Information Retrieval Journal (2022) 25:209–238
https://doi.org/10.1007/s10791-022-09404-z

SPECIAL ISSUE ON ECIR 2021

CoSearcher: studying the effectiveness of conversational search refinement and clarification through user simulation

Alexandre Salle[1] · **Shervin Malmasi**[2] · **Oleg Rokhlenko**[2] · **Eugene Agichtein**[2,3]

Received: 25 June 2021 / Accepted: 5 February 2022 / Published online: 10 March 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract

A key application of conversational search is refining a user's search intent by asking a series of clarification questions, aiming to improve the relevance of search results. Training and evaluating such conversational systems currently requires human participation, making it infeasible to examine a wide range of user behaviors. To support robust training/evaluation of such systems, we propose a simulation framework called CoSearcher Information about code/resources available at https://github.com/amzn/cosearcher that includes a parameterized user simulator controlling key behavioral factors like cooperativeness and patience. To evaluate our approach, we use both a standard conversational query clarification benchmark and develop an extended dataset using query suggestions from a popular Web search engine as a source of additional refinement candidates. Using these datasets, we investigate the impact of a variety of conditions on search refinement and clarification effectiveness over a wide range of user behaviors, semantic policies, and dynamic facet generation. Our results quantify the effects of user behavior variation, and identify conditions required for conversational search refinement and clarification to be effective. This paper is an extended version of our previous work, and includes new experimental results for comparing semantic similarity ranking strategies for facets, using enhanced representations of facets, learning from negative user responses, among other new results and more detailed experimental descriptions.

Keywords Conversational search · user simulation for conversational search · conversational query clarification

---

✉ Alexandre Salle
alex@alexsalle.com

Shervin Malmasi
malmasi@amazon.com

Oleg Rokhlenko
olegro@amazon.com

Eugene Agichtein
eugeneag@amazon.com

[1] Institute of Informatics, UFRGS, Porto Alegre, Brazil

[2] Amazon, Seattle, USA

[3] Emory University, Atlanta, USA

1 Introduction

As personalized information agents become ubiquitous, people increasingly expect to engage them in information-seeking dialogues, instead of having to formulate a precise query. A user's query to a search system often under-specifies the search intent (or *facet* of the information need, as is often referred to in the literature). A conversational system could elicit a more precise information need from a user, by asking her a series of *clarification questions* to narrow down the set of possible intents, ultimately to improve the relevance of the search results. Recent work (Aliannejadi et al. 2019) has shown the theoretical value of obtaining answers to such clarification questions to improve the final retrieval.

Search refinement is also critical in practice, namely for voice-based agents like Alexa or Siri. Generally, only a small number of results can be returned to the user via a voice modality, and matching the correct search intent is critical (Trippas et al. 2018). Furthermore, in applications such as e-commerce, successive search refinement is natural for narrowing down the choice of products using facets of the target item.

Unfortunately, conversational search refinement is highly challenging due to the reliance on human participation for developing, training, and evaluating system variants or parameters. Furthermore, some users may not be willing to provide additional information to the search system after the initial request, while others might be willing to collaborate with the system by engaging in a dialogue. To address these issues, training and evaluating such conversational systems with a large number of users or crowd workers has been the dominant strategy. This has two shortcomings: (1) High cost, especially when different variations of a search system must be tested; (2) The pool of human participants might not be representative of future participants, who might, for example, be less *cooperative* and/or *patient*. A key contribution of this paper is re-examining the underlying assumptions of conversational search, to quantify the effects of user *cooperativeness*, i.e., willingness to provide clarification information, and user *patience*, i.e., willingness to engage in a long dialogue with a search system. We quantify this intuition by developing a simple, yet powerful, stochastic user simulator CoSearcher for conversational search refinement, and investigate the implications of cooperativeness and patience of users by extensive simulation experiments that would not be feasible with human participants. This proposed simulator provides a way to better understand the effectiveness and limitations of the a given conversational search system, for a wider range of potential future users, without degrading their search experience.

Although our user simulator has only two parameters (cooperativeness and patience), and might thus be deemed *unrealistically simple* because humans have far more "variables", we argue that these are the characteristics directly responsible for the user behavior *observable* by a search system, and thus form an acceptable proxy for scalable evaluation of a conversational search system under a wide range of *realistic* configurations of complex latent search behavior factors.

This paper is an extension of our previous work (Salle et al. 2021). Compared to the earlier edition, this revision is not subject to the same space constraints and includes extended details and discussion about the experimental settings, as well as numerous new and insightful experimental results. Our contributions, including newly added results in this edition, consist of:

- We systematically investigate the task of conversational search intent clarification, comparing facet identification and ranking methods, for both static and dynamically generated candidate intents.
- We present a simple yet powerful conversational search simulator, CoSearcher, with key parameters of cooperativeness and patience, to enable systematic and scalable experimentation with conversational search refinement (Sect. 3.4).
- Using CoSearcher, we for the first time demonstrate using extensive simulation experiments, that modeling cooperation and patience of the searcher is fundamental for the success of conversational search, and identify the conditions where conversational search can be effective. This required evaluating results for hundreds of thousands parameter combinations for conversational experiments, which would not be feasible with human participants (Sect. 5).
- (Extended) Comparison between different strategies for semantic similarity ranking of candidate facets given dialogue context (Sect. 3.2).
- (Extended) Enhanced representation of candidate facets which incorporate information from the IR system, to give facet rankers more context about a facet (Sects. 3.2 and 5.4).
- (Extended) A facet ranker that accounts for uninformative responses: learning from a simple "No" (Sects. 3.2 and 5.5).
- (Extended) User dynamics, where patience and cooperativeness are functions of conversation length (Sects. 3.4.5 and 5.6).
- (Extended) In addition to measuring downstream IR performance, we perform direct evaluation of intent identification (Sects. 4.1 and 5.1).

Broadly, our work adds to the growing evidence of the importance of engaging in conversations with users to improve search performance, and provides the critical building block, the CoSearcher user simulator, for scalable evaluation of a given conversational search system under a variety of conditions. Next, we briefly review related work to place our contributions in context.

2 Related work

There is a large body of work in NLP and IR that addresses conversational systems (Weizenbaum 1966; Croft et al. 1987; Belkin et al. 1995; Young 2000). Advances in NLP and IR in the last few years have also been accompanied by a surge in research of conversational systems.

Agenda-based methods have been successfully used in dialogue management systems (Rudnicky and Xu 1999), and have proven particularly useful for task-oriented conversations where the user needs to complete a specific task such as making a reservation (Shah et al. 2016). Agenda-based user simulators have been deployed to bootstrap the training of such systems (Schatzmann et al. 2007). These methods factorize the user state into an agenda and a goal (Keizer et al. 2010), where the goal represents the task that the user wishes to achieve along with additional information about constraints and additional information sought by the user (Keizer et al. 2010). The agenda, represented by a stack data structure, consists of the pending dialogue acts that the user must perform. State transition models are used to update the agenda and goals as a conversation progresses. While this framework is suitable for task-oriented use cases with specific dialogue structures, they are less suited for open-domain information seeking dialogues, or cases where the user intent is ambiguous or exploratory in nature. Accordingly, our work does not attempt to model search refinement in an agenda-based manner.

Within the sub-field, understanding user behavior in conversational search is an important research direction: (Kiesel et al. 2018) find that interactivity between user and system agent is important to clarify the information need, and (Trippas et al. 2018) find that users do not mind, and can even enjoy being asked for clarification. However, neither work explicitly models the results for use in simulations. Additionally, (Sun and Zhang 2018; Zhang and Balog 2020) perform user simulation, but unlike our work focus solely on recommender systems and use a fixed user model. For chat systems and task-completion dialogues, developing user simulators has also been shown to be an effective way to reduce the required training data (Chandramohan et al. 2011; El Asri et al. 2016; Li et al. 2016), which inspired our efforts to adapt that general idea to search-oriented conversational systems. To the best of our knowledge, our paper is the first to propose a user simulator for *conversational search.*

A parallel line of work focuses on learning to ask clarification questions to fill in missing information (Mostafazadeh et al. 2016; Papangelis et al. 2017; Rao and Daume III 2018, 2019; Zamani et al. 2020). None of these, however, focus on intent refinement, nor do they make use of a variable user model for evaluation. Another related direction is faceted search, where a user reacts to the proposed facets to refine the information need or to restrict or change the set of results (Hearst et al. 2002; Yee et al. 2003; Hearst 2006; Tunkelang 2009; Kules et al. 2009; Fagan 2010; Kotov and Zhai 2010; Vandic et al. 2017).

Most similar to our work is that of Aliannejadi et al. (2019), which uses human annotation of clarification questions which are then used within an IR system to evaluate how they could help retrieval performance. They release the resulting dataset, called Qulac, which we use as the basis of our paper. Qulac makes use of the 198 topics, corresponding facets and relevance judgements from the TREC09-12 diversity track (Clarke et al. 2009, 2012), supplemented by crowdsourced human clarification questions and answers for each facet. For each topic, there are multiple human generated clarification questions corresponding to the each of the topic's facets, and for each *(topic,facet,question)* triple, there is an human generated answer where the human assumes the role of a searcher looking for the facet and answers the given question. Very recently, the Qulac dataset was expanded into ClariQ (Aliannejadi et al. 2020) via the addition of new data, including synthetic multi-turn conversations. Our work is evaluated using the original Qulac dataset which is sufficient to investigate the research questions posed here. Our other, expanded facet dataset, constructed from Bing query suggestions, complements Qulac and allows us to investigate additional challenges that arise with numerous query facets.

The Qulac paper (Aliannejadi et al. 2019) presents the Neural Question Selection (NeuQS) model, which given a conversation context (a series of questions/answers), selects the next question to ask from a candidate question database (the Qulac dataset). The human answer is then used to simulate the end of the conversation and the whole conversation is used as input to a query-likelihood IR system to evaluate the utility of the clarification question.

We differ from this work by focusing on intent refinement — the goal of our system is to narrow a set of candidate intents down to a specific intent — and by creating a user model and simulator, CoSearcher, which allows us to evaluate the utility of clarification questions not just on a specific set of human annotators, but rather a large set of simulated parameterized users. CoSearcher also enables the possibility of scalable training of conversational search systems, optimized for different types of users, and supporting sophisticated, yet data hungry, end-to-end deep learning approaches for conversational search, e.g., via Reinforcement Learning (Wen et al. 2017; Bordes et al. 2016).

In an effort to study the effect of clarification questions and answers on document ranking, (Krasakis et al. 2020) segment human answers in the Qulac dataset into positive ("Yes") and negative ("No") responses, finding that multi-word negative responses, which correspond to the *informative "No"* in our work, significantly improve retrieval performance. Our work also studies the effects of asking clarification questions on retrieval, but rather than looking at individual questions and answers in isolation (single turn dialogue), it does so in the context of complete dialogues generated through parameterizable user simulation.

Using a sophisticated Transformer document ranking model that incorporates clarification questions/answers and other external information sources, (Hashemi et al. 2020) also observe improved retrieval performance due to clarification questions and answers for both single and multi-turn dialogue. However, their multi-turn dialogues are not based on user simulation like our work, but on simple concatenation and shuffling of human data as in Aliannejadi et al. (2019).

3 Modeling conversational search intent refinement through user simulation

We now overview the conversational search intent refinement setting, following the recent formulation in (Aliannejadi et al. 2019), and our simulation-based approach for investigating this topic.

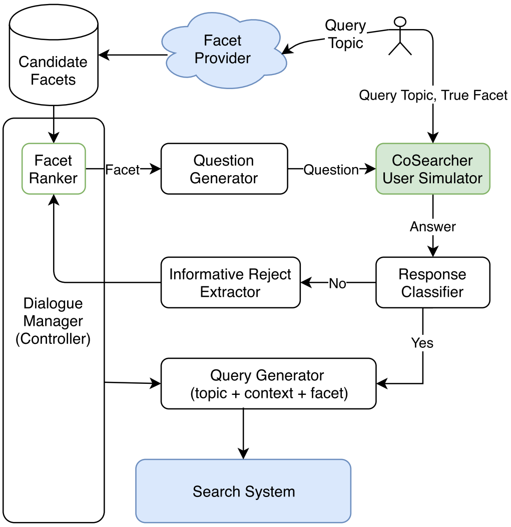
3.1 Problem setting: conversational search refinement

Often, a searcher (user) provides an under-specified query to the search system, which may reflect multiple information needs, or different facets of the same intent. A conversational search refinement system attempts to pinpoint the user's search intent via a series of *clarification questions*, which the Searcher can *choose* to answer cooperatively (by volunteering additional information about their intent), lazily ("yes/no") or not respond to the system at all, e.g., if the Searcher ran out of time or patience.

There are a number of use cases where we may not be able to rely on the user providing further information. In speech-based systems, such as voice assistants, shorter yes/no interactions are often associated with lower perceived levels of friction by users. Similarly, some user interfaces may elicit yes/no responses from users instead of requesting direct input. Yet in other cases the user may have an exploratory intent, seeking to discover a facet that was not known in advance.

After each turn, the search system may choose to ask additional clarification questions, or return search results, or both. An example conversational search dialogue is shown in Fig. 1b, for the initial under-specified query, where the system follows with a sequence of clarification questions to generate the result ranking using the expanded/refined query.

Formally, we assume that the searcher has an information need (topic) t (i.e., the initial search query), and a true information need facet or aspect f_t, which the system has to infer to properly rank the search results. We also assume that candidate facets C for the topic t is either known (e.g., from a knowledge base if the query is an entity), or can be dynamically generated (e.g., from query refinement logs of a search engine, or from popular entity



(a)

USER: Korean language *(the topic)*

BOT: **Are you looking to find resources on translating Korean to other languages?**
COSEARCHER: no
BOT: **Ok, are you looking to find resources on learning and teaching Korean?**
COSEARCHER: no
BOT: **Ok, are you looking to find a Korean/English bilingual dictionary?**
COSEARCHER: no i just want to know a few common korean phrases
BOT: **Are you looking for what are some common Korean phrases?**
COSEARCHER: YES

True intent: “What are some common Korean phrases?”

(b)

Fig. 1 **a** System overview, illustrating CoSearcher instantiated with (topic, intent facet), and a Facet Provider, which provides candidate facets that the search refinement system uses to converse with the CoSearcher to identify the intended facet; **b**: An actual simulated conversation with a partially cooperative CoSearcher instance

attributes). The goal of the search system, then, is to identify the intended topic facet f_t by asking clarification questions, and return a list of results relevant to f_t. Specifically, the search system picks the first candidate facet $c \in C$ and asks a clarification question: “Are you looking for c?”.

The user can respond with either “Yes” or “No”. If the answer is “Yes”, the agent stops, accepting c as its best guess for the searcher’s true information need. If the answer is “No”, the agent pops c from the list of candidate facets and adds c to the list of *dead facets* DB_{dead}. If the user’s “No” is *informative*, we add the answer to the *Informative No* list DB_{info}. An informative answer is one where the user volunteers additional information rather than simply providing a binary “Yes”/“No” response. For example, the answer “no i just want to know a few common korean phrases” in Fig. 1b is informative. Table 1 tracks the state of DB_{dead} and DB_{info} for the dialog in Fig. 1b. Candidates facets are then *re-ranked*, as described below, and this process repeated until either there are no more candidate facets or the user’s patience runs out.

Note that in our setup, we choose to model neutral responses (when the proposed facet is related to intended facet but not quite the same) as “No”, since the intended facet has not yet been identified.

3.2 Candidate facet ranking strategies

When many candidate refinement facets exist, choosing the best ones to clarify first is important to avoid taxing the user’s patience and effort. We consider three facet ranking strategies:

Rand A random baseline that orders facets randomly.

Table 1 Tracking the state of DB_{dead} and DB_{info} for the dialog in Fig. 1b

System Question	Answer	DB_{dead}	DB_{info}
Are you looking to find resources on translating Korean to other languages?	no	{Find resources on translating Korean to other languages.}	{ }
Ok, are you looking to find resources on learning and teaching Korean?	no	{Find resources on translating Korean to other languages, Find resources on learning and teaching Korean }	{ }
Ok, are you looking to find a Korean/English bilingual dictionary?	no i just want to know a few common korean phrase	{Find resources on translating Korean to other languages, Find resources on learning and teaching Korean, Find a Korean/English bilingual dictionary }	{ i just want to know a few common korean phrases }

Table 2 Qulac facet ranking using different sentence embedding methods

	P@1			MRR		
Model	Train	Dev	Test	Train	Dev	Test
Unsupervised						
fastText Crawl	0.8137	**0.8330**	0.8060	0.8924	**0.8996**	0.8851
Google News	0.8133	0.8003	0.7868	0.8919	0.8822	0.8742
GloVe 840B	0.6977	0.7504	0.6823	0.8170	0.8487	0.8069
LexVec n-gram	**0.8329**	**0.8330**	**0.8072**	**0.9027**	0.8990	**0.8857**
Supervised						
SBERT + BERT-large	**0.9812**	0.8554	0.8553	**0.9900**	0.9112	0.9118
CE + BERT-large	0.9580	**0.8967**	**0.9165**	0.9759	**0.9375**	**0.9498**

Unsupervised models are mean bag-of-vectors based on the corresponding word embeddings, and supervised models are fine-tuned on the labeled context-facet dataset

Sim A semantic similarity strategy, which assigns a score for each candidate facet c (for example, “What are some common Korean Phrases?”) by computing the similarity between each candidate facet and the conversation context:

$$\mathrm{score}(c, db) = \sum_{s \in db} cos(rep(c), rep(s))/|db| \tag{1}$$

where for this ranking strategy db is DB_{info}, and $rep(s)$ is the embedding of sentence s. If the conversation context is empty—which can happen if the conversation has just started or the user has returned no informative answers – facets are ranked randomly. We performed facet ranking evaluation using *unsupervised* sentence embeddings based on mean bag-of-vectors (BoV) using various word embeddings (GloVe (Pennington et al. 2014), Google News (Mikolov et al. 2013), fastText (Bojanowski et al. 2017; Mikolov et al. 2018), LexVec (Salle et al. 2016; Salle and Villavicencio 2018)) and the *supervised* Sentence-BERT (SBERT) model (Reimers and Gurevych 2019) based on Transformers. Given a target facet, an informative answer (db with a single element in Eq. (1)), and a set of candidate facets for the target facet’s topic (including the target facet),[1] $\mathrm{score}(c, db)$ is computed between this informative answer (db) and each candidate facet (c), measuring precision-at-1 (P@1) and mean reciprocal rank (MRR) of the target facet within the sorted list of candidates. In other words, this evaluates whether a facet ranking strategy (and underlying rep model) can leverage user responses to identify target facets. Unsupervised models do not make use of this labeled dataset, whereas supervised models use it in fine-tuning the underlying loss function to output the correct binary label. We only evaluate on Qulac facets since we do not have labeled data for Bing facets. We use the same topic splits as in Sect. 4.2.

Results are shown in Table 2. The LexVec n-gram embeddings outperform all other unsupervised embeddings. There is a small gap in performance between its P@1 and that of the SBERT model, and yet a smaller gap in MRR. As an upper bound, we also evaluate

---

[1] An alternative would be to consider all facets, not just from a single topic but from all topics, as candidates. In our use case, this does not make sense since our application is one where the user provides a topic. We assume that the topic is accurate and do not consider facets beyond it. In future work we could consider other scenarios, e.g. where the user provides no initial topic, there is uncertainty about the topic, or we begin with a very broad topic which may share facets with other topics.

a supervised cross encoder (CE) BERT model which does not generate sentence embeddings, but is instead trained to calculate score(.) directly from an input context-facet pair. This model has a larger performance gap compared to the BoV models.

If the computation cost of score(.) for a single context-facet pair is r, the number of turns in a dialogue is p, and the set of candidate facets is C, facet ranking over a whole dialogue incurs cost $O(rp|C|)$. This makes the CE model prohibitive since it requires running the entire BERT model (high r) for every context-facet pair and performance is a crucial aspect of simulation. In contrast, sentence embedding models perform only $O(r(p + |C|))$ computations to generate sentence embeddings for contexts and candidate facets, making $O(rp|C|)$ at ranking time cheap since r is but a cosine between vectors. Note that "Yes"/"No" classification only has cost $O(rp)$ (CoSearcher answers at most p questions in a dialogue), thus making the BERT model which has high r suitable for that usage, but not as a CE for facet ranking.

For this reason we focus on sentence embedding for facet ranking, and in particular use the LexVec n-gram BoV as our rep($\cdot$) (in Eq. (1)) for all subsequent experiments in this paper because (1) the performance gap to the SBERT model is small, in particular for MRR (2) we hypothesize that being unsupervised will help it generalize to Bing facets which look less like natural language than do Qulac facets using in the SBERT training set.

We also experiment with *enhanced facet representations*, where the input to $rep(\cdot)$ is the concatenation of the facet and the top-10 page titles when topic + facet is given as a query to the IR system. Table 5 shows some examples of enhanced facets. The intuition is that the additional context can help the semantic matching system, especially for shorter facets (Sect.3.3).

PNSim A positive and negative semantic similarity strategy:

$$\begin{aligned} posnegscore(f) =& \alpha \cdot score(f, DB_{\text{info}}) \\ & - (1 - \alpha) \cdot score(f, DB_{\text{dead}}) \end{aligned} \tag{2}$$

where DB_{info} is the set of informative "No" responses received from the user, and DB_{dead} is the set of proposed facets to which the user has responded with a "No". In this work, we only perform experiments with PNSim where user cooperativeness is set to 0. In this case, DB_{info} is always empty, and so any value of $\alpha < 1$ is equivalent (we set $\alpha = 0$). However, in future work on learning from simple "No" responses, this value should be tuned.

3.3 Dynamic facet generation

The previous state of the art approach—NeuQS (Aliannejadi et al. 2019) and similar methods — require knowing *a priori* a set of candidate questions and answers for a given facet, which is not realistic for most search topics or information needs. We now investigate how to abstract and generalize this approach to *dynamically generated a set of candidate facets* using a *facet provider*.

One example of such a facet provider is a search engine query suggestion mechanism, e.g., the Bing search engine Autosuggest available via an API[2] which, given an initial query, returns a set of 8 query completions. The Qulac topic is used as the initial query to the Autosuggest API, and the set of completions returned by this API are used as candidate

---

[2] https://azure.microsoft.com/en-us/services/cognitive-services/autosuggest/

facets. We experiment with two variants of this facet provider: (1) **S-Bing**, which uses a single call to Autosuggest, resulting in at most 8 facets per topic and (2) the superset **B-Bing**, which makes makes 26 additional calls for a topic by appending to the query each letter of the alphabet, resulting in $8 + 8 * 26 = 216$ candidate facets per topic. Note that this can be seen as a breadth-first-search of the Autosuggest API, where nodes are expanded by this letter-appending technique. Though we restrict ourselves to a single level, this search can go deeper, to allow for more in-depth and comprehensive exploration of the user intent refinement task, using a simulator described next.

3.4 CoSearcher: user simulator for conversational search

Our core contribution, CoSearcher, is the parameterized modeling of conversation search system users. The model is general, and is applicable to a broad set of conversational search tasks. It has two key components: (1) User Intent: a task-specific representation of the user's goals; and (2) User Parameters: values representing levels of cooperativeness and patience.

The overarching purpose of CoSearcher is to simulate varied user behavior, and evaluate systems under different conditions. The simulator relies on ground truth data to generate the user interactions; it does not replace the data needed to train or test a conversational system.

3.4.1 User intent

In our search intent refinement use case, the goal is a search intent known only to the user, and the goal of a system is to discover this intent through a series of questions. The simulator returns a Boolean response depending on whether the question matches the intent.

Formally, the user model has a function $g(topic, intent, question)$ that returns a similarity score between the topic/intent and the question. The "Yes"/"No" is then decided using a threshold that can be chosen using downstream performance, or intrinsically evaluated if there is labeled "Yes"/"No" data.

3.4.2 CoSearcher behavior parameters

CoSearcher has two core parameters: cooperativeness and patience. Cooperativeness is a key user characteristic which has been *assumed* by conversational systems, and represents the users willingness to help the agent. Patience, representing the maximum number of interactions a user is willing to have with the conversational system, is based on the observation that user willingness to examine results diminishes over time (Järvelin et al. 2002). Manipulating these two parameters via simulations enables us to expose the direct relationship between these key user behavior factors and conversational system results.

3.4.3 Cooperativeness

A user of a conversational system can be more cooperative by providing extra information (an *informative answer*) in addition to a minimal response. The informative answer can be task agnostic, by leaking the score from $g(\cdot)$ via answers such as "No, not even close"/"No, but you're close", or directly leaking *intent* (with or without rewording), such as "No, I'm looking for *$intent*". We define Cooperativeness as a Bernoulli random variable where p is the level of cooperativeness (i.e., a user with cooperativeness=0 only gives boolean answers, and a user with cooperativeness=1 always gives informative answers). If a Bernoulli trial returns 1, CoSearcher returns an informative answer, else it simply returns "No". Task-specific informative responses can be provided by making use of labeled data from human annotated informative answers, or by training a generative model using this data.

3.4.4 Patience

A user also has a patience level p, such that the conversation ends when the conversation exceeds a predefined number of turns p. This corresponds to the maximum amount of effort this user is willing to expand by interacting with the search system.

In this work, we explore a wide range of these values through simulation, thus exhaustively testing the effect of user behavior on the success of a conversational search refinement system.

3.4.5 User dynamics

In the simplest case we can assume that user cooperativeness remains static during the conversation. CoSearcher also allows for the cooperativeness parameter to be updated in order to capture user dynamics such as increasing or decreasing cooperation, which can be impacted by the length of the interaction, response latency, answer quality, *etc*.

In our reported intent refinement study we experiment with two variations of our user model that dynamically adjust the level of cooperativeness as the conversation is lengthened: (i) Increasing cooperativeness, where cooperativeness at turn t is $p_{\text{inc}}(t) = p(0) \cdot log_2(t + 1)$; and (ii) Decreasing cooperativeness, where $p_{\text{dec}}(t) = p(0)/log_2(t + 1)$. $p(0)$ is the initial cooperativeness level. We chose the logarithmic decay of cooperativeness inspired by the numerous empirical studies of Web search examination, where users' propensity to examine results further down the rank lists decays similarly, and is the basis of the nDCG measure (Järvelin et al. 2002), and the benefit of a relevant result is attenuated by the effort (time) required to read it (Smucker and Clarke 2012). We conjecture that users' willingness to respond to system questions will degrade in the similar fashion, where the number of conversation turns provides a similar indication of effort.

Other more complex dynamics of search user behavior have been proposed, notably those using economic models of effort and decision making (Azzopardi and Zuccon 2016), and models of dynamic user behavior changes during the search under various constraints, such as time pressure (Crescenzi et al. 2016). Nevertheless, the variations we study capture the main possible trends in user's willingness to "assist" a search system: constant, decreasing, and increasing cooperativeness; we plan to study more complex models, e.g., where the cooperativeness might be conditioned on the result quality, and is a promising area of future improvements to the supported CoSearcher behavior models.

4 Experimental setup

4.1 Resources and evaluation

Our study uses only publicly available resources. The main dataset used is the previously described Qulac benchmark dataset (Aliannejadi et al. 2019). Our "Yes"/"No" classifier fine-tunes the BERT-large uncased model from (Devlin et al. 2019). The similarity rankers use the LexVec (Salle and Villavicencio 2018) Common Crawl n-gram embeddings.[3] The IR search system is the same query-likelihood model used by (Aliannejadi et al. 2019)[4] indexed on ClueWeb09b.

Our evaluation has two parts: (1) *intent identification*: evaluating the intrinsic accuracy of matching a clarification question to a known user intent (using human annotations as ground truth), and (2) *downstream search performance*: evaluating the relevance of the search results to the underlying user intent.

Intent Identification we measure two distinct *success rates (accuracies)*: (1) Subjective (Subj) accuracy: Getting a “Yes” from the user *simulator* before patience runs out. We term this *subjective* because the simulator is accepting a facet proposed by the agent which might not match the correct facet. (2) Real accuracy: Getting a “Yes” from the user simulator to the *correct* facet before patience runs out. We note that real accuracy can only be measured for Qulac agent facets —not Bing—because it requires a matching between agent facets and the user model’s Qulac facet. We treat this as a classification task of matching a clarification question to the exact correct underlying user intent, and evaluate it using precision, recall, and the F1 score.

When both agent and user simulator use Qulac facets, the gap between subjective and real accuracy tells us how well our “Yes”/“No” classifier is performing in a full conversation context. If this gap is small, we can be more confident in the subjective accuracy as a proxy for real accuracy when the agent uses Bing facets and the user simulator Qulac facets, a scenario in which it is impossible to compute real accuracy since we do not have a human labeled mapping between Bing and Qulac facets.

Downstream Search Performance We measure the success of a dialogue by evaluating the relevance of the results retrieved using the enhanced query with identified user intent (topic + facet), using standard IR evaluation metrics: Mean Reciprocal Rank (MRR), Precision@k (P@k), and normalized Discounted Cumulative Gain@k (nDCG@k).

4.2 Conversational intent refinement simulations

We now describe the concrete implementation of CoSearcher used to evaluate a conversational search refinement system under a variety of conditions. Figure 1 shows the flow of an experiment for a given query topic and (hidden) true intent facet. For these experiments, the user intent is represented as a combination of topic and true intent facet, as described in Sect. 3.

To simulate cooperative users, we need a mechanism to provide informative answers that incorporate feedback. We achieve this through implementing for function $g(.)$ a simple heuristic to allow us to use a dataset such as Qulac (described above) to train CoSearcher. Specifically, we automatically label each instance *(topic, facet, question, answer)* in the Qulac dataset as follows: if *answer* contains “Yes”/“No” in its first three words, label *(topic, facet, question, answer, 1/0)* accordingly. Otherwise, if an answer does not contain a “yes” or a “no” in its first three words, the instance is excluded from the training dataset. For example, the instance (“Korean language”, “What are some common Korean phrases?", “Ok, are you looking to find a Korean/English bilingual dictionary?”, “no i just want to know a few common korean phrases”) is labeled with a 0.

---

[3] https://github.com/alexandres/lexvec

[4] Implementation distributed by authors at https://github.com/aliannejadi/qulac

Table 3 Descriptive statistics for the Qulac dataset

# Topics	198
# Facets	762
Mean facets per topic	3.85
Median facets per topic	4
# Unique questions	2593
# Question-answer pairs	11039
# "No" answers	773
# Informative "No" answers	4747
# "Yes" answers	1997
# Neither "Yes" or "No" answers	3522
Mean terms per question	9.49
Mean terms per answer	8.21

Were the answer for this same topic-intent-question triplet "yes", the instance would receive the label 1.

For CoSearcher to respond to a clarification question, we experiment with a variety of lexical and semantic matching mechanisms to determine a match between a question and a user's intended topic facet. We adapt the work on Semantic Textual Similarity (STS) for this task (Agirre et al. 2012, 2013, 2014, 2015, 2016; Cer et al. 2017). Specifically, we fine-tune the BERT-large model (Devlin et al. 2019) which achieves state of the art performance on the STS Benchmark (STS-B) (Cer et al. 2017). We use the same setup as used in (Devlin et al. 2019) for the STS-B task (linearly decaying learning rate set to $2e-5$ with 10% warm-up steps, 3 training epochs, batch size equal to 32, and maximum sequence length of 128 tokens), but train a binary classifier rather than a regressor. The input to the BERT model is *"topic . intent [SEP] question"* using WordPiece tokenization, and the output supervision is given through the 0/1 labels described above. At inference time, the output is a match score—if a threshold is exceeded, CoSearcher returns "Yes" to indicate that the correct facet was proposed, otherwise the system uses the following procedure to generate a negative response: a sample is drawn from the Bernoulli cooperativeness variable; if it is 0, CoSearcher simply returns "No"; else if it is 0, an informative "No" is randomly sampled from human answers for the corresponding topic-facet pair. This answer generation procedure is applied whether the system uses either Qulac or Bing as candidate facets, since CoSearcher (the user simulator component of the system) always assumes a Qulac facet, for which human answers are available.

We split Qulac's 198 topics into 100 training, 25 validation, and 73 test topics, using only training and validation topics for the intent match classifier training/evaluation, and reserving the test topics as hidden for the full conversational system evaluations. Table 3 gives some descriptive statistics for the Qulac dataset. At threshold 0.5, which we use throughout this paper, the classifier achieves an 0.63 F1 score. Figure 2 shows the resulting Precision/Recall curve of our trained classifier. For responding with informative answers, we calculate $g(topic, facet, question)$, decide if it is a "Yes"/"No" using the chosen threshold of .5, and randomly pick a human answer to the same topic and facet that has the right 1/0 label. This setup allows us to test our system with a fully configurable user. The system is run via a controller that selects the user parameters, including topic/facet and also initializes the interaction with the agent.

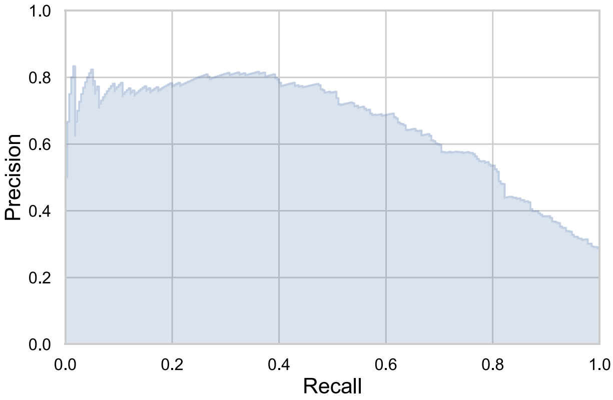


Fig. 2 Precision-Recall curve of BERT Yes/No classifier on Qulac validation set

5 Experimental results

We start our experiments by investigating whether or not our agent can interact with the user model to correctly identify the user's intent. We then study how much this conversation can improve downstream retrieval performance. All results reported are the average of 10 runs for a given agent/user model configuration.

5.1 Identification of user intent

We first set the user model to be fully cooperative ($p = 1$) with a patience of 3 turns, matching the cooperative nature and length of NeuQS (Aliannejadi et al. 2019) dialogues. We then vary the "Yes"/"No" classifier's threshold from 0 to 1 (step size 0.1). We test the random (Rand) and positive similarity (Sim) facet rankers since the fully cooperative user always returns an informative answer, reducing the need to use negative information.

Results are shown in Figs. 3a to c. We see that the Sim ranker is significantly better at identifying intent than the Rand ranker, and that for this level of user patience and cooperativeness, facet identification is highly successful. This improved performance is to be expected, since as shown in Sect. 3.2 the BoV facet ranker we use has high MRR. To avoid tuning the "Yes"/"'No" classifier threshold specifically to Qulac facets, we continue with the default of 0.5 for all experiments that follow.

5.2 Downstream search performance

Having shown that we can refine the user intent, we next assess whether this facet inference helps IR performance. We formulate the IR query with the topic and the first facet to which the user model answers "Yes", or only the topic if no "Yes" is received before user patience runs out. We use the exact same Query-Likelihood IR model/data as in the

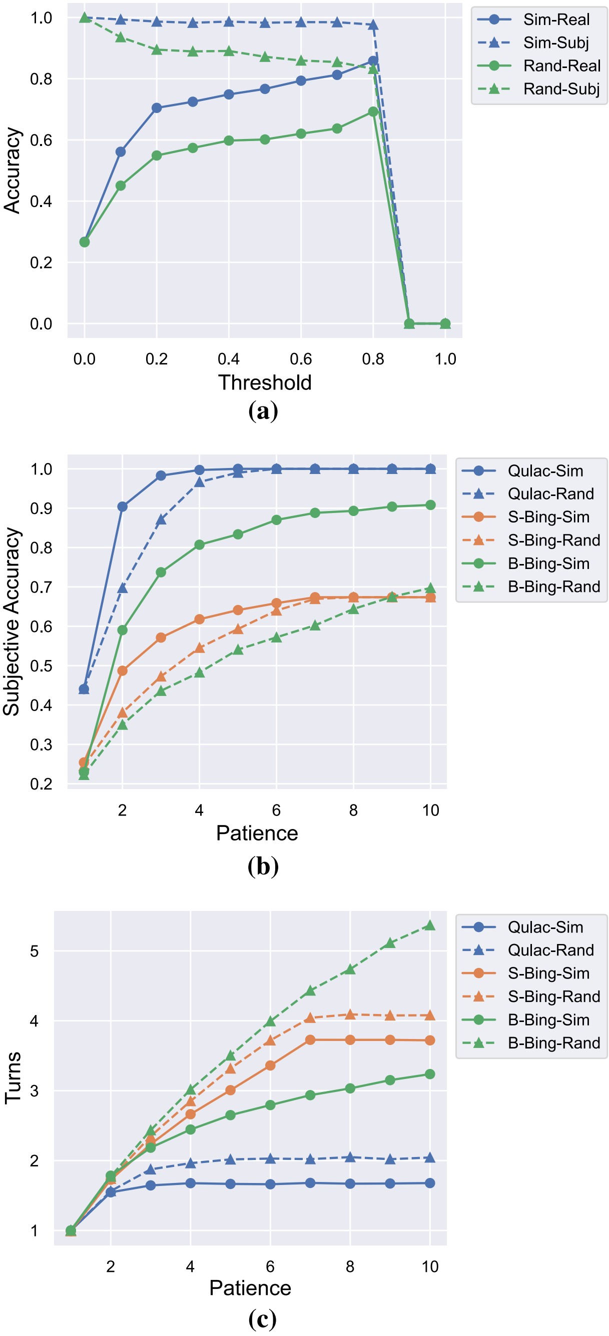


Fig. 3 **a** Real/Subjective accuracy on Qulac test set using Sim and Rand facet rankers on Qulac facets with cooperativeness=1, patience=3 with varying classifier thresholds. **b** Same as (**a**), but only subjective accuracy and using all 3 facet providers as patience is varied with fixed classifier threshold=0.5. **c** Same as (**b**), but evaluating number of turns instead of subjective accuracy (turns < patience means a "Yes" received before patience ran out, which plateaus for both Qulac and S-Bing because the number of candidate facets—on average 4 and 8 respectively—is smaller than patience.)

Table 4 Performance comparison between prior state of the art methods, including (Aliannejadi et al. 2019) (top, marked with *) and CoSearcher (bottom, marked with †)

Method	MRR	P@1	nDCG@1	nDCG@5	nDCG@20
Topic-only*	0.2715	0.1842	0.1381	0.1451	0.1470
σ-QPP*	0.3570	0.2548	0.1960	0.1938	0.1812
LambdaMART*	0.3558	0.2537	0.1945	0.1940	0.1796
RankNet*	0.3573	0.2562	0.1979	0.1943	0.1804
NeuQS*	**0.3625**	**0.2664**	**0.2064**	**0.2013**	**0.1862**
Topic-only†	0.2938	0.1900	0.1329	0.1456	0.1525
CoSearcher- Qulac†	**0.3999**	**0.3025**	**0.2263**	**0.2110**	**0.1908**
CoSearcher- S-Bing†	0.3136	0.2010	0.1415	0.1653	0.1597
CoSearcher- B-Bing†	0.3444	0.2366	0.1781	0.1769	0.1703

“Topic-only” refers to the baseline method issuing only the topic as the query to the search system, ignoring any facet information obtained through conversation. * and † results (example Topic-only* and Topic-only†) are not directly comparable because the authors use different train-test splits (* uses cross-validation, † uses single split)

NeuQS paper (Aliannejadi et al. 2019).[5] Although submitting the entire dialogue could potentially improve search performance, since it includes human user responses which often contain paraphrases of the search facet, we opt to use only the system’s best guess of what the correct facet is, as it excludes the previous (likely incorrect) facets discussed in the conversation.

We were not completely successful at reproducing the performance of the NeuQS system, so we compare our system using the Sim facet ranker to the results reported by Aliannejadi et al. (2019) on the same overall dataset. We mimic the combinatorially-generated dialogue used as input to NeuQS by setting CoSearcher cooperativeness to 1 and patience to 3. Results are given in Table 4. Our system using Qulac facets has a larger gap to the Topic-only baseline (+.1061) than NeuQS to its Topic-only baseline (+.0910). Dynamic facet generation outperforms the topic-only baseline; we see that having a large number of candidate facets is important: B-Bing has 26x more facets than S-Bing, allowing for *finer matching*.

5.3 Effects of patience and cooperativeness

We set cooperativeness to 1 and vary the patience of the user model. Results are shown in Fig. 4b. We note that similarity based ranking always outperforms random selection, and retrieval improves as patience increases. Random facet selection is feasible when the set of candidate facets is small, as is the case with Qulac and S-Bing. The performance degrades substantially, however, for the larger B-Bing facet generator, remaining close to the baseline topic MRR (see Table 4). In contrast, semantic similarity ranking shows clear improvements as the conversation progresses. Interestingly, MRR plateaus at around 4 turns for the Sim rankers, even for the B-Bing facets. This can be explained by the cooperativeness set

---

[5] Note that since they do not perform explicit intent refinement, they submit the entire dialogue context as a query to the IR system, whereas we submit only the topic and the refined facet.

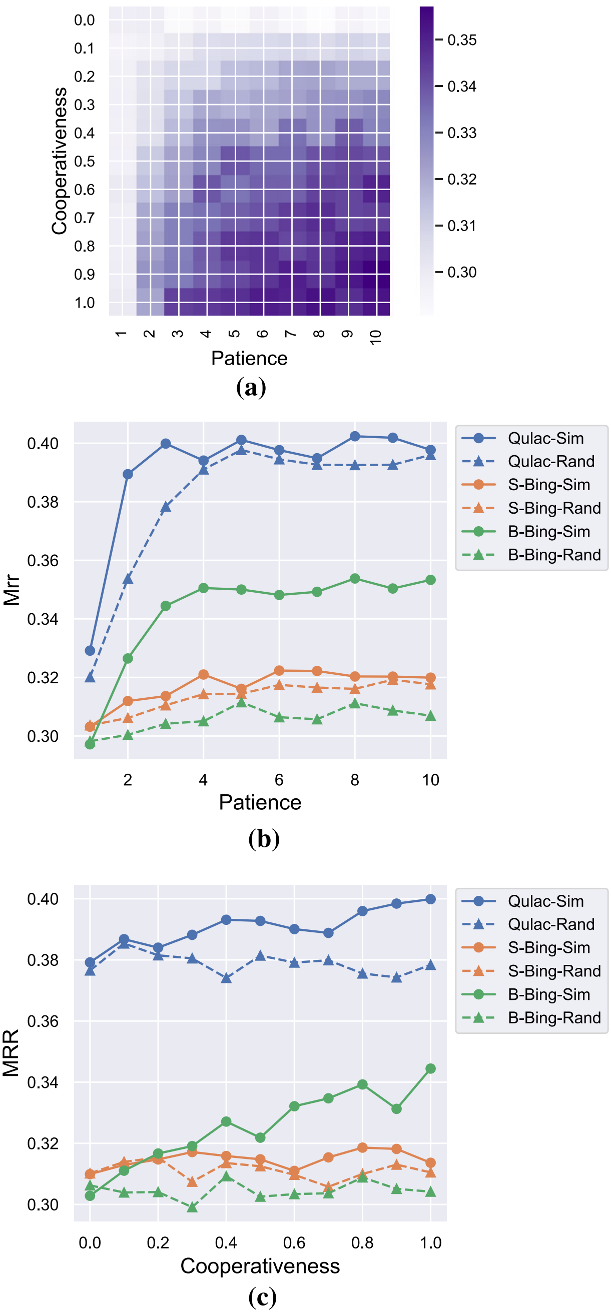


Fig. 4 **a**, **b**, **c** The effect of varying patience/cooperativeness: **a** Heatmap of MRR for B-Bing using similarity facet ranker as patience/cooperativeness vary. **b** MRR for all facet providers using Sim and Rand facet rankers for cooperativeness=1 as patience varied. **c** Same as (**b**), but fixing patience at 3 and varying cooperativeness

to 1 in this experiment, meaning that at every turn the user simulator is volunteering an informative response, sufficiently so as to allow the agent to identify the best matching candidate facet within 4 turns.

We repeat these experiments, but this time vary the cooperativeness rather than patience (which is now fixed at 3). Results are shown in Fig. 4c. The Sim ranker clearly benefits from higher cooperativeness, while Rand shows no improvement, as expected. Though cooperativeness has a positive impact on Qulac facets when using the Sim ranker, the impact is small when compared to B-Bing facets using the same ranker. We attribute this to patience being set to 3: since topics have on average 3.85 facets (see Table 3), even Rand facet ranker has strong performance given 3 turns of interaction. The considerable gap between B-Bing and S-Bing has a simple explanation: the user intent is less likely to be present in the small S-Bing set of facets than in the B-Bing superset, so additional cooperativeness helps one but not the other.

We next investigate the interaction between cooperativeness and patience, repeating the same setup from the previous IR experiments but this time varying both patience and cooperativeness. We study only B-Bing facets since these pose the hardest facet identification problem, requiring a deeper conversation to narrow down candidates. Results shown in Fig. 4a clearly indicate that *both* cooperativeness *and* patience are required to achieve maximal IR performance.

5.4 Enhanced facet representations

Enhanced representations, introduced in Sect. 3.2, are created by concatenating the facet representations with the titles of the top 10 documents returned by the IR system when it is queried for topic+facet. The intuition here is that these representations can provide the system with an expanded view of the facets, allowing it to better understand user queries that may be phrased differently. Table 5 shows two examples of facets and their enhanced representations.

Fixing cooperativeness at 1 and varying patience, we investigate the effect of using enhanced facet representations in the Sim facet ranker, called Sim-E. The results in Fig. 5a show minor improvements in Qulac, no change for S-Bing, and a detrimental effect for B-Bing.

To better understand these results, we perform a deeper analysis by focusing on the use of enhanced facets in S-Bing. We break down the aggregated performance and analyze the results at a more granular topic-facet level, observing that S-Bing-Sim-E outperforms S-Bing-Sim in 15% of the 279 topic-facet pairs, shows no change in 61% of cases, and degrades performance in 26% of the pairs. Analyzing the cases where performance improves, we observe cases where the enhanced representations include entries that are more similar to the language used by the user in formulating their responses. One such example is included in the first row of Table 5, where the enhanced representation helps the system match the a user's informative response "no im looking for information about tendinitis" to the Bing facet "forearm pain tendonitis". Ideally, as in this example, an enhanced representation created through retrieval should reflect the different ways in which a facet can be cast or rephrased. Here, the enhanced representation contains the valid alternative spelling "tendinitis" which matches the user feedback but differs from the spelling found in the facet description, leading to faster identification of the correct intent.

We also examine cases where performance degrades under enhanced representations. In such cases we often observe that the representations contain irrelevant documents

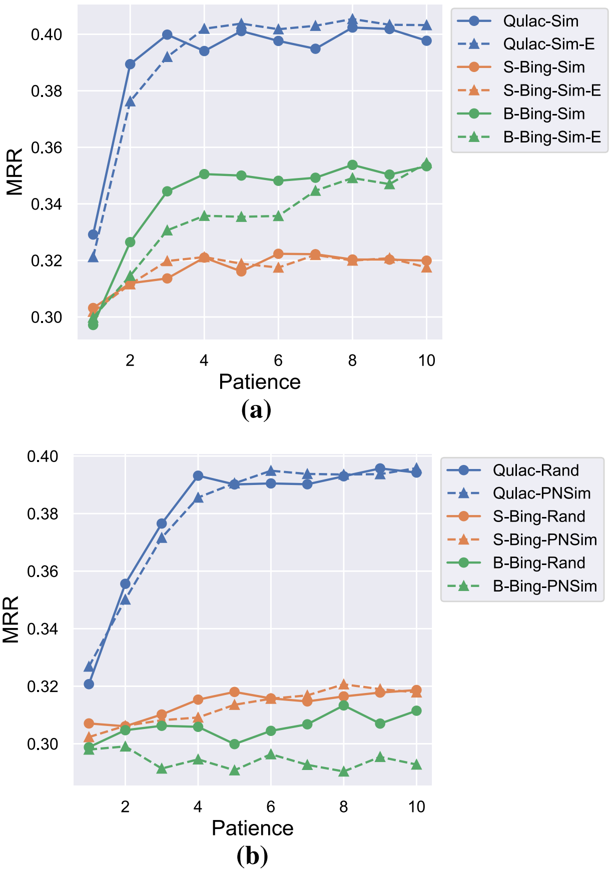


Fig. 5 MRR for all facet providers: **a** using Sim and Sim-E (enhanced representations) rankers **b** setting cooperativeness=0 and using PNSim facet ranker

returned by the IR system, essentially making the matching task noisier. In the second row of Table 5, we see an example that degrades performance of S-Bing-Sim-E compared to S-Bing-Sim: the enhanced representation of "diversity statement" is created from irrelevant documents, leading to a noisier facet representation than if no enhanced representation were used at all. We believe this is an inherent shortcoming of the IR system used in constructing enhanced representations, rather than the idea of enhanced representations, and in future work plan to experiment with more advanced IR models (Nogueira and Cho 2020) to see if the number of topic-facet pairs with degraded performance decreases.

5.5 Exploiting negative searcher feedback

We next study whether the "uninformative no" responses from the user, without any further explanation, can be used to improve candidate facet ranking. Our intuition is that if we receive a simple "No" to a facet guess, we should deprioritize guessing facets which are semantically similar or equally likely to receive yet another "No", wasting a turn of

Table 5 Examples of enhanced facet representations, which are formed by concatenating the **facet** with the top-10 document titles retrieved using an IR system using the topic+facet as input

Topic+Facet (in bold) and Top-10 document titles used in enhanced representation
forearm pain+forearm pain tendonitis
1. Houston Sports Medicine - Dr. Evan Collins Leading Orthopedic Surgeon
2. Self Care for RSI - Carpal Tunnel, Tendonitis, Thoracic Outlet, and more
3. Repetitive Strain Injuries
4. Pain Clinic - Nerve Pain - Entrapment Neuropathy - Trapped, pinched nerves, neuritis, neuropathy, neuralgia - Causes and treatment - Drugs, injections
5. Thermoflow Arm Band (One Size) : Online Pharmacy : Canadian Drugs Whole Sale
6. Climbing Injuries.com
7. Forearm Supports Reduce Upper Body Pain Linked To Computer Use
8. Tendonitis - *Tendinitis* - Weight Lifting
9. Tennis Elbow Glossary of Terms with Medical Definitions
10. tennis elbow
diversity+diversity statement
1. Best Colleges - Education - US News and World Report
2. Workplace diversity - Wikiversity
3. Diversity Action Plan - The Ohio State University
4. Browse Keywords: Diversity Consultant - Diversity Pg&e; |Juju Job Search
5. Diversity Elite 2008: The Top 15 - HispanicBusiness.com
6. Law Firm Diversity
7. OEOD
8. Chapter, Sections, and District Diversity Resources
9. Diversity and Libraries - EngagedIntellectual.org
10. Migration - Integration - Diversity: DiM in Unternehmen & Organisationen

Document numbers are included for reference only, and are not part of the representation

user patience. In other words, if a user is fully uncooperative, returning only simple "No" responses, would it be possible to outperform the Rand baseline?

Our hypothesis is that some of the candidate facets may have a semantic or hierarchical relationship (i.e. one facet is a subtype of another), and decreasing the selection likelihood of facets that are similar to a rejected one may be beneficial. Although the semantic relationships between our facets are not controlled in any way, we conduct experiments to explore the viability of this approach.

We repeat the search experiments, setting cooperativeness to 0, but varying patience, and compare the Rand and PNSim rankers (Sim ranker omitted since it only uses informative answers). Results are shown in Fig. 5b. Overall, there is no substantial gain when using the Qulac and S-Bing facets, but a clear loss with B-Bing facets. We hypothesize this drop is because the down-weighting of some candidate facets similar to the facet that the user rejected is too aggressive. Two facets can be very semantically similar but one receive a "Yes" and the other a "No".

We attribute the *uninformativeness* of the simple "No" to the weak semantic overlap between candidate facets: Qulac facets are quite distinct, and S-Bing search suggestions are purposefully diversified.

The outcomes are more subtle with B-Bing facet provider, where facets possess a tree hierarchy by design and thus have strong semantic similarity down tree paths. Getting a “No” to a parent node *could* preclude guesses of children, but this is a gamble. The “No” could mean (1) the guess was not specific enough, and children should be guessed or (2) the parent and its entire subtree are not what the user is looking for. Given that the “No” feedback is *hurting* performance of B-Bing (B-Bing-Sim vs B-Bing-PNSim in Fig. 5b), we assume (1) is occurring more often than (2), and the facet ranker is incorrectly down-ranking valid candidates. This is a fundamental and challenging problem as there is no way for the system to deduce whether the user means (1) or (2) from a “No”, again highlighting the need for cooperativeness. An alternative to full cooperativeness is education, where the user would be instructed to respond with graded negative responses such as “Not quite“ (1) and “No” (2). Or if the system is using voice, the tone of the “No” could be used for grading. A slow, reluctant “No” meaning (1), and quick, emphatic “No” meaning (2). We leave exploration of these ideas to future work.

5.5.1 Analyzing the effects of negative feedback

Having provided an explanation of why the negative feedback did not provide overall gains in results, we now perform a deeper quantitative and qualitative analysis of how CoSearcher utilizes this info, and how it impacts performance.

We first examine our results by considering the performance of individual topic-facet pairs in our test set, with the aim of understanding how this performance changes across all topics when negative feedback is incorporated. Our dataset and facets are not controlled for the relationship between candidate facets (i.e. we do not know in which cases the candidate facets are thematically related or not), and we believe that the system’s behavior may differ based on this criteria.

Using the S-Bing facet provider, we compare the MRR across all topic-facet pairs under two conditions: without using negative feedback (baseline), and with the PNSim model that uses it. Over 279 topic-facet pairs, we find that under PNSim MRR improves performance in 18% of the pairs, remains unchanged in 58%, and decreases compared to baseline in 24% of cases. This is an important finding as it shows that PNSim does change system behavior in some segments of the data.

To better understand this we qualitatively examine the conversations in the 3 segments. In the pairs that show gains, we often see several candidate facets that are related (i.e. can be semantically grouped). For example, under the topic “keyboard reviews” some of the facets are related to computer keyboards, and others to digital pianos. When PNSim receives negative information about one of the piano facets, the likelihood of the others also decreases, allowing the system to more rapidly narrow down to the user’s true intent. A more detailed example of how the system uses negative information is shown in Table 6.

In the other cases, where there is no change or a drop in performance, we generally note that the candidate facets are all equally diverse; eliminating one does not provide much information about the likelihood of the remaining ones. We also observe that under some conditions PNSim incorrectly down-weighs the likelihood of unrelated facets. This

Table 6 Tracking the state of the facet ranker for a dialogue (S-Bing-PNSim with cooperativeness set to 0 and patience set to 3; system uses S-Bing facets; gold user topic-facet are "euclidean"-"Take me to the homepage for the Euclid Chemical company") while accounting for negative feedback, where scores correspond to a softmax over Eq. (2)

System Question	User Answer	System Facet	Score
–	–	euclidean space	0.142
		euclid chemical company	0.142
		euclid ohio	0.142
		euclid chemical	0.142
		euclid's elements	0.142
		euclidean distance	0.142
		euclidean geometry	0.142
Are you looking for euclidean space?	no	euclid chemical company	0.178
		euclid ohio	0.176
		euclid chemical	0.174
		euclid's elements	0.173
		euclidean distance	0.150
		euclidean geometry	0.147
		euclidean space	0.000
Are you looking for euclid chemical company?	yes the euclid chemical company website	euclid chemical company	1.000
		euclid ohio	0.000
		euclid chemical	0.000
		euclid's elements	0.000
		euclidean distance	0.000
		euclidean geometry	0.000
		euclidean space	0.000

The first row shows the initial state, with all facets being equally likely. In the first turn the system asks "Are you looking for euclidean space?". The user answers "no". All facets related to mathematics have their scores decreased, and the system's next guess (the facet with maximal score "euclid chemical company") is correct when it asks "Are you looking for euclid chemical company?". The user's "yes..." answer ends the dialogue, and all scores but the guessed facet's are zeroed

behavior can be attributed to a failure of the similarity model in correctly capturing the relationship between facets.

These analyses demonstrate that using negative information can be beneficial, and our results are aligned with those reported in related work in the field. (Krasakis et al. 2020) also found single word negative responses to be detrimental to retrieval performance. And although (Hashemi et al. 2020) report positive results when using negative *or* short[6] clarification responses, their results for short responses conflate both short "Yes" and "No" answers. Since both question and answer are input to the retrieval system, it is intuitive that the question to a "Yes" response is highly informative, so these conflated performance numbers are lifted by including both positive and negative answers. Although it is possible that their system might perform well on simple "No" responses, we cannot conclude so from the results they report.

---

[6] The authors do not specify what short means, but for discussion we presume it to mean single word: a simple "Yes" or "No".

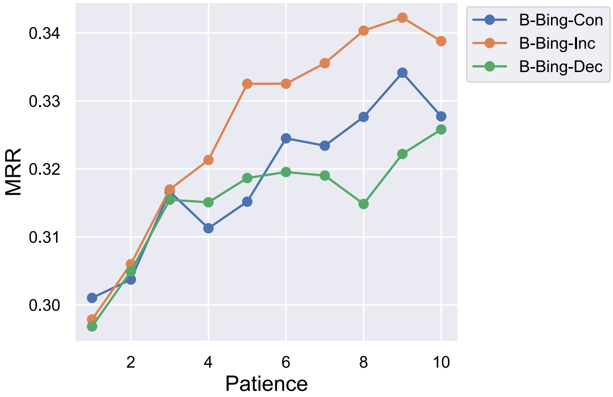


Fig. 6 The effect of dynamic cooperativeness in B-Bing (cooperativeness=0 and using PNSim facet ranker). We observe that increasing user cooperation results in improved results over constant cooperativeness. Decreasing the value has the opposite effect. -Con: constant cooperativeness, -Inc: increasing cooperativeness, -Dec: decreasing cooperativeness

In sum, our analysis shows that it is possible to learn from short negative responses under the right conditions, and leads us to some preliminary conclusions: (1) Negative information is most useful when some of the facets are related, and the clarification question is formed such that a “No” answer downranks more than a single facet. From an information theoretic perspective, where there is some probability distribution over facets, this corresponds to asking the question whose binary answer leads to the maximal reduction in entropy (information gain) of this distribution; (2) Errors in determining the semantic similarity between facets can cause the wrong facets to be down-weighted, resulting in cascading errors as the system pursues incorrect facets.

5.6 Dynamic user behavior

When users interact with a system, these interactions are impacted by feedback loops: the system output can change depending on user input (e.g. our usage of negative responses), but the converse can also occur, where user behavior changes in response to the system. Depending on the system output, the user may become more or less cooperative and/or patient. Our motivation is to see if a user simulator can be used to study such changes, and how these behaviors affect performance. To this end, our final experiment begins to investigate how users with dynamically varying cooperativeness behave, as a conversation progresses.

We use the models described in Sect. 3.4.5, setting $p(0) = .2$. We focus on B-Bing where cooperativeness is most important given the large number of candidate facets. Results are shown in Fig. 6. As expected, increasing cooperativeness yields the strongest results. Surprisingly, decreasing cooperativeness is only *slightly* detrimental when compared to the constant alternative. This can be explained, however, by the user’s initial cooperative response, i.e., before decay turns all responses to simple “No”’s, being sufficiently informative to narrow down the search space of candidate facets, and this ranking persists throughout the dialogue. Though a more cooperative user would return multiple informative responses containing paraphrases of his intent, some cooperativeness is sufficient, even in the challenging setting of many candidate facets of B-Bing facet provider.

These findings have practical implications for conversational search systems. System actions at each interaction that have a negative influence on user cooperativeness can be detrimental to the overall results. Note that in the current experimental setup, following prior work, the system does not return any search results until the (simulated) user confirms the facet, which prevents the simulator from conditioning behavior parameters on result quality. We plan to extend the CoSearcher model to include more sophisticated models of cooperativeness and patience e.g., conditioned on the result quality so far, which would require a different evaluation scheme and experimental setup.

In sum, we showed that different CoSearcher configurations (user configuration, facet providers, etc.) led to a wide range of IR performances, demonstrating the functionality and applicability of our framework.

6 Analysis and discussion

6.1 Characterization of successful conversational refinement

Using the conversations generated with a wide range of behavior simulator features, we can explore what makes for a successful conversational search session. It is clear that the topic of the query has some effect on the difficulty of the task. We attempt to quantify this intuition through semantic analysis of the properties of search topics and facets to gain insight into the system performance.

We observe that ambiguous entities are associated with lower success rates across all facet providers. Examples of such entities with multiple senses include: *iron* (chemical element, clothing iron, nutritional supplement), *Euclid* (person, multiple businesses), *Rice* (food, person name, e.g., Rice university). Conversely, unambiguous entities are associated with much higher success rates, e.g., *Universal Animal Cuts* (a product), or *solar panels*. To quantify this we simulate 100 dialogues for each facet and measure the ratio of successful conversations. Using a sample of 20 topics (10 ambiguous entities, 10 non-ambiguous) we observe an average success rate of 55% for the ambiguous ones, compared to 72% for the non-ambiguous entities.

Similarly, topic ambiguity is a key factor. Topics that are broad in nature, with a large number of potential facets, yield poorer results. One such example is the topic *cass county missouri* with the facet ‘What was the 2008 budget for Cass County, MO?’. For a sample of 10 topics with ≥ 5 Qulac facets, we observe a mean success rate of 58%, against 66% for 10 topics with ≤ 3 facets. We hypothesize that it can be difficult to refine the query to such a specific facet within a reasonable number of turns.

Finally, facets containing multiple entities and entities that are complex noun phrases were often associated with poorer performance. For a sample of 10 topics with complex entities, we observed an average success rate of 54%, compared to an overall average of 62%. These results indicate that entity extraction and disambiguation are key building blocks for successful conversational systems.

USER: South Africa *(the topic)*

BOT: **Are you looking to find pictures of South Africa?**
COSEARCHER: no but historical maps along with current ones would be good
BOT: **Are you looking to find pages that give the history of South Africa?**
COSEARCHER: no i want maps of south africa
BOT: **Are you looking to find maps of South Africa?**
COSEARCHER: yes

True intent = "Find maps of South Africa"

(a)

USER: Quit Smoking *(the topic)*

BOT: **Are you looking for quit smoking withdrawal symptoms?**
COSEARCHER: no i am interested if it is possible to quit smoking cold turkey
BOT: **Are you looking for quit smoking cold turkey?**
COSEARCHER: yes specifically the cold turkey method

True intent: "Can you quit smoking using the cold turkey method?"

(b)

Fig. 7 **a** an example of a successful conversation (cooperativeness=1, Qulac facets); **b** an example of a successful conversation (cooperativeness=1, Bing facets)

6.2 Qualitative analysis: case studies

We complement our analysis above by offering case studies to provide intuition on why conversational search succeeds and fails in different situations under various user "personas" with varying degrees of cooperativeness. First, we consider an example of a cooperative user interacting with a system using the Qulac (static) topic facets, shown in Fig. 1b and 7a. Recall that for high value of cooperativeness, the user (and the simulator) often volunteer information to the search system, even if the initial response or guess was not correct, i.e., provide "informative no" responses. As a result, we observe the search system quickly converging on the true searcher intent. Another successful example using the Bing query suggestion facets is shown in Fig. 7b. Given the large number of relevant

USER: iron *(the topic)*

BOT: **Are you looking for iron throne?**
COSEARCHER: no i want to know about iron deficiencies
BOT: **Are you looking for iron rich foods?**
COSEARCHER: no i would like information about iron deficiencies
BOT: **Are you looking for iron normal range?**
COSEARCHER: yes

True intent: "Find info about iron deficiencies"

(a)

USER: barbados *(the topic)*

BOT: **Are you looking to find a map of Barbados?**
COSEARCHER: no
BOT: **Are you looking to find the Barbados Underground homepage?**
COSEARCHER: no
BOT: **Are you looking to find information on Barbados Prime Minister David Thompson?**
COSEARCHER: no
BOT: **Are you look for what does the Barbados flag look like?**
COSEARCHER: no
BOT: **Are you looking to suggest tourist activities in Barbados?**
COSEARCHER: no

True intent: "Find info on Barbados history"

(b)

Fig. 8 **a** an example of a matching error (cooperativeness=1, Bing facets). The user incorrectly accepts a facet that is very closely related to the true intent. **b** An unsuccessful conversation with a non-cooperative user (cooperativeness=0, Qulac facets)

facets available via the external search provider, the system is able to match the Qulac facet within 2 turns.

The example in Fig. 7b highlights the importance of realistically modeling "informative rejection" via our proposed cooperativeness parameter. In this example, a *cooperative* user volunteers her intent immediately, as soon as the system asks a clarification question. This is a known limitation of the Qulac dataset (which is crowdsourced with highly cooperative "users"), but may not be realistic. A more common scenario is that a user may not be able to fully specify her intent (hence the vague original query), but can easily recognize the topic facets she is, or is *not* interested in when prompted. The CoSearcher framework explicitly models and allows to automatically identify such cases. Consider a failed conversation (Fig. 8a), also with a cooperative user, using the Bing query suggestions (dynamic facets) as candidate facets. In the simulated conversation example below, the search system continues to ignore the search intent refinements volunteered by the cooperative CoSearcher user model, until the user simulator finally accepts the (incorrect) intent suggestion, likely resulting in non-relevant results. Finally, in Fig. 8b we present an example of a similarly uncooperative user, but where the conversation ends before a match is found.

These examples provide additional intuition about the challenges in conversational search refinement, and illustrate the range of conversations and interactions that CoSearcher can support to simulate different types of users and search tasks.

7 Conclusions and future work

We investigated the effectiveness of conversational search refinement, a key task for conversational search systems. We hypothesized that the success of conversational search depends significantly on the users' behavior and the search task characteristics. To accomplish this, we introduced a parameterized conversational search user simulator, CoSearcher, to systematically probe the boundaries of conversational search intent refinement. CoSearcher was used to evaluate the effectiveness of our query facet identification algorithm under a variety of conditions corresponding to different types of users. Our experiments on an existing benchmark (Qulac) and a new, dynamically generated dataset of search intent facets, demonstrate the power and generality of CoSearcher, exhibiting a new state of the art performance.

We also systematically explored the space of conversational search refinement outcomes for different types of search tasks and users. Specifically, we characterized the semantic differences between search topics and intents which are more (or less) amenable to conversational search refinement; We also empirically showed that (1) For the interesting real-world scenario where set of facets is large and a non-random facet ranker is used (B-Bing-Sim), cooperation on the user's part is fundamental for the success of conversational search refinement (in Fig. 4c, a uncooperative user's MRR in 3-turn-or-less dialogue is nearly identical to the .2938 topic-only baseline, improving up to .3444 as cooperativeness increases); and as illustrated in Fig. 4a), the effort (characterized by patience and cooperativeness) vs. benefit (MRR) tradeoff can be quantified: linear regression gives $MRR = .0038 \times patience + .034 \times cooperativeness + .29$ with $R^2 = 0.77$. (2) A simple semantic policy is effective for identifying searcher intent: in all experiments, it outperforms Random facet selection; in particular for B-Bing-Sim in Fig. 4b, MRR plateauing at 4 turns indicates that the best matching facet of the 216 candidates facets has been identified; (3) Dynamic search intent facet generation is feasible: MRR of .3444 for B-Bing-Sim is much higher than the topic-only baseline of .2938, suggesting a promising direction for future extensions by considering other sources of search intent facets.

New experimental results in this edition of the paper also provide further insights into the task. Our experiments with enhanced representations demonstrated that additional information can extend the conversational agent's understanding of the topic, allowing it to link synonyms and related terms to user requests. On the other hand, we also observed that when the underlying IR system used to retrieve documents for the enhanced facets provides irrelevant documents, this noise can actually degrade our performance. Future work in this direction should focus on ensuring that high quality documents are used to generate extended facet representations.

We also presented results using a facet ranker that tries to leverage uninformative "no" responses from users by attempting to exploit semantic similarity between the rejected facet and remaining candidate facets in order to down-weigh those that may be similar, which increases the diversity of our clarification questions. We showed that this approach can improve results in cases where several candidate facets come from the same sub-topic, but not in cases where there is little thematic overlap among the facets.

Another experiment demonstrated how conversational agents can be tested under changing user behaviors, proposing a new setting for evaluating systems, and yet again demonstrating the importance of user cooperation in the success of the refinement system.

Our proposed simulation framework can be adapted to search and recommendation tasks in other domains, if they can be modeled through topics and facts or sub-topics. For example, instead of information-seeking dialogues, an e-commerce product search scenario could be simulated by CoSearcher. Similarly, a media recommendation agent could be modeled in the same manner.

A conversational system trained using a simulator will only be as effective as the simulator itself. Consequently, the design and development of an effective simulator is a problem equally important and challenging as developing a conversational dialogue system. We emphasize that our described results and analysis required simulating hundreds of thousands of conversational search refinement experiments, enabled by the presented CoSearcher simulator. However, previous work, like that of Schatzmann et al. (2007), has demonstrated such systems can effectively bootstrap the training of conversational agents, but additional effort is needed to achieve satisfactory results.

A promising future direction is to expand CoSearcher to support more sophisticated behavior dynamics, which could be conditioned on the conversation length, search result quality, task characteristics, or other contextual factors. Additionally, CoSearcher is naturally suited for scenarios where the user intent is in natural language, but the system represents facets as database queries (e.g., over an e-commerce catalog) and must select or generate these queries through dialogue. Finally, recent advances in models for text generation allow more advanced simulations through rephrasing of the ground-truth responses to create more diverse inputs used to test conversational systems.

The combination of the new state of the art results, our empirical insights, and the newly introduced flexible CoSearcher framework – complemented by the new dynamic search intent dataset to be released, provide significant progress towards more intelligent and effective conversational search systems.

Funding Not applicable

Declarations

Conflict of interest Not applicable

References

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Wiebe, J. (2014, August). SemEval-2014 task 10: Multilingual semantic textual similarity. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014) (pp. 81-91). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S14-2010 10.3115/v1/S14-2010

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Wiebe, J. (2015, June). SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 252–263). Denver, Colorado: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S15-2045 10.18653/v1/S15-2045

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Wiebe, J. (2016, June). SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 497–511). San Diego, California: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S16-1081 10.18653/v1/S16-1081

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W. (2013, June). *SEM 2013 shared task: Semantic textual similarity. *Second joint conference on lexical and computational semantics (*SEM), volume 1: Proceedings of the main conference and the shared task: Semantic textual similarity* (pp. 32–43). Atlanta, Georgia, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S13-1004

Agirre, E., Diab, M., Cer, D., Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. *Proceedings of the first joint conference on lexical and computational semantics-volume 1: Proceedings of the main conference and the shared task, and volume 2: Proceedings of the sixth international workshop on semantic evaluation* (pp. 385–393).

Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., Burtsev, M. (2020). *ConvAI3: Generating Clarifying Questions for Open-Domain Dialogue Systems (ClariQ). arXiv preprint.* arXiv:2009.11352.

Aliannejadi, M., Zamani, H., Crestani, F., Croft, W.B. (2019). Asking clarifying questions in open-domain information-seeking conversations. *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval* (pp. 475–484).

Azzopardi, L., & Zuccon, G. (2016). An analysis of the cost and benefit of search interactions. *Proceedings of the 2016 acm international conference on the theory of information retrieval* (pp. 59–68).

Belkin, N. J., Cool, C., Stein, A., & Thiel, U. (1995). Cases, scripts, and informationseeking strategies: On the design of interactive information retrieval systems. *Expert systems with applications, 9*(3), 379–395.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146. Retrieved from https://www.aclweb.org/anthology/Q17-1010https://doi.org/10.1162/tacl_a_00051

Bordes, A., Boureau, Y.-L., Weston, J. (2016). Learning end-to-end goal-oriented dialog. *arXiv preprint*arXiv:1605.07683.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L. (2017, August). SemEval2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 1–14). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S17-2001https://doi.org/10.18653/v1/S17-2001

Chandramohan, S., Geist, M., Lefevre, F., Pietquin, O. (2011). User simulation in dialogue systems using inverse reinforcement learning. *Twelfth annual conference of the international speech communication association.*

Clarke, C.L., Craswell, N., Soboroff, I. (2009). *Overview of the trec 2009 web track* (Tech. Rep.). WATERLOO UNIV (ONTARIO).

Clarke, C.L., Craswell, N., Voorhees, E.M. (2012). *Overview of the trec 2012 web track* (Tech. Rep.). NATIONAL INST OF STANDARDS AND TECHNOLOGY GAITHERSBURG MD.

Crescenzi, A., Kelly, D., Azzopardi, L. (2016). Impacts of time constraints and system delays on user experience. *Proceedings of the 2016 acm on conference on human information interaction and retrieval* (pp. 141-150).

Croft, W. B., & Thompson, R. H. (1987). I3r: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science, 38*(6), 389–404.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

El Asri, L., He, J., & Suleman, K. (2016). A sequence-to-sequence model for user simulation in spoken dialogue systems. *Interspeech, 2016*, 1151–1155.

Fagan, J. C. (2010). Usability studies of faceted browsing: A literature review. *Information Technology and Libraries, 29*(2), 58–66.

Hashemi, H., Zamani, H., & Croft, W.B. (2020). Guided transformer: Leveraging multiple external sources for representation learning in conversational search. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval* (p. 1131-1140). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3397271.3401061

Hearst, M. (2006). Design recommendations for hierarchical faceted search interfaces. *Acm sigir workshop on faceted search* (pp. 1–5).

Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., & Yee, K.-P. (2002). Finding the flow in web site search. *Communications of the ACM, 45*(9), 42–49.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS), 20*(4), 422–446.

Keizer, S., Gasic, M., Jurcicek, F., Mairesse, F., Thomson, B., Yu, K., Young, S. (2010). Parameter estimation for agenda-based user simulation. In Proceedings of the sigdial 2010 conference (pp. 116–123).

Kiesel, J., Bahrami, A., Stein, B., Anand, A., Hagen, M. (2018). Toward voice query clarification. The 41st international acm sigir conference on research & development in information retrieval (pp. 1257–1260).

Kotov, A., & Zhai, C. (2010). Towards natural question guided search. In Proceedings of the 19th international conference on world wide web (pp. 541–550).

Krasakis, A.M., Aliannejadi, M., Voskarides, N., Kanoulas, E. (2020). Analysing the effect of clarifying questions on document ranking in conversational search. Proceedings of the 2020 acm sigir on international conference on theory of information retrieval (p. 129–132). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3409256.3409817

Kules, B., Capra, R., Banta, M., Sierra, T. (2009). What do exploratory searchers look at in a faceted search interface? Proceedings of the 9th acm/ieee-cs joint conference on digital libraries (pp. 313–322).

Li, X., Lipton, Z.C., Dhingra, B., Li, L., Gao, J., Chen, Y.-N. (2016). A user simulator for task-completion dialogues. arXiv preprint arXiv:1612.05688.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A. (2018). Advances in pre-training distributed word representations. In Proceedings of the international conference on language resources and evaluation (lrec 2018).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. InProceedings of the 26th international conference on neural information processing systems - volume 2 (p. 3111–3119). Red Hook, NY, USA: Curran Associates Inc.

Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., Vanderwende, L. (2016). Generating natural questions about an image. In Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers) (pp. 1802–1813).

Nogueira, R., & Cho, K. (2020). Passage re-ranking with bert.

Papangelis, A., Papadakos, P., Kotti, M., Stylianou, Y., Tzitzikas, Y., Plexousakis, D. (n.d.). Ld-sds: Towards an expressive spoken dialogue system based on linked-data. arXiv:171002973v1

Pennington, J., Socher, R., Manning, C. (2014, October). GloVe: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://doi.org/10.3115/v1/D14-1162

Rao, S., & Daume III, H. (2018). Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers) (pp. 2737–2746).

Rao, S., & Daume III, H. (2019). Answer-based adversarial training for generating clarification questions. In Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers) (pp. 143–155).

Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp) (pp. 3982-3992). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D19-1410https://doi.org/10.18653/v1/D19-1410

Rudnicky, A., & Xu, W. (1999). An agenda-based dialog management architecture for spoken language systems. Ieee automatic speech recognition and understanding workshop (Vol. 13).

Salle, A., Idiart, M., Villavicencio, A. (2016). Enhancing the lexvec distributed word representation model using positional contexts and external memory.

Salle, A., Malmasi, S., Rokhlenko, O., Agichtein, E. (2021). Studying the effectiveness of conversational search refinement through user simulation. In: D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, & F. Sebastiani (Eds) Advances in information retrieval. Springer International Publishing, Cham, pp. 587–602.

Salle, A., & Villavicencio, A. (2018, June). Incorporating subword information into matrix factorization word embeddings. In: Proceedings of the second workshop on subword/character LEvel models (pp. 66–71). New Orleans: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-1209

Schatzmann, J., Thomson, B., Weilhammer, K., Ye, H., Young, S. (2007, April). Agenda-based user simulation for bootstrapping a POMDP dialogue system. Human language technologies 2007: The conference of the north American chapter of the association for computational linguistics; companion volume, short papers (pp. 149-152). Rochester, New York: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N07-2038

Shah, P., Hakkani-Tur, D., Heck, L. (2016). Interactive reinforcement learning for task-oriented dialogue management. *Workshop on Deep Learning for Action and Interaction.*

Smucker, M.D., & Clarke, C.L. (2012). Time-based calibration of effectiveness measures. In Proceedings of the 35th international acm sigir conference on research and development in information retrieval (pp. 95–104).

Sun, Y., & Zhang, Y. (2018). Conversational recommender system. In The 41st international acm sigir conference on research & development in information retrieval (pp. 235–244).

Trippas, J.R., Spina, D., Cavedon, L., Joho, H., Sanderson, M. (2018). Informing the design of spoken conversational search: perspective paper. In Proceedings of the 2018 conference on human information interaction & retrieval (pp. 32-41).

Tunkelang, D. (2009). Faceted search. *Synthesis lectures on information concepts, retrieval, and services, 1*(1), 1–80.

Vandic, D., Aanen, S., Frasincar, F., & Kaymak, U. (2017). Dynamic facet ordering for faceted product search engines. *IEEE Transactions on Knowledge and Data Engineering, 29*(5), 1004–1016.

Weizenbaum, J., et al. (1966). Eliza-a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45.

Wen, T., Vandyke, D., Mrksšíc, N., Gasšíc, M., Rojas-Barahona, L., Su, P., Young, S. (2017). A network-based end-to-end trainable task-oriented dialogue system. 15th conference of the european chapter of the association for computational linguistics, eacl 2017-proceedings of conference (Vol. 1, pp. 438-449).

Yee, K.-P., Swearingen, K., Li, K., Hearst, M. (2003). Faceted metadata for image search and browsing. In Proceedings of the sigchi conference on human factors in computing systems (pp. 401-408).

Young, S. J. (2000). Probabilistic methods in spoken-dialogue systems. Philosophical Transactions of the Royal Society of London. *Series A Mathematical Physical and Engineering Sciences, 358*(1769), 1389–1402.

Zamani, H., Lueck, G., Chen, E., Quispe, R., Luu, F., Craswell, N. (2020). Mimics: A large-scale data collection for search clarification. In Proceedings of the 29th ACM international conference on information & knowledge management (pp. 3189–3196).

Zhang, S., & Balog, K. (2020). Evaluating conversational recommender systems via user simulation. In Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining (pp. 1512–1520)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.