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.

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].

Existing efforts in conversational search have started in late 1970's, with a dialogue-based approach for reference retrieval [35]. Since then, research in IR has focused on strategies—such as exploiting relevance feedback [47], query suggestions [5] and exploratory search [34, 58]—to make the search engine result page more interactive, which can be considered as a very crude approach to conversational search systems. User studies [13, 23, 25, 52, 54, 57] have been conducted to understand how people interact with agents (simulated by humans) and inform the design of CSSs.

A popular approach to conversational search is retrieval-based: given an ongoing conversation and a large corpus of historic conversations, retrieve the response that is best suited from the corpus (i.e., conversation response ranking [38, 59, 61, 62]). This retrieval-based approach does not require task-specific semantics by domain experts [21], and it avoids the difficult task of dialogue generation, which often suffers from uninformative, generic responses [26] or responses that are incoherent given the dialogue context [27]. However, the current offline[3] benchmarks (cf. Table 1) for conversation response ranking are overly simplified: they mostly require models to retrieve the correct response from a small set of 10 candidates.

In this paper we first formally describe the three main approaches to CSS based on previous work on conversational search. We then take a critical look at the premises of their offline evaluation schemes, e.g. '*The correct response is always in the candidates list.*', discuss their implications and suggest future directions to cope with their limitations.

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.

2.1 Conversation Ranking Tasks

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 $\mathcal{D} = \{(\mathcal{U}_i, \mathcal{R}_i, \mathcal{Y}_i)\}_{i=1}^{N}$ be a data set consisting of N triplets: dialogue context, response candidates and response relevance labels. The dialogue context $\mathcal{U}_i$ is composed of the previous utterances $\{u^1, u^2, ..., u^\tau\}$ at the turn τ of the dialogue. The candidate responses $\mathcal{R}_i = \{r^1, r^2, ..., r^k\}$ are either ground-truth

[1] https://sites.google.com/view/cair-ws/home
[2] https://scai.info/

[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.

Converse'20, August 24, 2020, San Diego, CA

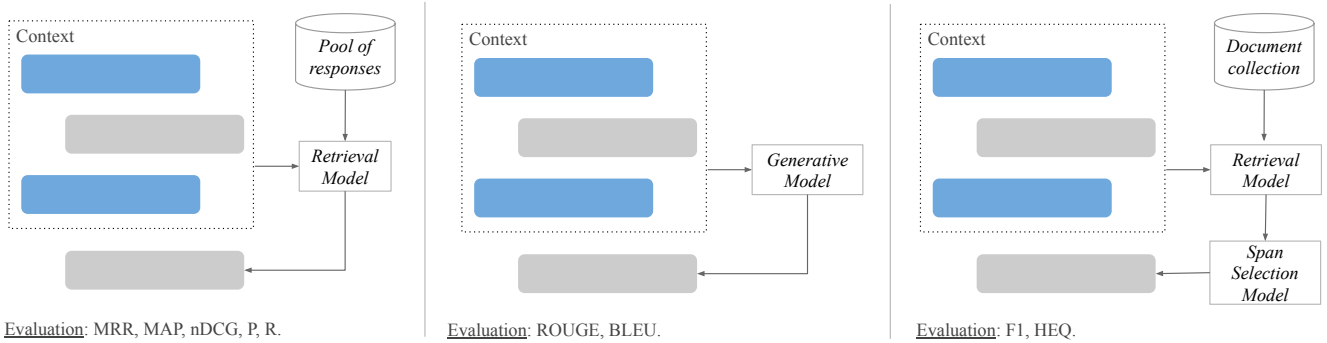

Figure 1: Different approaches for conversational search systems. The inputs are the previous utterances in the conversation (the context $\mathcal{U}$) and the model output is the system response r. They encompass, from left to right respectively, the conversational search tasks of ranking, generation and question answering employed in a end-to-end system.

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

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

responses or negative sampled candidates, indicated by the relevance labels $\mathcal{Y}_i = \{y^1, y^2, ..., y^k\}$. Typically, the number of candidates $k \ll K$, 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 $\mathcal{R}_i$ based on their predicted relevance scores $f(\mathcal{U}_i, r)$.

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 *conversation passage retrieval* [8, 31][4]. 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).

[4]We do not include TREC CAsT 2019 in 1, since it differs from other datasets by doing TREC style pooling and judgements.

2.2 Conversation Generation 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 $\mathcal{D} = \{(\mathcal{U}_i, r_i)\}_{i=1}^{N}$ be a data set consisting of N tuples: dialogue context and response. The dialogue context $\mathcal{U}_i$ is composed of the previous utterances $\{u^1, u^2, ..., u^\tau\}$ at the turn τ 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

Table 2: Premises and respective conversational tasks.

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

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 $\{u^1, u^2, ..., u^\tau\}$ at the turn τ 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.

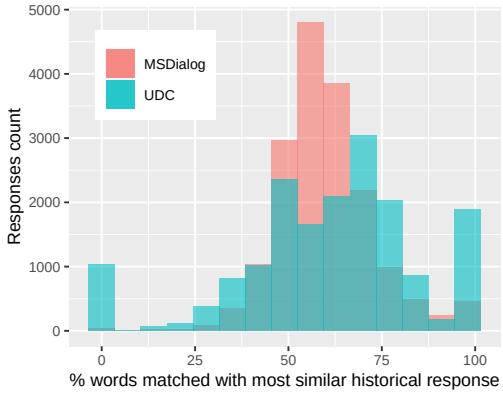

Figure 2: Amount of word intersection between response and the most similar historical responses.

(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 adhoc 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 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 though 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 performance prediction [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

[1] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. arXiv:cs.CL/2001.09977
[2] Alex Bălan. 2019. MANtIS: a novel information seeking dialogues dataset. *Master's thesis, TU Delft* (2019).
[3] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *JMLR* 14, 1 (2013), 3207–3260.
[4] Chris Buckley and Ellen M Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 25–32.
[5] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. 2008. Context-aware query suggestion by mining click-through and session data. In *SIGKDD*. 875–883.
[6] Ben Carterette and James Allan. 2007. Semiautomatic evaluation of retrieval systems using document similarities.. In *CIKM*, Vol. 7. 873–876.
[7] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036* (2018).
[8] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. Cast 2019: The conversational assistance track overview. In *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC*. 13–15.
[9] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2019. Survey on Evaluation Methods for Dialogue Systems. *arXiv preprint arXiv:1905.04071* (2019).
[10] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241* (2018).
[11] Jianxiong Dong and Jim Huang. 2018. Enhance word representation for out-of-vocabulary on ubuntu dialogue corpus. *arXiv preprint arXiv:1802.02614* (2018).
[12] Jiachen Du, Wenjie Li, Yulan He, Ruifeng Xu, Lidong Bing, and Xuan Wang. 2018. Variational autoregressive decoder for neural response generation. In *EMNLP*. 3154–3163.
[13] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, and Sylvain Daronnat. 2018. Investigating how conversational search agents affect user's behaviour, performance and search experience. In *CAIR*.
[14] Yulan Feng, Shikib Mehri, Maxine Eskenazi, and Tiancheng Zhao. 2020. "None of the Above":Measure Uncertainty in Dialog Response Retrieval. arXiv:cs.CL/2004.01926
[15] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational AI. In *SIGIR*. 1371–1374.
[16] Jia-Chen Gu, Tianda Li, Quan Liu, Xiaodan Zhu, Zhen-Hua Ling, Zhiming Su, and Si Wei. 2020. Speaker-Aware BERT for Multi-Turn Response Selection in Retrieval-Based Chatbots. *arXiv preprint arXiv:2004.03588* (2020).
[17] Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2019. Utterance-to-Utterance Interactive Matching Network for Multi-Turn Response Selection in Retrieval-Based Chatbots. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2019), 369–379.
[18] Chulaka Gunasekara, Jonathan K Kummerfeld, Lazaros Polymenakos, and Walter Lasecki. 2019. Dstc7 task 1: Noetic end-to-end response selection. In *Proceedings of the First Workshop on NLP for Conversational AI*. 60–67.
[19] Claudia Hauff. 2010. Predicting the effectiveness of queries and retrieval systems. In *SIGIR Forum*, Vol. 44. 88.

[20] Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, et al. 2019. A Repository of Conversational Datasets. *arXiv preprint arXiv:1904.06472* (2019).

[21] Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Ivan Vulić, et al. 2019. ConveRT: Efficient and Accurate Conversational Representations from Transformers. *arXiv preprint arXiv:1911.03688* (2019).

[22] Thorsten Joachims and Adith Swaminathan. 2016. Counterfactual evaluation and learning for search, recommendation and ad placement. In *SIGIR*. 1199–1201.

[23] Hyunhoon Jung, Changhoon Oh, Gilhwan Hwang, Cindy Yoonjung Oh, Joonhwan Lee, and Bongwon Suh. 2019. Tell Me More: Understanding User Interaction of Smart Speaker News Powered by Conversational Search. In *CHI*. 1–6.

[24] Anjuli Kannan and Oriol Vinyals. 2017. Adversarial evaluation of dialogue models. *arXiv preprint arXiv:1701.08198* (2017).

[25] Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. 2018. Toward voice query clarification. In *SIGIR*. 1257–1260.

[26] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL*. 110–119.

[27] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A Persona-Based Neural Conversation Model. In *ACL*. 994–1003.

[28] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep Reinforcement Learning for Dialogue Generation. In *EMNLP*. 1192–1202.

[29] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial Learning for Neural Dialogue Generation. In *EMNLP*. 2157–2169.

[30] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81.

[31] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Query Reformulation using Query History for Passage Retrieval in Conversational Search. arXiv:cs.CL/2005.02230

[32] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP*. 2122–2132.

[33] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909* (2015).

[34] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[35] Robert N Oddy. 1977. Information retrieval through man-machine dialogue. *Journal of documentation* 33, 1 (1977), 1–14.

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*. 311–318.

[37] Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. Introducing MANtIS: a novel Multi-Domain Information Seeking Dialogues Dataset. *arXiv preprint arXiv:1912.04639* (2019).

[38] Gustavo Penha and Claudia Hauff. 2019. Curriculum Learning Strategies for IR: An Empirical Study on Conversation Response Ranking. *arXiv preprint arXiv:1912.08555* (2019).

[39] Chen Qu, Liu Yang, W Bruce Croft, Johanne R Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and characterizing user intent in information-seeking conversations. In *SIGIR*. 989–992.

[40] Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W Bruce Croft, and Mohit Iyyer. 2019. Attentive History Selection for Conversational Question Answering. In *CIKM*. 1391–1400.

[41] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 784–789. https://doi.org/10.18653/v1/P18-2124

[42] Sudha Rao. 2017. Are you asking the right questions? Teaching Machines to Ask Clarification Questions. In *ACL Student Research Workshop*. 30–35.

[43] Sudha Rao and Hal Daumé III. 2018. Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information. In *ACL*. 2737–2746.

[44] Sudha Rao and Hal Daumé III. 2019. Answer-based Adversarial Training for Generating Clarification Questions. In *NAACL*. 143–155.

[45] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207* (2018).

[46] Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *ACL* 7 (2019), 249–266.

[47] Joseph John Rocchio. 1971. Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing* (1971), 313–323.

[48] Gerard Salton. 1972. The "generality" effect and the retrieval evaluation for large collections. *Journal of the American Society for Information Science* 23, 1 (1972), 11–22.

[49] Lifeng Shang, Tetsuya Sakai, Zhengdong Lu, Hang Li, Ryuichiro Higashinaka, and Yusuke Miyao. 2016. Overview of the NTCIR-12 Short Text Conversation Task.. In *NTCIR*.

[50] Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. Multi-Representation Fusion Network for Multi-Turn Response Selection in Retrieval-Based Chatbots. In *WSDM*. 267–275.

[51] Christopher Tegho, Paweł Budzianowski, and Milica Gašić. 2017. Uncertainty Estimates for Efficient Neural Network-based Dialogue Policy Optimisation. arXiv:stat.ML/1711.11486

[52] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *CAIR*.

[53] Zhiliang Tian, Wei Bi, Xiaopeng Li, and Nevin L Zhang. 2019. Learning to Abstract for Memory-augmented Conversational Response Generation. In *ACL*. 3816–3825.

[54] Johanne R Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *CHIIR*. 325–328.

[55] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* (2015).

[56] Ellen M Voorhees. 2001. Overview of the TREC-9 question answering track. In *In Proceedings of the Ninth Text REtrieval Conference (TREC-9*. Citeseer.

[57] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *CHI EA*. 2187–2193.

[58] Ryen W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–98.

[59] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In *ACL*. 496–505.

[60] Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A Hybrid Retrieval-Generation Neural Conversation Model. In *CIKM*. 1341–1350.

[61] Liu Yang, Minghui Qiu, Chen Qu, Cen Chen, Jiafeng Guo, Yongfeng Zhang, W Bruce Croft, and Haiqing Chen. 2020. IART: Intent-aware Response Ranking with Transformers in Information-seeking Conversation Systems. *arXiv preprint arXiv:2002.00571* (2020).

[62] Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *SIGIR*. 245–254.

[63] Emine Yilmaz and Javed A Aslam. 2006. Inferred AP: estimating average precision with incomplete judgments. In *CIKM*. 102–111.

[64] Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop Selector Network for Multi-turn Response Selection in Retrieval-based Chatbots. In *EMNLP*. 111–120.

[65] Hamed Zamani, Mostafa Dehghani, W Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *CIKM*. 497–506.

[66] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243* (2018).

[67] Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling Multi-turn Conversation with Deep Utterance Aggregation. In *ACL*. 3740–3752.

[68] Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *ACL*. 1118–1127.