Studying the Effectiveness of Conversational Search Refinement through User Simulation

Alexandre Salle[1⋆], Shervin Malmasi[2], Oleg Rokhlenko[2], and Eugene Agichtein[2,3]

[1] Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil
alex@alexsalle.com
[2] Amazon, Seattle, WA, USA {malmasi,olegro,eugeneag}@amazon.com
[3] Emory University, Atlanta, GA, USA

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 unfeasible to examine a wide range of user behaviors. To support robust training/evaluation of such systems, we propose a simulation framework called CoSearcher[4] that includes a parameterized user simulator controlling key behavioral factors like cooperativeness and patience. Using a standard conversational query clarification benchmark, we experiment with a range of user behaviors, semantic policies, and dynamic facet generation. Our results quantify the effects of user behaviors, and identify critical conditions required for conversational search refinement to be effective.

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

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 [7] 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 [31]. 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

⋆Work conducted during an internship at Amazon, Seattle, WA, USA.

[4]Information about code/resources available at https://github.com/alexandres/CoSearcher

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, 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 "variables".

In summary, our contributions include:

- 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 (Section 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. (Section 5).

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 [34, 14, 8, 37]. Advances in NLP and IR in the last few years have also been accompanied by a surge in research of conversational systems.

Within the sub-field, understanding user behavior is an important research direction. [31] and [21] performed user studies to understand what kind of user behavior is useful for conversational search, but they did not explicitly model the results for use in simulations. Additionally, [30, 39] 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 [11, 16, 24], 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 [25–28, 38]. 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 [19, 36, 18, 32, 23, 17, 22, 33].

Most similar to our work is that of [7], 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 [12, 13], 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 [6] 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 and manual annotations, complements Qulac and allows us to investigate additional challenges that arise with numerous query facets.

The Qulac paper [7] 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 [35, 9].

3 Modeling Conversational Search Intent Refinement through User Simulation

We now overview the conversational search intent refinement setting, following the recent formulation in [7], and our simulation-based approach for investigating this topic.

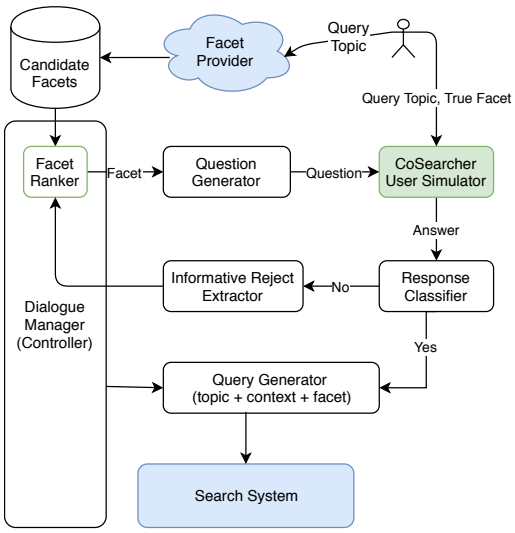
3.1 Problem Setting: Conversational Search Refinement

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. After each turn, the search system may chose to ask additional clarification questions, or return search results, or both. An example conversational search dialogue is shown in Fig. 3a, 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 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 selects the next candidate facet c from the list of candidate facets. If the user's "No" is *informative* (has additional information which might guide refinement, such as "No, I'm looking for...") we add the answer to the current context to be used for re-ranking. 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

We consider two facet ranking strategies: **(1) Rand**: a random baseline that orders facets randomly. **(2) Sim**: a semantic similarity strategy, which assigns a score for every candidate facet by computing the mean cosine similarity between the facet and each infor-



(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.

mative "No" in the conversation context, using mean bag-of-vectors as sentence embeddings. We use the LexVec n-gram subword vectors [29] to represent each word.

3.3 Dynamic Facet Generation

The previous state of the art approach — NeuQS [7] 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, [5] which, given an initial query, returns a set of 8 query completions. The query topic is used the initial query and the returned set of completions as candidate 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

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

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;

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 $g(\textit{topic}, \textit{intent}, \textit{question})$ that returns a similarity score between the topic/intent and the question. The "Yes"/"No" is then decided using a threshold that be chosen using downstream performance, or intrinsically evaluated if there is labeled "Yes"/"No" data.

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 [20]. Manipulating these two parameters via simulations enables us to expose the direct relationship between these key user behavior factors and conversational system results.

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). 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.

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.

While in this paper we fix a user's patience and cooperativeness parameters throughout a conversation session, CoSearcher can also be configured to update these values dynamically, which can increase or decrease cooperativeness or patience of the user as the session progresses. 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.

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 [7]. Our "Yes"/"No" classifier fine-tunes the BERT-large uncased model from [15]. The similarity rankers use the LexVec [29] n-gram embeddings.[6] The IR search system is the same query-likelihood model used by [7][7] indexed on ClueWeb09b.

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 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 Section 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, else ignore it.

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 [5, 4, 2, 1, 3, 10]. Specifically, we fine-tune the BERT-large model [15] which achieves state of the art performance on the STS Benchmark (STS-B) [10]. We use the same setup as used in [15] for the STS-B task, but train a binary classifier rather than a regressor. The input to BERT model is *"topic . intent [SEP] question"* using WordPiece tokenization, and the output is a match score - if a threshold is exceeded, CoSearcher returns "Yes" to indicate that the correct facet was proposed, and "No" otherwise (potentially with additional information as described above).

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.

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

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

At threshold 0.5, which we use throughout this paper, the classifier achieves an 0.63 F1 score. Figure 2a shows the resulting Precision/Recall curve of our trained classifier. For responding with informative answers, we calculate $g(\textit{topic}, \textit{facet}, \textit{question})$, 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.

5 Results and Discussion

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 NeuQS paper [7].[8] 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 adapting NeuQS to our exact problem setting (*explicit* intent refinement), so we compare our system using the Sim facet ranker to the results reported by [7] on the same overall IR task and dataset. We mimick the combinatorially-generated dialogue used as input to NeuQS by setting CoSearcher cooperativeness to 1 and patience to 3. Results are given in Table 1. 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*.

5.1 Effects of Patience and Cooperativeness

We set cooperativeness to 1 and vary the patience of the user model. Results are shown in Fig. 2c. 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 1). In contrast, semantic similarity ranking shows clear improvements as the conversation progresses.

We repeat these experiments, but this time vary the cooperativeness rather than patience (which is now fixed at 3). Results are shown in Fig. 2d. The Sim ranker clearly benefits from higher cooperativeness, while Rand shows no improvement, as expected. 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 no the other.

[8] 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.

Table 1: Performance comparison between prior state of the art methods, including [7] (top) and CoSearcher (bottom). "Topic-only" refers to the baseline method issuing only the topic as the query to the search system system, ignoring any facet information obtained through conversation.

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

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. 2b clearly indicate that *both* cooperativeness *and* patience are required to achieve maximal IR performance.

In sum, we showed that different CoSearcher configurations (user config, 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.

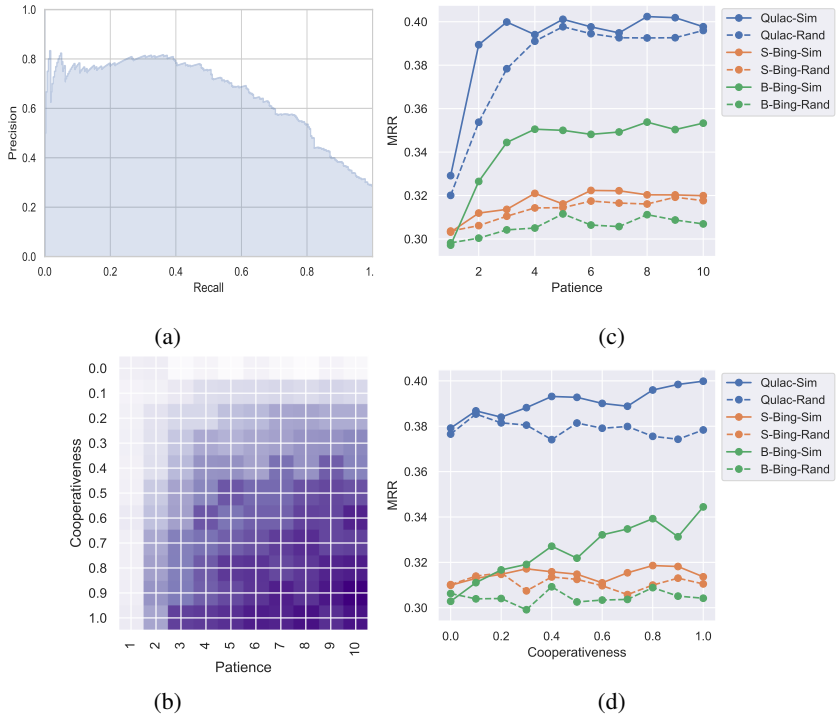


Fig. 2: (a) Precision-Recall curve of BERT Yes/No classifier on Qulac validation set. (b, c, d) The effect of varying patience/cooperativeness: (b) Heatmap of MRR for B-Bing using similarity facet ranker as patience/cooperativeness vary. (c) MRR for all facet providers using Sim and Rand facet rankers for cooperativeness=1 as patience varied. (d) Same as (c), but fixing patience at 3 and varying cooperativeness.

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 indicates that entity extraction and disambiguation are key building blocks for successful conversational systems.

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

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?"

(a)

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 user intent: "Find info about iron deficiencies"

(b)

Fig. 3: (a) an example of a successful conversation (cooperativeness=1, Bing facets); (b) 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.

"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. 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. 3a. Given the large number of relevant facets available via the external search provider, the system is able to match the Qulac facet within 2 turns.

The example in Figure 3a 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. 3b), 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.

We can see that in the above examples the system uses the information from the user to identify the true intent within a few turns. 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 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. 2d, 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. 2b), 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. 2c, 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.

We emphasize that the described results and analysis required simulating hundreds of thousands of conversational search refinement experiments, enabled by the presented CoSearcher simulator. In the future, we plan 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 catalogue) and must select or generate these queries through dialogue.

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.

References

1. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., Wiebe, J.: SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 252–263. Association for Computational Linguistics, Denver, Colorado (Jun 2015). https://doi.org/10.18653/v1/S15-2045, https://www.aclweb.org/anthology/S15-2045
2. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., Wiebe, J.: SemEval-2014 task 10: Multilingual semantic textual similarity. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 81–91. Association for Computational Linguistics, Dublin, Ireland (Aug 2014). https://doi.org/10.3115/v1/S14-2010, https://www.aclweb.org/anthology/S14-2010
3. Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., Wiebe, J.: SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 497–511. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/S16-1081, https://www.aclweb.org/anthology/S16-1081
4. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W.: *SEM 2013 shared task: Semantic textual similarity. In: 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. Association for Computational Linguistics, Atlanta, Georgia, USA (Jun 2013), https://www.aclweb.org/anthology/S13-1004
5. Agirre, E., Diab, M., Cer, D., Gonzalez-Agirre, A.: Semeval-2012 task 6: A pilot on semantic textual similarity. In: 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. Association for Computational Linguistics (2012)
6. Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., Burtsev, M.: ConvAI3: Generating Clarifying Questions for Open-Domain Dialogue Systems (ClariQ) (2020)
7. Aliannejadi, M., Zamani, H., Crestani, F., Croft, W.B.: Asking clarifying questions in open-domain information-seeking conversations. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 475–484. ACM (2019)
8. Belkin, N.J., Cool, C., Stein, A., Thiel, U.: Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. Expert systems with applications **9**(3), 379–395 (1995)
9. Bordes, A., Boureau, Y.L., Weston, J.: Learning end-to-end goal-oriented dialog. arXiv preprint arXiv:1605.07683 (2016)
10. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 1–14. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/S17-2001, https://www.aclweb.org/anthology/S17-2001
11. Chandramohan, S., Geist, M., Lefèvre, F., Pietquin, O.: User simulation in dialogue systems using inverse reinforcement learning. In: Twelfth Annual Conference of the International Speech Communication Association (2011)

12. Clarke, C.L., Craswell, N., Soboroff, I.: Overview of the trec 2009 web track. Tech. rep., WATERLOO UNIV (ONTARIO) (2009)
13. Clarke, C.L., Craswell, N., Voorhees, E.M.: Overview of the trec 2012 web track. Tech. rep., NATIONAL INST OF STANDARDS AND TECHNOLOGY GAITHERSBURG MD (2012)
14. Croft, W.B., Thompson, R.H.: I3r: A new approach to the design of document retrieval systems. Journal of the american society for information science **38**(6), 389–404 (1987)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. 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. 4171–4186 (2019)
16. El Asri, L., He, J., Suleman, K.: A sequence-to-sequence model for user simulation in spoken dialogue systems. Interspeech 2016 pp. 1151–1155 (2016)
17. Fagan, J.C.: Usability studies of faceted browsing: A literature review. Information Technology and Libraries **29**(2), 58–66 (2010)
18. Hearst, M.: Design recommendations for hierarchical faceted search interfaces. In: ACM SIGIR workshop on faceted search. pp. 1–5. Seattle, WA (2006)
19. Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., Yee, K.P.: Finding the flow in web site search. Communications of the ACM **45**(9), 42–49 (2002)
20. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems (TOIS) **20**(4), 422–446 (2002)
21. Kiesel, J., Bahrami, A., Stein, B., Anand, A., Hagen, M.: Toward voice query clarification. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 1257–1260. ACM (2018)
22. Kotov, A., Zhai, C.: Towards natural question guided search. In: Proceedings of the 19th international conference on World wide web. pp. 541–550 (2010)
23. Kules, B., Capra, R., Banta, M., Sierra, T.: What do exploratory searchers look at in a faceted search interface? In: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries. pp. 313–322 (2009)
24. Li, X., Lipton, Z.C., Dhingra, B., Li, L., Gao, J., Chen, Y.N.: A user simulator for task-completion dialogues. arXiv preprint arXiv:1612.05688 (2016)
25. Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., Vanderwende, L.: 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 (2016)
26. Papangelis, A., Papadakos, P., Kotti, M., Stylianou, Y., Tzitzikas, Y., Plexousakis, D.: Ld-sds: Towards an expressive spoken dialogue system based on linked-data
27. Rao, S., Daumé III, H.: 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 (2018)
28. Rao, S., Daumé III, H.: 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 (2019)
29. Salle, A., Villavicencio, A.: Incorporating subword information into matrix factorization word embeddings. In: Proceedings of the Second Workshop on Subword/Character LEvel Models. pp. 66–71. Association for Computational Linguistics, New Orleans (Jun 2018). https://doi.org/10.18653/v1/W18-1209
30. Sun, Y., Zhang, Y.: Conversational recommender system. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 235–244. ACM (2018)

31. Trippas, J.R., Spina, D., Cavedon, L., Joho, H., Sanderson, M.: Informing the design of spoken conversational search: perspective paper. In: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval. pp. 32–41. ACM (2018)
32. Tunkelang, D.: Faceted search. Synthesis lectures on information concepts, retrieval, and services **1**(1), 1–80 (2009)
33. Vandic, D., Aanen, S., Frasincar, F., Kaymak, U.: Dynamic facet ordering for faceted product search engines. IEEE Transactions on Knowledge and Data Engineering **29**(5), 1004–1016 (2017)
34. Weizenbaum, J., et al.: Eliza—a computer program for the study of natural language communication between man and machine. Communications of the ACM **9**(1), 36–45 (1966)
35. Wen, T., Vandyke, D., Mrkšíc, N., Gašíc, M., Rojas-Barahona, L., Su, P., Ultes, S., Young, S.: A network-based end-to-end trainable task-oriented dialogue system. In: 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017-Proceedings of Conference. vol. 1, pp. 438–449 (2017)
36. Yee, K.P., Swearingen, K., Li, K., Hearst, M.: Faceted metadata for image search and browsing. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 401–408 (2003)
37. Young, S.J.: 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 (2000)
38. Zamani, H., Lueck, G., Chen, E., Quispe, R., Luu, F., Craswell, N.: Mimics: A large-scale data collection for search clarification (2020)
39. Zhang, S., Balog, K.: Evaluating conversational recommender systems via user simulation. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1512–1520 (2020)