Challenges in the evaluation of conversational search systems

Penha, Gustavo; Hauff, Claudia

Publication date
2020

Document Version
Final published version

Published in
KDD 2020 Workshop on Conversational Systems Towards Mainstream Adoption, KDD-Converse 2020

Citation (APA)

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.
Challenges in the Evaluation of Conversational Search Systems

Gustavo Penha
TU Delft
g.penha-1@tudelft.nl

Claudia Hauff
TU Delft
c.hauff@tudelft.nl

ABSTRACT
The area of conversational search has gained significant traction in the IR research community, motivated by the widespread use of personal assistants. An often researched task in this setting is conversation response ranking, that is, to retrieve the best response for a given ongoing conversation from a corpus of historic conversations. While this is intuitively an important step towards (retrieval-based) conversational search, the empirical evaluation currently employed to evaluate trained rankers is very far from this setup: typically, an extremely small number (e.g., 10) of non-relevant responses and a single relevant response are presented to the ranker. In a real-world scenario, a retrieval-based system has to retrieve responses from a large (e.g., several millions) pool of responses or determine that no appropriate response can be found. In this paper we aim to highlight these critical issues in the offline evaluation schemes for tasks related to conversational search. With this paper, we argue that the currently in-use evaluation schemes have critical limitations and simplify the conversational search tasks to a degree that makes it questionable whether we can trust the findings they deliver.

ACM Reference Format:

1 INTRODUCTION

Conversational search is concerned with creating agents that fulfill an information need by means of a mixed-initiative conversation through natural language interaction, rather than the traditional turn-taking models exhibited in a traditional search engine’s results page. It is an active area of research (as evident for instance in the recent CAIR\(^1\) and SCAI\(^2\) workshop series) due to the widespread deployment of voice-based agents, such as Google Assistant and Microsoft Cortana. Voice-based agents are currently mostly used for simple closed domain tasks such as fact checking, initiating calls and checking the weather. They are not yet effective for conducting open domain complex and exploratory information seeking conversations [15].

\(^1\)https://sites.google.com/view/cair-wx/home
\(^2\)https://scai.info/

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

KDD Converse’20, August 2020,
© 2020 Copyright held by the owner/author(s).

2 TASKS AND EVALUATION SCHEMES

We describe next sub-tasks of the CSS pipeline. First let us consider Figure 1, where we display three different end-to-end CSS approaches. On the left a retrieval-based system uses the conversational context to select amongst a pool of responses the most adequate (conversation ranking tasks). On the center we have a generative model that directly generates the responses from the conversational context (conversation generation tasks). On the right the system encompasses a model to retrieve documents followed by a model that select spans in such documents (conversation question answering). Next we discuss common assumptions used when employing such tasks to evaluate models and we highlight their shortcomings. On Table 1 we describe popular benchmarks for conversational ranking tasks with statistic such as the number of response candidates, and on Table 2 we describe the relation between the premises discussed in this section and the tasks they relate to.

\(^3\)We do not consider here the online evaluation of conversational search systems, which although is more reliable than offline evaluation, it is expensive, time consuming and non-repeatable.
The task of conversation response ranking \([11, 17, 20, 21, 37, 38, 50, 59, 61, 62, 64, 67, 68]\) (also known as next utterance selection), concerns retrieving the best response given the dialogue context. Formally, let \(D = \{(U_i, R_i, Y_i)\}_{i=1}^N\) be a data set consisting of \(N\) triplets: dialogue context, response candidates and response relevance labels. The dialogue context \(U_i\) is composed of the previous utterances \(\{u^1, u^2, ..., u^t\}\) at the turn \(r\) of the dialogue. The candidate responses \(R_i = \{r^1, r^2, ..., r^K\}\) are either ground-truth responses or negative sampled candidates, indicated by the relevance labels \(Y_i = \{y^1, y^2, ..., y^K\}\). Typically, the number of candidates \(K \ll N\), where \(K\) is the number of available responses and by design the number of ground-truth responses is usually one, the observed response in the conversational data. The task is then to learn a ranking function \(f(.)\) that is able to generate a ranked list for the set of candidate responses \(R_i\) based on their predicted relevance scores \(f(U_i, r)\).

Other similar ranking tasks related to conversational search are clarification question retrieval \([42, 43]\), where the set of responses to be retrieved are always clarification questions, conversation document ranking \([37]\), where the item to be retrieved is a document that contains the answer to the dialogue context and conversion passage retrieval \([8, 31]\). A successful model for the ranking tasks retrieves the ground-truth response(s) first in the ranked list, and thus the evaluation metrics employed are standard IR metrics such as MAP and \(R_N@K\) (where \(N\) is the number of candidate responses and \(K\) is the list cutoff threshold).

### 2.1 Conversation Ranking Tasks

The task of conversation response generation, also known as dialogue generation \([1, 12, 28, 29, 53]\), is to generate a response given the dialogue context. Formally, let \(D = \{(U_i, r_i)\}_{i=1}^N\) be a data set consisting of \(N\) tuples: dialogue context and response. The dialogue

---

**Table 1: Overview of conversational search benchmarks for which the task requires ranking.**

<table>
<thead>
<tr>
<th>Task</th>
<th>Benchmark</th>
<th># of candidates</th>
<th>average # of relevant</th>
<th>negative sampling procedure</th>
<th>relevant is always present</th>
<th>independent instances</th>
<th>maximum eval. metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversation</td>
<td>MSDialog [39]</td>
<td>10</td>
<td>1</td>
<td>scoring function</td>
<td>yes</td>
<td>yes</td>
<td>0.836 MAP [61]</td>
</tr>
<tr>
<td></td>
<td>E-commerce [67]</td>
<td>10</td>
<td>1</td>
<td>scoring function</td>
<td>yes</td>
<td>yes</td>
<td>0.704 R_{10}@1 [16]</td>
</tr>
<tr>
<td></td>
<td>UDC [39]</td>
<td>10</td>
<td>1</td>
<td>random</td>
<td>yes</td>
<td>yes</td>
<td>0.855 R_{10}@1 [16]</td>
</tr>
<tr>
<td></td>
<td>Doublan [59]</td>
<td>10</td>
<td>1.18</td>
<td>scoring function</td>
<td>yes</td>
<td>yes</td>
<td>0.619 MAP [16]</td>
</tr>
<tr>
<td></td>
<td>MANtIS [37]</td>
<td>11</td>
<td>1</td>
<td>scoring function</td>
<td>yes</td>
<td>yes</td>
<td>0.733 MAP [38]</td>
</tr>
<tr>
<td></td>
<td>MANtIS [37]</td>
<td>51</td>
<td>1</td>
<td>scoring function</td>
<td>yes</td>
<td>yes</td>
<td>0.519 MAP [37]</td>
</tr>
<tr>
<td></td>
<td>PolyAI-AQA [20]</td>
<td>100</td>
<td>1</td>
<td>random</td>
<td>yes</td>
<td>yes</td>
<td>0.843 R_{100}@1 [21]</td>
</tr>
<tr>
<td></td>
<td>PolyAI-Reddit [20]</td>
<td>100</td>
<td>1</td>
<td>random</td>
<td>yes</td>
<td>no</td>
<td>0.718 R_{100}@1 [21]</td>
</tr>
<tr>
<td></td>
<td>DSTC7-NOESIS-5 [18]</td>
<td>100</td>
<td>1</td>
<td>random</td>
<td>yes</td>
<td>no</td>
<td>0.822 MRR [18]</td>
</tr>
<tr>
<td></td>
<td>DSTC7-NOESIS-2 [18]</td>
<td>120,000</td>
<td>1</td>
<td>random</td>
<td>yes</td>
<td>no</td>
<td>0.253 MRR [18]</td>
</tr>
<tr>
<td>Conversation</td>
<td>MANtIS [37]</td>
<td>11</td>
<td>1.13</td>
<td>scoring function</td>
<td>yes</td>
<td>yes</td>
<td>0.672 MAP [2]</td>
</tr>
<tr>
<td>doc. ranking</td>
<td>MANtIS [37]</td>
<td>50</td>
<td>1.13</td>
<td>scoring function</td>
<td>yes</td>
<td>yes</td>
<td>0.487 MAP [2]</td>
</tr>
<tr>
<td>Clarifying question</td>
<td>StackExchange [43]</td>
<td>10</td>
<td>1</td>
<td>scoring function</td>
<td>yes</td>
<td>yes</td>
<td>0.492 MAP [43]</td>
</tr>
</tbody>
</table>

---

\(\text{df} \) we do not include TREC CAsT 2019 in 1, since it differs from other datasets by doing TREC style pooling and judgements.
context $\mathcal{U}_i$ is composed of the previous utterances \{u$_1$, u$_2$, ..., u$_T$\} at the turn $\tau$ of the dialogue. The response $r_i$ is the u$_{\tau+1}$ utterance, i.e., the ground-truth. The task is then to learn a model $f(.)$ that is able to generate the response $r_i$ based on the dialogue context $\mathcal{U}_i$. The majority of the research conducted in response generation relies on data sets that are not information-seeking, e.g., movies subtitles or chit-chat [10, 45, 55, 66].

Other generation tasks from conversational search that share the same evaluation scheme of conversation response generation are clarification question generation [44], the response is generated on the go, and query reformulation [8], suggestions of follow-up queries are generated. The evaluation of generative models relies on word-overlap metrics inspired by machine translation, e.g., BLEU [36], or text summarization, e.g., ROUGE [30]. Such metrics have been extensively studied and criticized by the natural language processing (NLP) community. There is empirical evidence that word-overlap metrics do not correlate well with human judgments [32].

The complexity of the generation task evaluation is so high that it is common to resort to expensive human evaluation, through crowd-sourcing, lab experiments or in-field experiments [9].

### 2.3 Conversational Question Answering

This task is also known as conversational machine reading comprehension [7, 40, 46], and it concerns selecting spans from a document as a response to the dialogue context. Formally, let $\mathcal{D} = \{(\mathcal{U}_i, a_i, p_i)\}_{i=1}^N$ be a data set consisting of $N$ triplets: dialogue context, the answer span and the context passage. The dialogue context $\mathcal{U}_i$ is composed of the previous utterances \{u$_1$, u$_2$, ..., u$_T$\} at the turn $\tau$ of the dialogue. The answer span $a_i$ is composed of the ground-truth start and finish indexes of the passage $p_i$ containing the correct answer. The task is then to learn a model $f(.)$ that is able to predict the answer span $a_i$ based on the dialogue context $\mathcal{U}_i$ and the context passage $p_i$. The evaluation is based on the amount of words that are correct in the selected span, using classification metrics such as the F-score. Extractive (span-based) question answering is a very similar NLP task, for which similar problems arise, such as unanswerable questions [41].

### 2.4 Premises and Limitations

**I** There is a complete pool of adequate responses that endure over time. Our ranking tasks assume access to a pool of responses that contains at least one appropriate answer to a given information need. If we resort only to historical responses the maximum effectiveness of a system would be very low. For example, in popular benchmarks such as UDC [33] and MSDialog [39] the number of responses that are exact matches with historical responses are less than 11% and 2% respectively. As we see at Figure 2, most conversations have only 50–60% words match, when compared to the most similar historical response. This indicates that the maximum accuracy achieved by a real-world system would be small, since only the responses that semantically match a previous one can be employed effectively. We also see that such exact matches are often uninformative: 40% are utterances for which the intent is to show gratitude, e.g. ‘Thank you!’, compared to the 20% overall rate in MSDialog. Another concern is that responses that were never given before, e.g. questions about a recent Windows update, would not be answerable by such a system even though this information might be available on the web.

**II** The correct answer is always in the candidate responses list. Neural ranking models are generally employed for the task of re-ranking a set of documents in adhc retrieval, obtained from a recall-oriented and efficient first stage ranker [65]. While such multi-stage approach offers a practical approach for conversational response ranking, 12 of 13 benchmarks analyzed at Table 1 always include the relevant response in the small candidate list to be retrieved and none require models to do full corpus retrieval.

**III** The effectiveness of models for small candidate lists generalize to large collections. While in adhoc retrieval we have to rank from a pool of millions of documents, current benchmarks require models to retrieve responses from a list of 10–100 candidates (12 out of 13 use less than 100 candidates, and 7 use only 10 candidates). This makes the task unreasonably easy, as demonstrated by

### Table 2: Premises and respective conversational tasks.

<table>
<thead>
<tr>
<th>Premise</th>
<th>Holds for</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I) There is a complete pool of adequate responses that endure over time.</td>
<td>Conversation ranking tasks</td>
</tr>
<tr>
<td>(II) The correct answer is always in the candidate responses list.</td>
<td>Conversation ranking tasks and conversational QA</td>
</tr>
<tr>
<td>(III) The effectiveness of models for small candidate lists generalize to large collections.</td>
<td>Conversation ranking tasks</td>
</tr>
<tr>
<td>(IV) Test instances from the same dialogue are considered as independent.</td>
<td>Conversation ranking tasks, conversation generation tasks and conversational QA</td>
</tr>
<tr>
<td>(V) There is only one adequate answer.</td>
<td>Conversation ranking tasks, conversation generation tasks and conversational QA</td>
</tr>
</tbody>
</table>

Figure 2: Amount of word intersection between response and the most similar historical responses.
the 80% drop in performance from subtask 5 (120000 candidates) and subtask 2 (100 candidates) of DTSC7-NOESIS [18]. Additionally, 5 of the 13 tasks sample instances randomly as opposed to using a scoring function such as BM25, making the task even easier as evidenced by the higher maximum evaluation metrics for random negative sampled benchmarks in Table 1.

(IV) Test instances from the same dialogue are considered as independent. When creating conversational datasets [20, 33, 39] the default is to generate multiple instances from one dialogue: one instance for each answer provided by the information provider composed of the last information seeker utterance, and the dialogue history. Even when multiple utterances come from the same dialogue, they are evaluated independently, e.g. an inappropriate response in the beginning of a conversation does not change the evaluation of a response given later by the system in the same dialogue. All benchmarks analyzed in Table 1 evaluate instances from the same dialogue independently. In a real-world scenario, if a model fails in the start of the conversation, it has to recover from unsatisfactory responses.

(V) There is only one adequate answer. Traditional offline evaluation cannot handle counterfactuals [3] such as what would have happened if another response was given instead of the ground-truth one. Due to the high cost of human labels, it is common to use only one relevant response per context (the observed human response). However, multiple responses could be correct to a given context with different levels of relevance. Multiple answers can be right because they provide semantically similar responses or because they are different but appropriate responses to an information-need.

3 CONCLUSION AND FUTURE DIRECTIONS

In this paper we argue that current evaluation schemes in conversational search, as instantiated through popular tasks and benchmarks, are extreme simplifications of the actual problem. Based on our observations, we encourage work on the following directions for each of the issues we described:

• (I) Creation of a pool of responses: creation of a comprehensive pool of responses from historical responses and other sources e.g. creating responses from web documents. Study whether hybrids of ranking and generation [60] that generate the pool of responses to be ranked is a viable alternative to using only historical responses.

• (II) Handling dialogue contexts that are unanswerable: study the effect of candidate lists for which no adequate response to the dialogue context exist, and how to automatically detect such cases, e.g. through a prediction model [19], none-of-the-above prediction [14, 18, 56] and uncertainty estimation [51]. Detecting if the current information-need still needs further clarification and elucidation in order to make it answerable is also an important research direction.

• (III) Ranking beyond 100 responses: methods for effective retrieval from the entire pool of responses such as multi-stage approaches that apply a recall-oriented first stage ranker [65]. Traditional IR methods which are efficient might not be effective for retrieval of responses to be re-ranked [49]. Investigations of the effectiveness of conversational search tasks for large corpus retrieval, i.e. the generality effect [48].

• (IV) Take into account the dialogue evolution: When evaluating a model for retrieving responses, instead of having several independent instances for each dialogue (one for each information-provider response), consider a dialogue uniquely. For instance by introducing evaluation metrics that take into account the other responses from the same dialogue given by the system, e.g. ranking relevant responses in the initial turns of the dialogue leads to higher gains than ranking relevant responses in the last turns of the dialogue.

• (V-a) Expanding the number of relevant responses: how to expand the number of relevant response candidates, e.g. paraphrases [18] and adversarial examples [24], for information-seeking conversations. In IR, the evaluation in a scenario of limited relevance judgments has been studied [4, 63].

• (V-b) Counterfactual evaluation of dialogue: how to tell what would have happened if answer B was given instead of A, when there is no relevance label for answer B [22]. For example, Carterette and Allan [6] proposed a evaluation scheme that takes advantage of the similarity between documents with the intuition that closely associated documents tend to be relevant to the same information need.

Acknowledgements. This research has been supported by NWO projects SearchX (639.022.722) and NWO Aspasia (015.013.027).

REFERENCES


