSA. Nonetheless, the number of possible sequences of conversational turns exponentially increases with the number of turns. This presents an open challenge in evaluating Conversational Search Agents.

3.2 Instantiating the User Model for CS

In order to make the problem tractable, we need to reduce the number of possibilities so that we can simulate and then evaluate the CS process. Grounded by observed behaviours from [60], we propose two strategies for conversational search:

- **Feedback First (FF)**: where the user performs a query-feedback loop before assessing. That is, after querying the user given F rounds of feedback, before assessing A items.
- **Feedback After (FA)**: where the user performs assessment-feedback loops, where after assessing A items, the user gives feedback, and then repeats the process F times.

These two interaction models represent two "*pure*" strategies that users/agents might evolve/apply. The first approach, Feedback-First, represents a CSA that is like a Librarian or Booking agent.Here, the agent asks the user a number of clarifying questions or makes a number of suggestions to refine the user's information need before presenting results to the user. The second approach, of Feedback

After, represents a more exploratory search setting where the user learns about the topic, and then provides feedback to the agent to progress their search through the topic space. While in practice it is likely that the optimal CS strategy would be a mixture of FA and FF, investigating these strategies is feasible, and has not been previously evaluated. Given these two strategies, we aim to draw insights into how and when they are more successful and under what conditions. For example, how do the performance of the initial query, the cost of turns, the type of feedback, and the searcher's strategy interact and influence performance?

3.3 Evaluating the Gain and Cost of CS

While evaluation in a traditional IR setting is primarily concerned with measuring the expected utility of a ranked list, CS introduces an interaction space that grows exponentially with interaction. This makes evaluating different strategies and methods more complicated, because the different interactions have different costs and provide different benefits. For example, giving feedback comes at a cost, on the hope that it will lead to accruing more gain later on. Obviously, if feedback turns are expensive, and they don't lead to greater increased gain, then the "conversational" part of the search may not be beneficial. To represent the costs associated with each action, we model the cost for the conversational turns τ_Q, τ_F and τ_A as $c(\tau_Q)$, $c(\tau_F)$ and $c(\tau_A)$, respectively. The cost associated with each turn will depend on the response and the modality of the CSA. Consequently, when considering which CS strategy or which CSA is better than another, we cannot be agnostic to the cost of the conversation. Both cost and gain arising from the CS need to be measured. In the CS setting, we can generalise the cumulative gain metric from traditional IR evaluations [28] to turns, where the *Turn-based Cumulative Gain* of a sequence of conversational turns is: $G(t_1,...,t_T) = \sum_{i=1}^{i=T} g(t_i)$ where T is the total number of turns, and $g(t_i)$ is the gain obtained from the ith turn, and t_i is either τ_Q, τ_F or τ_A. As each turn comes at a cost, then the total cost, is: $C(t_1,...,t_T) = \sum_{i=1}^{i=T} c(t_i)$, where $c(t_i)$ is the cost of performing the ith turn. The subsequent rate of gain, can then be calculated as total gain divided by the total cost: $R(t_1,...,t_T) = \frac{G(t_1,...,t_T)}{C(t_1,...,t_T)}$ While discounting or session based metrics could be applied [11, 29, 38], we leave such directions for further work, as it is not clear how the discounts would or should be applied in this context.

4 RESEARCH QUESTIONS

In the context of conversational information seeking, where a user wants to explore a topic, and find out about various facets of the topic, through a conversational chat bot interface (like the one in Fig. 1), we aim to obtain insights into the following research questions:

- How does the conversational strategy (Feedback-First or Feedback-After) affect performance?
- How does the mixed initiative approach (Clarification or Suggestion) affect the performance?
- How is the strategy and/or approach affected by the quality of the initial query?
- How is the strategy and/or approach affected by changes in the cost of turns?

5 EXPERIMENTAL METHOD

To answer our research questions, we have opted to undertake a simulated analysis as done in previous works on IIR [8, 27, 31, 33, 41, 42, 71]. This is because the space of possible interaction sequences is very large and evaluating the different combinations would not be feasible in a user study. However, we do ground our analysis by conducting a user study to obtain estimates of the costs of performing different turns using a text based CSA (as in Fig. 1).

Collection. Following Aliannejadi et al. [3] we use the topics created as part of the TREC Web Track from 2009 to 2012, based on the ClueWeb09-Category B collection. The collection consists of 198 topics. Each topic consists of a series of facets that the user would like to explore – making them suitable to explore in a conversational manner because clarifications and suggestions can help refine or redirect the search towards the different facets that the user wants to explore. To ensure having a reasonable space of exploration, we filter out the topics that have fewer than four facets or fewer than ten relevant documents. These steps lead us to 49 topics with a total of 211 facets (approx 4.3 per topic).

Conversational Search Agent. For our study, the CSA is defined by: (i) the conversational strategy that it employs either *Feedback-First* (FF) or *Feedback-After* (FA), (ii) the mixed initiative approach of *Query Clarification* (QC) or *Query Suggestion* (QS), (iii) the number of rounds of feedback that it offers (F), and (iv) the number of result items it presents to be assessed by the user (A).

Retrieval of Results. Given the query, and any subsequent clarifications or suggestions, we pre-process the query terms (i.e. stopword removal and stemming) and submit it to the retrieval system. To retrieve the ranked list of documents, we use an extension of the *Query Likelihood Model* (QLM) for CS proposed in Aliannejadi et al. [3] with the suggested parameters. The model is a linear interpolation of the language model based on the query submitted by the user, and the language model based on the feedback. Once the results are retrieved, the result lists are filtered, and only previously unseen result items are presented to the user. We assume the CSA has a memory of what results the user has already seen.

User Interactions. Following the user model presented in Fig. 2, we assume that the user follows the search strategy given by the specific CSA. Below we describe how our simulated users generate queries which they issue during query turns, and then describe the feedback presented to them during feedback turns.

Query Generation (Q). To generate the queries we employed the approach given by [8, 31]. For each topic, a language model is created given the set of documents relevant to the topic. Then, to generate a query of length L, terms are sampled without replacement from the top 20 terms given their relative entropy in the language model.

Feedback (F) - Query Clarifications and Query Suggestions. Given the query issued, we assume that the agent is able to either (i) ask clarifying questions or (ii) provide query suggestions. The answers to the clarifying questions, or selection of the query suggestions, are then used to improve the query representation. Each τ_F turn is expected to lead to a better query representation, which in turn should lead to improved query performance. We take two approaches for simulating feedback:

- **Query Clarifications.** For clarifications, we used the query clarifications from the Qulac dataset [3] along with the human responses. We followed [37] and pre-processed the data to remove redundant clarifications and low quality answers.

- **Query Suggestions.** For suggestions, we used the same query generation algorithm as before to generate additional terms used as suggestions. As shown in Fig. 1, the user is presented with four query suggestions, and when giving feedback the user selects a suggestion at random.

Each successive round of feedback given, adds additional terms to the original query. We checked the performance of the resulting expanded queries given the query clarifications or suggestions and found that there was no significant difference between the two approaches at neither P@10 nor P@20.

Calculating the Gain. To calculate the gain, we follow Section 3.3, where we assume that the user only accumulates gain on an assessment turn (τ_A), where $g(\tau_A)=1$ when the user assesses a previously unseen relevant item, otherwise $g(\tau_A)=0$ and for the other turns: $g(\tau_Q)=g(\tau_F)=0$. That is, a user only received gain when they are provided with relevant and novel information during the conversation (as done in [7, 13, 41, 54]).

Estimating the Cost. To ground the estimation of the costs for each of the conversational turns, we conducted a user study where we designed four crowdsourcing tasks (HITs) on Amazon Mechanical Turk[1]. In all of our tasks, we first showed the user a search topic description (from a total of five search topics from the TREC Web Track). Our choice of topic was based on their difficulty and type (informational and faceted), aiming to cover a wide spectrum of search tasks in the study. Each search session started with a query from the user. Once the user clicks the Search button, they were shown either a result snippet or document and asked to judge its relevance (definitely relevant, possibly relevant, non-relevant). As soon as the worker assessed one snippet (or document), we show the next snippet (or document). We repeated this process for five results. After assessing the fifth result, we instructed the workers to either: (i) reformulate their query to look for a different facet of the topic, (ii) provide feedback by answering a clarifying question, or (iii) select one of the four query suggestions. Each HIT provided data for 20 result assessments, 1 or 4 queries, and 3 rounds of feedback. We had 81 workers undertake the HITs, who submitted 144 queries, assessed 1,280 result snippets, 1000 result web pages, and provided 268 responses to feedback. The average time taken to issue a query was 29.3 seconds, to assess a result snippet was 6.3 seconds, to assess a result web page was 17.0 seconds, while the average time to provide feedback was 8.3 seconds. While in practice the cost of selecting suggestions vs. providing clarifications will differ depending on the implementation we wanted to compare the two approaches as fairly as possible – and thus kept the feedback costs the same between mixed initiatives.

To calculate the total cost, we follow Section 3.3 where we set the costs as: $c(\tau_Q)=29.3$, and $c(\tau_F)=8.3$. For estimating $c(\tau_A)$ we draw upon past work [13, 54], where the cost of assessing an item depends on its relevance, such that: $c(\tau_A)=c(s)+c(d)\big(P(C=1|R=1)+P(C=1|R=0)\big)$ where the cost of inspecting a snippet is $c(s)=6.3$, the cost of inspecting the document is $c(d)=17.0$ and $P(C=1|R)$ is the probability of clicking on the item given its relevance. In this work, we set $P(C=1|R=1)=1$ and $P(C=1|R=0)=0$, and thus assume the user only inspects relevant items but always pays the cost of examining the snippet regardless of relevance. One could explore more sophisticated click models, e.g., to account for position and trust bias. We leave exploration of these options for future work. While this is a very optimistic setting – we found including mis-clicks on non-relevant items or lower probabilities of clicking relevant items had little impact on which strategy/approach resulted in a higher rate of gain – only that changes lowered the overall rate of gain of all conditions. Instead, our findings show that the relative costs between querying and giving feedback play a much larger role in the choice of strategy/approach (see §6.4). We leave modelling other variations in cost for future work.

Simulated Analysis. To perform the analysis, we first decided on the CS strategy (i.e. FF or FA) and a mixed initiative approach (i.e. QC or QS) the agent would adopt, and then simulated the interaction as follows. For each topic, we assume that a user submits a query to the agent. The user either gives feedback and then examines results, or examines results and then gives feedback depending on the CS strategy. We recorded costs and gains as the number of queries (Q) is varied from 1 to 15, the number of rounds of feedback (F) is varied from 1 to 10, and the number of results assessed (A) is varied from 1 to 20. The total number of conversational turns for the FF strategy is $Q\times(F+A)$ and for FA strategy is $Q\times(F+1)\times A$. The entire process was repeated 20 times for each of the 49 topics, for each strategy and mixed initiative (2x2). To explore the influence of query quality on CS we varied the length of queries during the generation process from 1 to 4. This resulted in over 12 million simulated CS sessions being generated for our analysis[2].

6 RESULTS AND ANALYSIS

To focus the presentation of our results, we will constrain our reports to the interactions within ten minutes (of simulated time) – and unless stated otherwise, present the results when the starting query is of length two (L=2).

6.1 Conversational Trade-offs

To provide some insights into the trade-off between the different conversational turns, in Fig. 3 we have plotted queries (Q) vs assessments (A) for the different rounds of feedback (F) for each conversational strategy and mixed initiative approach. The plots show the interactions for approx. 10 minutes of simulated time. From the plots we can see that as more rounds of feedback are included, the number of queries and the number of assessments decrease – because given a conversational search session of a similar length, taking an F turn comes at the expense of taking an alternative turn. When we compare the Feedback-First strategy (left) to a Feedback-After strategy (right), we can see that Q and A decrease by a greater amount. Recall, though, the subtle difference between conditions: in the FF strategy users perform F rounds of feedback, then examine A items, while in the FA strategy every round of feedback means they examine A items. If we examine the query plots, we can see that as the number of rounds of feedback (F) increases, then the number of queries issued (Q) and the number of assessments performed (A) decreases. For the FA strategy the number of possible queries decreases at a much faster rate as F increases because of the successive rounds of assessing after each round of feedback. Given the space of possible

[1] http://mturk.com

[2] Code and data: https://github.com/i2lab/cikm21-conversational-search-strategies.

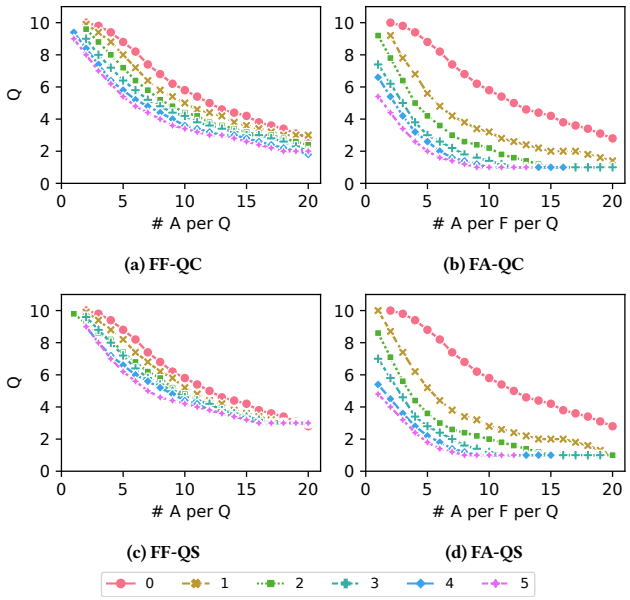


Figure 3: The plots show the trade-off between querying (Q) and assessing (A) for different levels of feedback (F) for the two strategies and two mixed initiatives. Left: Feedback First and Right: Feedback After. Top: Query Clarification and Bot.: Query Suggestion. For clarity, only F≤5 and Q≤10 is shown.

conversational sequences, we now turn our attention to comparing how well the different combinations of strategy and initiative perform. To make our comparisons we will be reporting the rate of gain, because different combinations lead to different session lengths, depending on the conversation turns taken and the relevant items found – also reporting the rate of gain also means we can visualize the performance w.r.t the different number of interactions.

6.2 Conversational Strategy vs. Mixed Initiative

To answer our main research question of how performance is affected by the conversational strategies: Feedback-First (FF) or Feedback-After (FA), vs. the Mixed Initiative (MI) approaches: Query Clarification (QC) or Query Suggestion (QS), we considered how the rate of gain (R) changed as the number of assessments (A) and levels of feedback (F) were varied for different CSA combinations.

First, we can how the different MI approaches perform for the FF strategy by inspected the left hand plots in Fig. 4. The top left plot shows that when query clarifications are employed it leads to substantial increases in the rate of gain over the baseline (i.e. when no feedback is given/provided $F=0$). Additional rounds of feedback increase the rate of gain but with diminishing returns. While as the number of assessments per query (A) increases, the rate of gain also increases. This makes sense, because the investment in improving the query means that more relevant information is surfaced later on. However, when query suggestions are offered with the FF strategy, we observed a similar trend in the bottom left plot, but not as pronounced. In fact, after two rounds of query suggestions the rate of gain starts to decrease such that five iterations results in similar gain to the no feedback baseline.

Table 1: For each combination of CSS x MI and for the no feedback condition (F= 0), the best settings (query Q*, assessment A*, feedback F*, cost C*, and rate of gain R*), on average, to achieve a gain (G) of 1, 5 and 9 when the starting query is length 2.

CSSxMI	G	Q*	A*	F*	C*	R*
No Feedback	1	1	7	0	96	0.014
FF-QC	1	1	5	1	89	0.015
FA-QC	1	1	5	1	123	0.012
FF-QS	1	1	5	1	90	**0.016**
FA-QS	1	1	2	1	80	0.015
No Feedback	5	3	9	0	338	0.014
FF-QC	5	1	11	5	209	**0.023**
FA-QC	5	2	10	1	409	0.013
FF-QS	5	2	10	1	278	0.017
FA-QS	5	1	5	3	256	0.019
No Feedback	9	7	10	0	791	0.011
FF-QC	9	2	11	5	407	**0.022**
FA-QC	9	4	11	1	842	0.011
FF-QS	9	5	10	1	635	0.014
FA-QS	9	1	9	5	541	0.016

The plots on the right hand side of Fig. 4, show how the two mixed initiatives perform under the FA strategy. When query clarifications are offered after the user assesses items, then the rate of gain initially is higher than the baseline until A increases past three result items, then the strategy becomes less effective and the rate of gain drops below baseline (top-right plot). Interestingly, for query suggestions we see that rate of gain is much higher when suggestions are taken afterwards, and it is only if the user assesses more items per round of feedback does the rate of gain start to decrease and tend towards the baseline (bottom-right plot). Here, we see the query suggestions improve the initial query and bring more relevant information back in subsequent assessment turns – but crucially assessing only a few items and then providing feedback leads to the highest rate of gain for this mixed initiative approach.

To directly compare the different combinations of search strategy and mixed initiative, we have plotted the best performing combinations for each in Fig. 5. Here we can see that the FA-QC (with $F=1$) combination is clearly inferior, while the FF-QS (with $F=1$) leads to a small increase over the baseline. More interestingly, we see that FA-QS (with $F=3$) outperforms the baselines and the other two combinations mentioned. However, FF-QC (with $F=5$) leads to the highest rate of gains overall, if the user is willing to assess five or more results per query. This suggests that there is no dominant strategy/approach but two competing combinations. Thus, for the remainder of our analysis, we focus on these two superior combinations: FF-QC ($F=5$), and FA-QS ($F=3$).

Table 1 provides similar insights as described above, where we have listed the configurations that lead to the higher rate of gain for different strategies/approaches for three levels of gain. The table shows that as the amount of gain desired increases then the FF-QC combination results in the highest rate of gain.

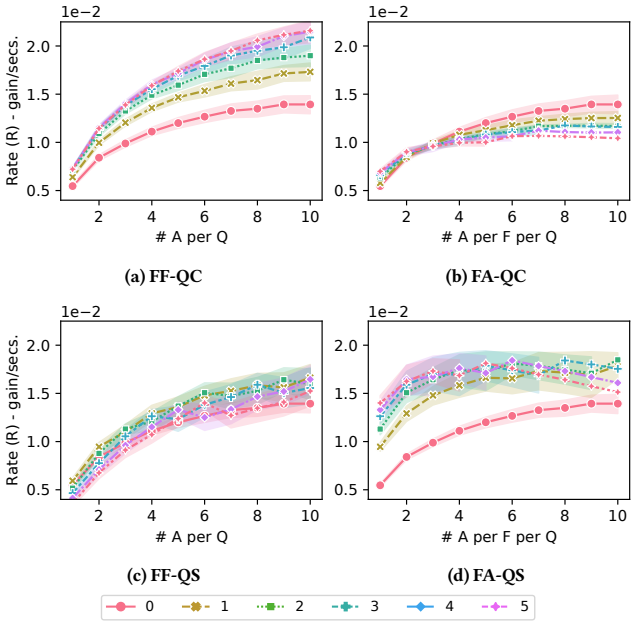


Figure 4: The Rate of Gain (R) by the number of Assessments (A) for different levels of feedback (F). Top: Query Clarification, Bot: Query Suggestion, Left: Feedback First, and Right: Feedback After. For clarity, only F≤5 and Q≤10 is shown.

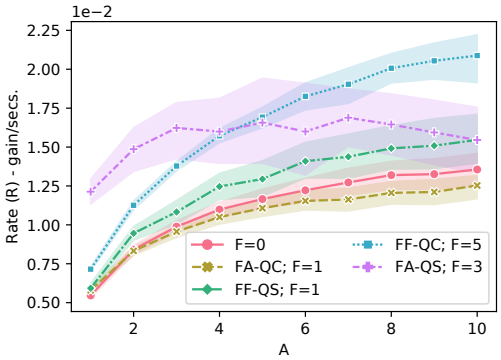


Figure 5: The Rate of Gain (R) by the number of Assessments (A) for the F value that yields the high rate of gain for each combination. FA-QC and FF-QS are clearly inferior to FF-QC and FA-QS, respectively. For clarity, only A≤10 is shown.

6.3 Query Length

To explore our next research question on how the quality of queries influences the choice of strategy and mixed initiative approach taken, we examined how the rate of gain for each combination changed when we varied query length (and consequently the retrieval performance), see. Fig. 6. In the plots, we can see that as query length increases from L=1 to L=4, the rate of gain also increases regardless of condition – which is to be expected [15].

In Fig. 6a we have plotted the rate of gain for FF-QC($F=5$). We can see that the increase in query length leads to a higher rate of gain. However, when compared to the no feedback condition, the rate of gain is similar when A is less than 3, but after providing feedback leads to higher rates of gain (when L=4).

A different story emerges in Fig. 6b (right), where we have plotted the rate of gain for FA-QS ($F=3$). Here, when the length of the starting query is short (L=1), obtaining feedback from query suggestions leads to dramatic improvements in the rate of gain. However, if the starting query is longer (L=4), then the benefit of obtaining feedback via query suggestions leads to smaller increases in the rate of gain. Finally, as the number of assessments a user is willing to make increases, the benefit of feedback rounds from query suggestions diminishes and leads to a similar rate of gain as the no feedback baseline. Essentially, going deeper mitigates conversational interactions.

6.4 Cost of Conversational Turns

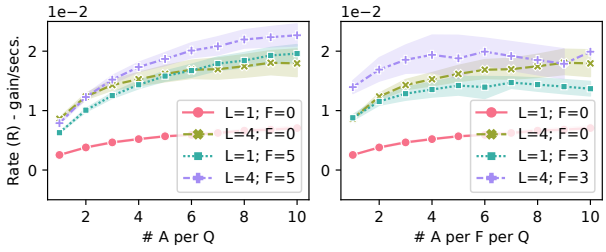
To answer our final question, we explore whether changes to the costs affected the viability of the strategy or approach. So far we have used costs grounded by our user study, but what happens if the average cost associated with the different conversational turns changes? To explore how this might affect behaviours, we varied the cost of feedback, in two ways: (i) by halving it and (ii) by doubling it. For FF-QC, Fig. 7a shows that as the feedback cost decreases it leads to a higher rate of gain. And, as the cost of feedback increases, we see that rate of gain decreases, making the combination less attractive when A is low. For FA-QS, Fig. 7c shows that as feedback cost decreases, the rate of gain also increases, and this makes the combination worthwhile up until A is around 5-6 assessments. But when feedback cost increases, then the viability of the combination diminishes quickly. And, in fact, it eventually becomes worse than no feedback at all.

In terms of changes to query cost, when we reduce the cost of querying, then the rate of gain for the baseline increases (as users reach relevant material sooner) – and so we have updated the baselines in Fig. 7b and 7d. For FF-QC, while previously providing clarifications resulted in a higher rate of gain, the decrease in query costs, means that FF-QC is only effective when A is less than 2, after that point re-querying results in a higher rate of gain. For FA-QS, the suggestions still result in a higher rate of gain than the no feedback baseline – but the difference between the feedback and no feedback condition is considerably reduced – and as A gets larger the differences between becomes smaller and smaller. Essentially, once queries become cheap enough, then issuing a series of queries, even if some are poor, is likely to lead to a higher rate of gain (as previously observed in [34] during session search), rather than trying to refine the query through feedback.

Regardless of the combination, we found that if assessment cost decreases then the rate of gain (R) increases, as less time is needed to extract relevant information, and conversely as the assessment cost increase then the rate of gain decreases. However, changing the cost of assessment didn't impact when to give feedback relative to the number of assessments (plots not shown).

7 DISCUSSION AND FUTURE WORK

In this paper, we have explored how different CS strategies and different MI approaches combine in the context of a text based CSA where we have simulated CS sessions. In order to do so, we first built upon existing models of IIR to develop a model of the CS process which explicitly includes the core conversational concept of mixed initiative. From the model, we derived two different CS strategies, which have been previously observed in conversational settings. While these strategies reduced the evaluation space, it is still largely intractable



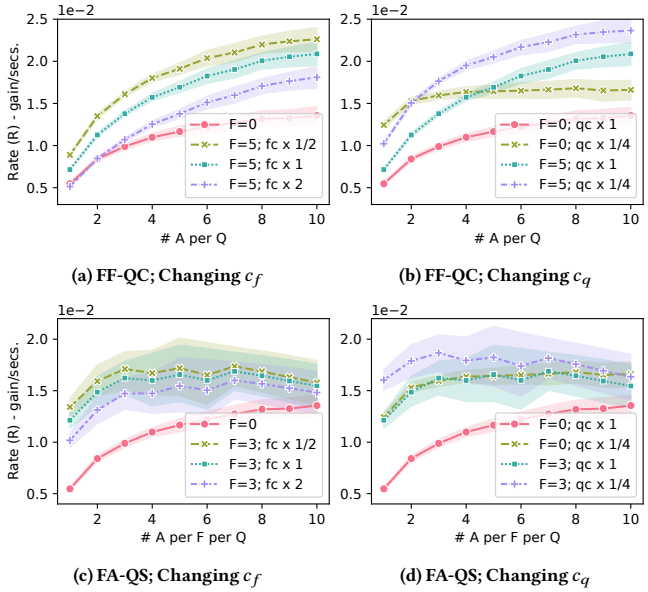
(a) FF-QC; F=5 (b) FA-QS; F=3

Figure 6: The Rate of Gain (R) vs Assessments (A) for query lengths L=1 & L=4 with no feedback (F=0) & the F that yields the higher rate of gain. Left: Feedback First - Query Clarifications, Right: Feedback After - Query Suggestions. For clarity, only A≤10 is shown.

to explore all possible factors and so we focused on the most salient (i.e. number of A, F and Q, given the different conditions).

With respect to the different conditions, we found that there was no dominant CS strategy and MI approach combination. However, we did observe that certain combinations were clearly inferior (e.g. FF-QS and FA-QC), while the choice of combination FA-QC led to higher rates of gain when A was lower, whereas for FF-QC higher rates of were observed when A was greater. Nonetheless, the viability of these combinations was dependent upon the initial query submitted, and the relative cost of giving feedback vs. the cost of querying. In sum, if (i) the length/quality of the initial queries increases, (ii) the cost of giving feedback increases, (iii) the cost of querying decreases, or (iv) a combination of, then providing feedback regardless of combination becomes less beneficial (resulting in a lower rate of gain), and it may even be detrimental where the rate of gain drops below the no feedback / non-conversational baseline. These findings begin to illuminate the complexities and trade-offs involved in conversational search, where it is clear that certain criteria need to be met for conversational search to be beneficial in terms of the rate of gain.

It should be noted, however, that our findings need to be considered in context. We evaluated one particular type of CSA – a chat/text based CSA like those proposed in [6, 32, 67] – where we employed the traditional IR evaluation approach in a conversational setting. We also used simulation based methodology so that we could begin to explore the large evaluation space (which would be near impossible to do so within a user study). Even so, we could only explore a subset of possibilities and focused on pure strategies with fixed rounds of feedback, etc.. Nonetheless, by evaluating and comparing pure strategies combined with the different mixed initiative approaches, we were still able to observe the strengths and weaknesses of the combinations and better understand the different trade-offs. In practice, however, it is clear that a mixture of different strategies and approaches will be employed and required to optimize the rate of the gain experienced during a CS session. As more interaction data becomes available from deployed CSAs it will be possible to instantiate more nuanced interaction models, and to evaluate other conversational search settings where the costs and gains vary. Clearly, this would change the pay-off dynamics associates with the different conversational turns – and so evaluating different types of CSAs that, for example, try to surface relevant information directly would invariably lead to different strategies evolving. We have also made an assumption that CS should be as efficient as possible (following



(a) FF-QC; Changing c_f (b) FF-QC; Changing c_q

(c) FA-QS; Changing c_f (d) FA-QS; Changing c_q

Figure 7: The Rate of Gain (R) vs Assessments (A) as the query cost of querying and the cost of feedback is varied. For clarity, only F≤5 and A≤10 is shown.

Grice's maxims of conversation [24]) and that the users of CSAs and the CSAs will adapt/evolve to maximise the rate of gain (as per Information Foraging Theory [48]). However, it is possible that the conversation itself has additional benefits leading to greater user satisfaction which may not be captured by focusing solely on gain, cost or rate measures. For example, in previous work, they found that asking a relevant clarification increased user satisfaction in voice-only conversations [35] and so this may lead to other trade-offs emerging with satisfaction. Also, in this work, we solely relied on the TREC assessments. In a more realistic experimental setup, one could compute gain based on the amount of useful information given the agent's response. But, these are emerging challenges within the context of CS that need to be addressed through the development of more fine grained test collections before we can evaluate such scenarios.

In this paper, we have shown that the choice of search strategy and mixed initiative depends upon a number of factors: the quality of the starting query, the relative costs of querying vs. giving feedback, the number of results the user is willing to assess, and the amount of gain desired. While more work is needed to explore and investigate the effectiveness of different CSA configurations, the methods they used, and the strategies they employ, we have provided a model and framework for evaluating and simulating the conversational search process in an offline/batch setting. This will enable researchers to explore the complexities and trade-offs of design decisions before developing and deploying them in practice.

Acknowledgements. This work was supported in part by the NWO Innovational Research Incentives Scheme Vidi (016.Vidi.189.039), the NWO Smart Culture - Big Data / Digital Humanities (314-99-301), the H2020-EU.3.4. - SOCIETAL CHALLENGES - Smart, Green And Integrated Transport (814961), and in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

REFERENCES

[1] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeffrey Dalton, and Mikhail Burtsev. 2021. Building and Evaluating Open-Domain Dialogue Corpora with Clarifying Questions. In *EMNLP*.

[2] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail S. Burtsev. 2020. ConvAI3: Generating Clarifying Questions for Open-Domain Dialogue Systems (ClariQ). *CoRR* abs/2009.11352 (2020).

[3] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *SIGIR*. 475–484.

[4] JE Allen, Curry I Guinn, and Eric Horvtz. 1999. Mixed-Initiative Interaction. *IEEE Intell. Syst.* 14, 5 (1999), 14–23.

[5] Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2019. Conversational Search (Dagstuhl Seminar 19461). *Dagstuhl Reports* 9, 11 (2019), 34–83.

[6] Sandeep Avula. 2017. Searchbots: Using Chatbots in Collaborative Information-Seeking Tasks. In *SIGIR*. 1375.

[7] Leif Azzopardi. 2011. The Economics in Interactive Information Retrieval. In *SIGIR*. 15–24.

[8] Leif Azzopardi, Maarten de Rijke, and Krisztian Balog. 2007. Building Simulated Queries for Known-Item Topics: An Analysis using Six European Languages. In *SIGIR*. 455–462.

[9] Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeffery Dalton. 2018. Conceptualizing Agent-Human Interactions during the Conversational Search Process. In *CAIR*.

[10] Leif Azzopardi, Diane Kelly, and Kathy Brennan. 2013. How Query Cost Affects Search Behavior. In *SIGIR*. 23–32.

[11] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the Utility of Search Engine Result Pages: An Information Foraging Based Measure. In *SIGIR*. 605–614.

[12] Leif Azzopardi, Paul Thomas, and Alistair Moffat. 2019. cwl_eval: An Evaluation Tool for Information Retrieval. In *SIGIR*. 1321–1324.

[13] Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. 2012. Time Drives Interaction: Simulating Sessions in Diverse Searching Environments. In *SIGIR*. 105–114.

[14] Nicholas J Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. 1995. Cases, Scripts, and Information-Seeking Strategies: On the Design of Interactive Information Retrieval Systems. *Expert Syst. Appl.* 9, 3 (1995), 379–395.

[15] Nicholas J. Belkin, D. Kelly, G. Kim, J.-Y. Kim, H.-J. Lee, G. Muresan, M.-C. Tang, X.-J. Yuan, and C. Cool. 2003. Query Length in Interactive Information Retrieval. In *SIGIR*. 205–212.

[16] Gabriela Bosetti, Sergio Firmenich, Alejandro Fernández, Marco Winckler, and Gustavo Rossi. 2017. From Search Engines to Augmented Search Services: An End-User Development Approach. In *ICWE*. 115–133.

[17] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017. What Do You Mean Exactly?: Analyzing Clarification Questions in CQA. In *CHIIR*. 345–348.

[18] Ben Carterette. 2011. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In *SIGIR*. 903–912.

[19] W. Bruce Croft. 2019. The Importance of Interaction for Information Retrieval. In *SIGIR*. 1–2.

[20] W. Bruce Croft and R. H. Thompson. 1987. I^3R: A New Approach to the Design of Document Retrieval Systems. *J. Am. Soc. Inf. Sci.* 38, 6 (1987), 389–404.

[21] J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. 2018. Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum* 52, 1 (2018), 34–90.

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

[23] M. Dubiel, M. Halvey, L. Azzopardi, D. Anderson, and S. Daronnat. 2020. Conversational Strategies: Impact on Search Performance in a Goal-Oriented Task Mateusz. In *CAIR*.

[24] Herbert P Grice. 1975. Logic and conversation. In *Speech acts*. Brill, 41–58.

[25] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2020. Guided Transformer: Leveraging Multiple External Sources for Representation Learning in Conversational Search. In *SIGIR*. 1131–1140.

[26] Yulan He and Steve J. Young. 2005. Semantic Processing Using the Hidden Vector State Model. *Comput. Speech Lang.* 19, 1 (2005), 85–106.

[27] Bouke Huurnink, Katja Hofmann, Maarten de Rijke, and Marc Bron. 2010. Validating Query Simulators: An Experiment Using Commercial Searches and Purchases. In *CLEF*. 40–51.

[28] Kalervo Järvelin and Jaana Kekäläinen. 2017. IR Evaluation Methods for Retrieving Highly Relevant Documents. *SIGIR Forum* 51, 2 (2017), 243–250.

[29] Kalervo Järvelin, Susan L. Price, Lois M. L. Delcambre, and Marianne Lykke Nielsen. 2008. Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. In *ECIR*. 4–15.

[30] Michael Johnston, John Chen, Patrick Ehlen, Hyuckchul Jung, Jay Lieske, Aarthi Reddy, Ethan Selfridge, Svetlana Stoyanchev, Brant Vasilieff, and Jay Wilpon. 2014. MVA: The Multimodal Virtual Assistant. In *SIGDIAL*. 257–259.

[31] Chris Jordan, Carolyn R. Watters, and Qigang Gao. 2006. Using Controlled Query Generation to Evaluate Blind Relevance Feedback Algorithms. In *JCDL*. 286–295.

[32] Abhishek Kaushik, Vishal Bhat Ramachandra, and Gareth J.F. Jones. [n. d.]. An Interface for Agent Supported Conversational Search. In *CHIIR*. 452–456.

[33] Heikki Keskustalo, Kalervo Järvelin, and Ari Pirkola. 2008. Evaluating the Effectiveness of Relevance Feedback based on a User Simulation Model: Effects of a user Scenario on Cumulated Gain Value. *Inf. Retr.* 11, 3 (2008), 209–228.

[34] Heikki Keskustalo, Kalervo Järvelin, Ari Pirkola, Tarun Sharma, and Marianne Lykke. 2009. Test Collection-based IR Evaluation Needs Extension Toward Sessions–A Case of Extremely Short Queries. In *AIRS*. 63–74.

[35] Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. 2018. Toward Voice Query Clarification. In *SIGIR*. 1257–1260.

[36] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting User Satisfaction with Intelligent Assistants. In *SIGIR*. 45–54.

[37] Antonios Minas Krasakis, Mohammad Aliannejadi, Nikos Voskarides, and Evangelos Kanoulas. 2020. Analysing the Effect of Clarifying Questions on Document Ranking in Conversational Search. In *ICTIR*. 129–132.

[38] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2021. How Am I Doing?: Evaluating Conversational Search Systems Offline. *ACM Trans. Inf. Syst.* 39, 4 (2021).

[39] Tom Lotze, Stefan Klut, Mohammad Aliannejadi, and Evangelos Kanoulas. 2021. Ranking Clarifying Questions Based on Predicted User Engagement. *CoRR* abs/2103.06192 (2021).

[40] Gary Marchionini. 1997. *Information Seeking in Electronic Environments*. Cambridge University Press, USA.

[41] David Maxwell and Leif Azzopardi. 2016. Agents, simulated users and humans: An analysis of performance and behaviour. In *CIKM*. 731–740.

[42] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. 2015. Searching and Stopping: An Analysis of Stopping Rules and Strategies. In *CIKM*. 313–322.

[43] Michael F. McTear. 2002. Spoken Dialogue Technology: Enabling the Conversational User Interface. *ACM Comput. Surv.* 34, 1 (2002), 90–169.

[44] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus Models: What Observation Tells us about Effectiveness Metrics. In *CIKM*. 659–668.

[45] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. on Inf. Sys.* 27, 1 (2008), 2:1–2:27.

[46] R.N. Oddy. 1977. Information Retrieval Through Man-Machine Dialogue. *J. Documentation* 33, 1 (1977), 1–14.

[47] Roberto Pieraccini, Evelyne Tzoukermann, Z. Gorelov, Jean-Luc Gauvain, Esther Levin, Chin-Hui Lee, and Jay Wilpon. 1992. A Speech Understanding System based on Statistical Representation of Semantics. In *ICASSP*. 193–196.

[48] Peter Pirolli and Stuart K. Card. 1995. Information Foraging in Information Access Environments. In *CHI*. 51–58.

[49] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *CHIIR*. 117–126.

[50] Sudha Rao and Hal Daumé. 2018. Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information. In *ACL (1)*. 2736–2745.

[51] Nuzhah Gooda Sahib, Anastasios Tombros, and Tony Stockman. 2012. A Comparative Analysis of the Information-Seeking Behavior of Visually Impaired and Sighted Searchers Nuzhah. *J. Assoc. Inf. Sci. Technol.* 63, 2 (2012), 377–391.

[52] Ivan Sekulic, Mohammad Aliannejadi, and Fabio Crestani. 2021. User Engagement Prediction for Clarification in Search. In *ECIR*. 619–633.

[53] Stefan Sitter and Adelheit Stein. 1992. Modeling the Illocutionary Aspects of Information-Seeking Dialogues. *Inf. Process. Manag.* 28, 2 (1992), 165–180.

[54] Mark D. Smucker and Charles L. A. Clarke. 2012. Time-based Calibration of Effectiveness Measures. In *SIGIR*. 95–104.

[55] Damiano Spina, Johanne R. Trippas, Lawrence Cavedon, and Mark Sanderson. 2017. Extracting Audio Summaries to Support Effective Spoken Document Search. *J. Assoc. Inf. Sci. Technol.* 68, 9 (2017), 2101–2115.

[56] Paul Thomas, Alistair Moffat, Peter Bailey, and Falk Scholer. 2014. Modeling Decision Points in User Search Behavior. In *IIiX*. 239–242.

[57] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search. In *CHIIR*. 32–41.

[58] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search: Perspective Paper. In *CHIIR*. 32–41.

[59] Johanne R. Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a Model for Spoken Conversational Search. *Inf. Process. Manag.* 57, 2 (2020), 102162.

[60] Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. 2019. QRFA: A Data-Driven Model of Information Seeking Dialogues. In *ECIR*. 541–557.

[61] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. 2017. Exploring Conversational Search With Humans, Assistants, and Wizards. In *CHI Extended Abstracts*. 2187–2193.

[62] Somin Wadhwa and Hamed Zamani. 2021. Towards System-Initiative Conversational Information Seeking. In *DESIRES*.

[63] Marilyn A. Walker, Rebecca J. Passonneau, and Julie E. Boland. 2001. Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems. In *ACL*. 515–522.
[64] Muuo Wambua, Stefania Raimondo, Jennifer Boger, Jan Polgar, Hamidreza Chinaei, and Frank Rudzicz. 2018. Interactive Search through Iterative Refinement. In *CAIR*.
[65] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System. In *SIGIR*. 55–64.
[66] Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response Ranking with Deep Matching Networks and External Knowledge in Information-seeking Conversation Systems. In *SIGIR*. 245–254.
[67] Hamed Zamani and Nick Craswell. 2020. Macaw: An Extensible Conversational Information Seeking Platform. In *SIGIR*. 2193–2196.
[68] Hamed Zamani, Susan T. Dumais, Nick Craswell, Paul N. Bennett, and Gord Lueck. 2020. Generating Clarifying Questions for Information Retrieval. In *WWW*. 418–428.
[69] Hamed Zamani, Gord Lueck, Everest Chen, Rodolfo Quispe, Flint Luu, and Nick Craswell. 2020. MIMICS: A Large-Scale Data Collection for Search Clarification. In *CIKM*. 3189–3196.
[70] Hamed Zamani, Bhaskar Mitra, Everest Chen, Gord Lueck, Fernando Diaz, Paul N. Bennett, Nick Craswell, and Susan T. Dumais. 2020. Analyzing and Learning from User Interactions for Search Clarification. In *SIGIR*. 1181–1190.
[71] Shuo Zhang and Krisztian Balog. 2020. Evaluating Conversational Recommender Systems via User Simulation. In *KDD*. ACM, 1512–1520.
[72] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *CIKM*. 177–186.