Hierarchical Dependence-aware Evaluation Measures for Conversational Search

Guglielmo Faggioli
University of Padova
guglielmo.faggioli@phd.unipd.it

Marco Ferrante
University of Padova
ferrante@math.unipd.it

Nicola Ferro
University of Padova
ferro@dei.unipd.it

Raffaele Perego
National Research Council
raffaele.perego@isti.cnr.it

Nicola Tonellotto
University of Pisa
nicola.tonellotto@unipi.it

ABSTRACT

Conversational agents are drawing a lot of attention in the information retrieval (IR) community also thanks to the advancements in language understanding enabled by large contextualized language models. IR researchers have long ago recognized the importance of a sound evaluation of new approaches. Yet, the development of evaluation techniques for conversational search is still an underlooked problem. Currently, most evaluation approaches rely on procedures directly drawn from ad-hoc search evaluation, treating utterances in a conversation as independent events, as if they were just separate topics, instead of accounting for the conversation context. We overcome this issue by proposing a framework for defining evaluation measures that are aware of the conversation context and the utterance semantic dependencies. In particular, we model the conversations as Direct Acyclic Graphs (DAG), where self-explanatory utterances are root nodes, while anaphoric utterances are linked to sentences that contain their missing semantic information. Then, we propose a family of hierarchical dependence-aware aggregations of the evaluation metrics driven by the conversational graph. In our experiments, we show that utterances from the same conversation are 20% more correlated than utterances from different conversations. Thanks to the proposed framework, we are able to include such correlation in our aggregations, and be more accurate when determining which pairs of conversational systems are deemed significantly different.

CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**; **Retrieval effectiveness**.

KEYWORDS

Evaluation; Conversational Search; Conversation modelling;

ACM Reference Format:
Guglielmo Faggioli, Marco Ferrante, Nicola Ferro, Raffaele Perego, and Nicola Tonellotto. 2021. Hierarchical Dependence-aware Evaluation Measures for Conversational Search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3463090

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463090

1 INTRODUCTION

The *Information Retrieval (IR)* scientific community widely acknowledges the importance of sound, theoretically well-founded and generalizable evaluation protocols. Researchers devoted a lot of effort in developing sound methodologies, measures and guidelines to correctly assess and compare the performance of different systems [12, 22]. Moreover, it is recognized that an evaluation approach developed to measure the performance of a system on a specific task might not be suited to assess the performance on a different task [5, 17]. In principle, each IR task requires an ad-hoc evaluation methodology. On the other hand, the development of conversational systems has recently received a lot of attention from the IR community. The availability of effective machine learned methods and contextualized language models such as BERT, have in fact enabled the development of brand new agents capable of seemingly converse with the user in a multi-turn interaction made of a sequence of utterances and answers. Nevertheless, the evaluation of such conversational systems is still struggling to reach satisfying results. In particular, at the current time, the procedures to evaluate the quality of conversational systems are still in their seminal state. Typically, researchers adopt techniques drawn from machine translation, machine learning or ad-hoc search evaluation that difficultly capture the specificity of the task and the differences among conversational systems. Conversational systems are instead peculiar and the tasks they address cannot be assimilated to other IR tasks. Some systems are focused mainly on the management of the dialogue with the user [26, 27], but from an IR perspective we are more interested to task-oriented conversational systems [3, 13, 19, 23, 25]. In these systems, the user has a final goal in mind, for example purchasing something or learning about a specific topic. The system replies to user utterances driving the user through the dialogue toward satisfying her need. Due to their complexity, user studies are the best evaluation strategy for such conversational agents [8, 9, 11, 14, 15]. However, user studies tend to be very expensive and an offline evaluation is commonly preferred, at least in the first phases of the development of new conversational systems.

An issue that impairs the offline evaluation of the conversational agents is linked to the availability of public datasets. Researchers however recently invested a lot of effort in the construction of publicly available collections [6, 18, 20]. Among these initiatives, the

TREC campaign on Conversational Assistance Track (CAsT) [7] notably promote good practices for a fair offline evaluation of different systems. Even for this initiative however, utterances are evaluated independently by using *Normalized Discounted Cumulated Gain (nDCG)* with a cut-off at 3. After that, to score the system at the conversation level, they consider either mean-based or turn-based aggregation approaches. The former simply computes the mean score on all utterances. The latter weights the measure according to the distance of the utterance from the beginning of the conversation. The turn-based measure relies on the assumption that the further the utterance is from the beginning of the conversation, the harder it is to keep track of the context of the conversation and effectively retrieve documents useful to answer the utterance. The community has recognized the drawbacks linked to adapting traditional evaluation approaches to the conversational setup [1, 16, 19]. In some cases a low correlation between traditional metrics and user satisfaction has been observed [2, 16]. We note that often utterances of the same conversation share the same context. Different utterances might be connected to the same semantic entity or have common relevant documents. This gives rise to inter-dependencies between utterances that might invalidates most of the statistical evaluation approaches. To solve the above-mentioned limitations, this work proposes a novel evaluation framework. The framework relies on the concept of "conversation graph": each conversation can be modelled as a *Directed Acyclic Graph (DAG)*. In detail, each utterance is linked to the previous ones containing the contextual information that can help understand the utterance itself. Using such graph we define a family of measures, dubbed *Hierarchical Dependency-aware Aggregation (HDA)*, which relies on the topology of the conversational graph to compute the performance for each conversation. The proposed approach has two main advantages over the state-of-the-art of conversational systems evaluation: *i)* By excluding the temporal component of utterance issuing, it allows generalization to multiple conversations on the same topic; *ii)* it accounts for the intrinsic interdependence between utterances that cause over-estimation of the performance. The remainder of this work is organized as follows: Section 2 describes our approach for the annotation of the conversations and the recursive evaluation framework. Section 3 shows our empirical results, while Section 4 concludes the paper.

2 THE PROPOSED FRAMEWORK

2.1 Annotating the Collection

A task-driven conversation between a human and an agent is composed of a sequence of user utterances. These utterances express the (faceted) information need that the user desires to satisfy, while the system tries to answer it. Utterances are either self-contained or contextual. A self-contained utterance states a clear intent, and can be submitted directly to the system to search for the answer. Contextual utterances, on the other hand, are not so straightforward to be answered. They contain references to entities and concepts mentioned in the previous utterances. We typically identify such utterances as "anaphoric" or "elliptic". Anaphoric utterances include terms e.g., pronouns, that explicitly reference previously cited elements of the conversations. Elliptic utterances include missing elements inferable by the context. Based on such classification of the utterances, we can model a conversation as a graph highlighting the dependencies among utterances. More formally, a conversation can be defined as a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of utterances in the conversation, while $\mathcal{E}$ is a set of direct edges modelling the dependencies between pairs of utterances. In particular $\mathcal{E} = \{(u, v) | u, v \in \mathcal{V}$, iff u contains the context to understand $v\}$. Note that $\mathcal{G}$ is not necessarily a connected graph: we can easily imagine cases in which the conversation involves a topic drift where the user swiftly changes the context. In such a case, two sets of utterances refer to two different contexts and the corresponding subgraphs will likely be not connected. Furthermore, we observe that the most common structure for $\mathcal{G}$ or its subgraphs is a tree. Given an information need, the user starts the conversation with a self-explanatory utterance. Since the information searched for can be complex and faceted, the dialogue generally evolves through a sequence of further utterances investigating aspects related to the context of the first one. Such conversations, very in conversational datasets, clearly originate graphs with the shape of a tree.

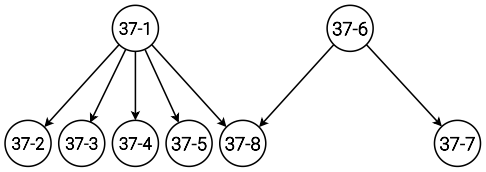


Figure 1: Graph originated by the conversation reported in Table 1. Links indicate a contextual dependence (e.g., anaphoric or elliptical) between utterance nodes.

Table 1: Coversation 37 of the CAsT 2019 collection.

id	prison psychology studies
37-1	What was the Stanford Experiment?
37-2	What did it show?
37-3	Tell me about the author of the experiment.
37-4	Was it ethical?
37-5	What are other similar experiments?
37-6	What happened in the Milgram experiment?
37-7	Why was it important?
37-8	What were the similarities and differences between the studies?

For the following experimental analysis, authors manually annotated the conversations available in CAsT 2019. Table 1 and Figure 1 depict a conversation included in such dataset and the corresponding graph annotation, respectively. The task of building automatically such graphs is a complex task that requires human supervision. In future, we can imagine the usage of machine learning algorithms to construct the conversation graph. Nowadays, the collections could be annotated by the practitioner or directly at construction time by assessors when judging the documents.

2.2 The Hierarchical Dependence-aware Aggregation

Now we propose an approach to aggregate the utterances scores by exploiting the conversational graph. Our approach is dubbed *Hierarchical Dependency-aware Aggregation (HDA)* since it accounts for semantic and contextual hierarchical dependencies between utterances to weight their scores in the global aggregation. The HDA framework relies on propagating the performance information, according to an arbitrary evaluation measure, for an utterance

to its neighbours in the conversational graph. Such information can flow either from children to parents or in the opposite direction: we will refer to the two situations with HDA-backward (HDA-b) and HDA-forward (HDA-f), respectively. Furthermore, HDA can be instantiated with any evaluation measure. Throughout our experiments, we use nDCG@3, which is the reference measure for CAsT 2019 [7]. Let $\mathcal{U}$ be the set of utterances, with u indicating a single utterance. Let $\mathcal{S}$ be the set of conversational systems to be compared, with s representing a single system. We identify with C_u and P_u the children and parents of the utterance u in the graph of the conversation as defined in Section 2.1. Furthermore, we define $m_s(u)$ the value of an arbitrary IR measure m on utterance u for system s. Hereinafter we assume measure m to be normalized with 1 indicating the highest performance. We consider $m_s(u)$ the probability that the user information need has been completely satisfied thanks to the answers provided by s for the utterance u. Therefore, $1 - m_s(u)$ corresponds to the probability for the user information need to be not satisfied. The first formulation of our framework, HDA-backward induced by measure m (HDA-b_m), accumulates the performance information on the root nodes of the conversational graph. In HDA-b_m, $1 - m_s(u)$ corresponds to the probability that the user will pose new (contextually linked) utterances C_u after u to solve her need. With probability $1 - m_s(u)$ the user will gain additional information about the current conversation topic in a measure that depends on how utterances in C_u are answered. Therefore, we define the gain $\mathrm{g}_s(u)$ experienced on u using s as:

$$\mathrm{g}_s(u) = \begin{cases} m_s(u) & \text{if } C_u = \emptyset, \\ m_s(u) + (1 - m_s(u)) \cdot \frac{\sum_{v \in C_u} \mathrm{g}_s(v)}{|C_u|} & \text{otherwise.} \end{cases} \quad (1)$$

For a single conversation C, the global score for s is computed as:

$$\text{HDA-}b_m(C, s) = \frac{1}{|C_r|} \sum_{u \in C_r} \mathrm{g}_s(u), \quad (2)$$

where C_r are the root utterances of the conversational graph. Conversely, HDA-f_m aggregates the performance information on leaves nodes. Ideally, we can assume that, for an utterance u, part of the $1-m_s(u)$ information that is missing from the complete satisfaction of the user is available in the parents P_u of u. In fact, we can expect that parent utterances will be more general and contain pieces of information relevant also to the children. In this sense, $1 - m_s(u)$ is the probability that the user will gain further information by considering the parents of an utterance. Similarly to HDA-b_m, the gain for HDA-f_m is defined as:

$$\mathrm{g}_s(u) = \begin{cases} m_s(u) & \text{if } P_u = \emptyset, \\ m_s(u) + (1 - m_s(u)) \cdot \frac{\sum_{v \in P_u} \mathrm{g}_s(v)}{|P_u|} & \text{otherwise.} \end{cases} \quad (3)$$

For a single conversation C, the global score is then computed as:

$$\text{HDA-}f_m(C, s) = \frac{1}{|C_l|} \sum_{u \in C_l} \mathrm{g}_s(u), \quad (4)$$

where C_l are the leaf utterances of the conversational graph. As a rule-of-thumb, the backward approach HDA-b_m should be favoured in settings where conversations are "exploratory": root nodes contain the overall description of the conversation, while leave utterances regard highly-specific sub-aspects of the topic. Conversely, HDA-f_m should be preferred when the user starts with a generic or unclear information need and adjusts it by gaining new information, such as in the Q&A scenario.

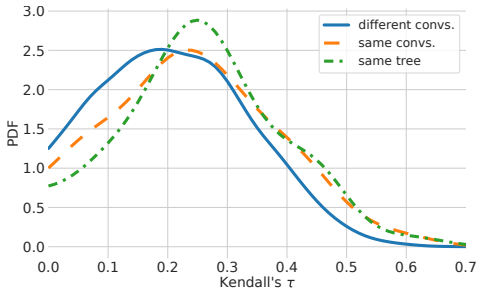


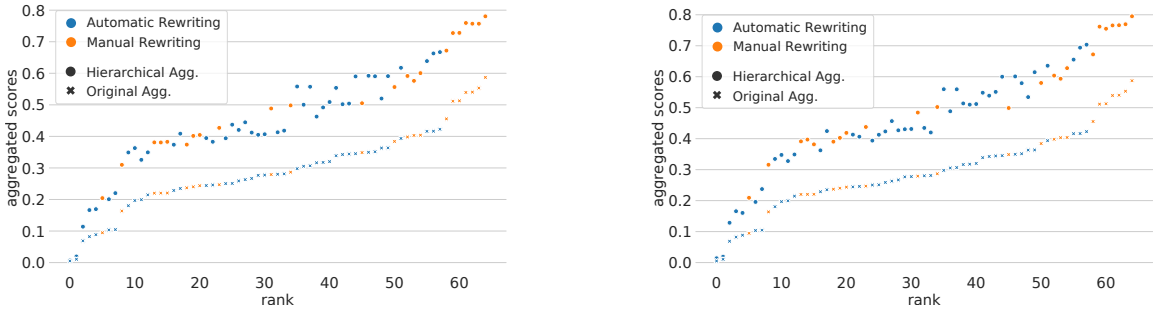
Figure 2: Probability Density Function of the Kendall's τ correlation between pairs of utterances of different groups.

3 EXPERIMENTS

To assess its behaviour, we apply the proposed framework to the CAsT 2019 Collection [7] and the 65 runs submitted originally to the TREC track[1]. The collection contains 20 test conversations, with, on average, 8.65 utterances, and 173 judged utterances. Akin [7], in our experiments, we use nDCG with cut off at 3.

Figure 2 shows the probability density function obtained with the Kendall's τ correlations measured for the different systems on all the possible pairs of utterances. Specifically, we considered the nDCG@3 scores and three partitions of utterance pairs: utterances belonging to different conversations, utterances from the same conversation, and utterances from the same conversation tree as defined in Subsection 2.1. Utterances from the same conversation have a statistically higher correlation (20.2% higher mean correlation with p-value of $< 10^{-10}$) than those from different conversations. Similarly, we observe that the correlation between utterances belonging to the same tree tends to be slightly shifted toward the right compared to utterances from the same conversation (4.16% higher mean correlation). Note that, in this case, the correlation is not significantly higher. When computing the mean of the scores over utterances from the same conversation, we are including in the computation the same magnitude multiple times. This observation provides a strong evidence that a simple mean does not model correctly the system performance. Figure 3a shows the aggregated scores for the 65 runs submitted to CAsT 2019 using HDA-$b_{nDCG@3}$ and nDCG@3 averaged over utterances and conversations. We highlight in blue "Automatic" runs for which the utterance rewriting was performed by a system. "Manual" runs (in orange) are instead those runs for which anaphora resolution and query rewriting was performed manually. It is possible to observe that scores aggregated with HDA-$b_{nDCG@3}$ tend to be always higher compared to the original ones. It is worth noting the multiple changes in the rank of the systems induced by HDA-$b_{nDCG@3}$ w.r.t. mean nDCG@3. In particular, these position swaps tend to be much more frequent on automatic runs than on manual ones. More in details, we observe that HDA-$b_{nDCG@3}$ swaps 5.2% of the pairs of systems in total, with 13% pairs of systems swapped among the automatic runs and 4.7% of the systems switched among the

[1]To ease reproducibility, the annotated collection and code are available at https://github.com/guglielmof/HDA



(a) HDA- $b_{nDCG@3}$ **scores** **(b)** HDA- $f_{nDCG@3}$ **scores**

Figure 3: Aggregated scores for each run submitted to CAsT 2019, observed using the proposed HDA framework against the classical mean aggregation. Runs are sorted according to the original rank induced by mean aggregation of the nDCG@3.

manual runs. This indicates that manual runs are more stable: the results on anaphoric utterances are similar to those observed for self-explanatory ones. Such behaviour is not a surprise since, for manual runs, the difference in quality between self-explanatory and anaphoric utterances is only due to the retrieval engine. For automatic runs, we have instead in the rewriting engine an additional source of possible errors. The measure presents a Kendall's τ correlation with the original nDCG@3 of 0.8962 (p-value $< 10^{-10}$).

Figure 3b shows the aggregated scores using HDA- $f_{nDCG@3}$ against the one observed by using the traditional nDCG@3 mean aggregation. In this case, we observe a correlation of 0.9019 (p-value $< 10^{-10}$). The higher correlation is likely to be due to the higher number of utterances included in the aggregation (eq. 4). Similarly to the previous case, 4.9% of the pairs of systems are swapped by HDA- $f_{nDCG@3}$. Among the automatic runs, we observe 11.6% pairs swapped, while only 5.5% of the manual pairs have been swapped. To compare the systems, we apply *ANalysis Of VAriance (ANOVA)* [4, 10, 21] over the HDA scores. In particular, we define the following *two-ways* ANOVA model:

$$y(C, s) = \mu_{..} + \gamma_C + \alpha_s + \varepsilon,$$

where $y(C, s)$ is the performance score for the conversation C using the conversational system s. y can be computed either as the mean over nDCG@3 values or by using one of the proposed HDA scores. $\mu_{..}$ is the grand mean among overall observations, γ_C is the effect of the conversation C, α_s is the effect of the system s and ε is the approximation error. Then, we apply Tukey's posthoc analysis [24] to carry out a pairwise comparison of the systems. Using the above-mentioned approach, we observe 966 statistically significantly different pairs of systems for the mean aggregation of nDCG@3, while we observe 844 (-13.6%) and 855 (-11.5%) pairs of statistically significantly different runs for HDA- $b_{nDCG@3}$ and HDA- $f_{nDCG@3}$, respectively. Note that, as shown in Figure 2, the pairs for the mean aggregation are overestimated: by averaging over correlated magnitudes, we are artificially inflating the means of the nDCG over different utterances. Therefore we are wrongfully boosting some systems and penalizing others. In this sense, both HDA- b and HDA- f appear to be more reliable in discriminating the systems. Table 2 reports the comparison between aggregation approaches. *Active Agreements (AA)* are those pairs of systems for which both aggregations agree that system A is significantly better than B. *Passive Agreements (PA)* are those pairs of systems for which neither the first nor the second consider the difference between systems A and B significant. *Passive Disagreements (PD)* are the cases where only either the first or the second aggregation states that A is significantly better than B. Compared to HDA- b, HDA- f tend to have a higher agreement with the mean aggregation.We observe a high AA, indicating that overall, we achieve similar conclusions with all aggregations. As expected, looking at the PD1 of the first two rows we observe that the traditional nDCG@3 aggregation is too loose and considers different an higher number of pairs of systems than both HDA- b and HDA- f.

Table 2: Agreement on significantly different systems pairs found by the proposed aggregations. AA: Active Agreement; PA: Passive Agreement, PD: Passive Disagreement. 2080 pairwise comparisons in total.

comparisons	AA	PA	PD1	PD2	total
nDCG@3 vs HDA- b	806	1076	160	38	2080
nDCG@3 vs HDA- f	817	1076	149	38	2080
HDA- b vs HDA- f	819	1200	25	36	2080

4 CONCLUSIONS AND FUTURE WORK

We presented a novel framework for the evaluation of conversational systems. Such framework relies on the definition of a *conversational graph*, that models a conversation as a DAG where edges indicate a semantic relation between utterances. With our experiments, we highlighted that performance scores measured for utterances in the same conversation are correlated and thus they cannot be safely aggregated using the mean. Therefore, we extended the proposed framework with two possible aggregation approaches of the scores for a conversation, HDA- b and HDA- f, that overcame some of the limitations exhibited by the traditional mean aggregation. We plan to develop machine learning algorithms capable of learning the conversational graph automatically. Furthermore, we intend to deeply explore the proposed framework, making it more general, by adding a Markovian exploration of the conversational graph.

Acknowledgments: This research was partially supported by the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence), by the University of Pisa in the framework of the AUTENS project (Sustainable Energy Autarky), and by the DAta BenchmarK for Keyword-based Access and Retrieval (DAKKAR) Starting Grants project sponsored by University of Padua and Fondazione Cassa di Risparmio di Padova e di Rovigo.

REFERENCES

[1] Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational Search (Dagstuhl Seminar 19461). In *Dagstuhl Reports*, Vol. 9. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[2] Ron Artstein, Sudeep Gandhe, Jillian Gerten, Anton Leuski, and David Traum. 2009. *Semi-formal Evaluation of Conversational Characters*. Springer Berlin Heidelberg, Berlin, Heidelberg, 22–35.

[3] S. Bangalore, G. Di Fabbrizio, and A. Stent. 2008. Learning the Structure of Task-Driven Human–Human Dialogs. *IEEE Transactions on Audio, Speech, and Language Processing* 16, 7 (2008), 1249–1259. https://doi.org/10.1109/TASL.2008.2001102

[4] D. Banks, P. Over, and N.-F. Zhang. 1999. Blind men and elephants: Six approaches to TREC data. *Inf. Retr.* 1, 1 (1999), 7–34.

[5] Ben Carterette. 2011. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 903–912. https://doi.org/10.1145/2009916.2010037

[6] 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:cs.CL/1808.07036

[7] Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. *CAsT-19: A Dataset for Conversational Information Seeking*. Association for Computing Machinery, New York, NY, USA, 1985–1988. https://doi.org/10.1145/3397271.3401206

[8] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review* (2020), 1–56.

[9] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, and Sylvain Daronnat. 2018. Investigating how conversational search agents affect user's behaviour, performance and search experience. In *The second international workshop on conversational approaches to information retrieval*.

[10] N. Ferro and G. Silvello. 2018. Toward an anatomy of IR system component performances. *jasist* 69, 2 (2018), 187–200.

[11] Asbjørn Følstad and Petter Bae Brandtzaeg. 2020. Users' experiences with chatbots: findings from a questionnaire study. *Quality and User Experience* 5 (2020), 1–14.

[12] Norbert Fuhr. 2018. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum* 51, 3 (Feb. 2018), 32–41. https://doi.org/10.1145/3190580.3190586

[13] Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2020. Utterance-to-Utterance Interactive Matching Network for Multi-Turn Response Selection in Retrieval-Based Chatbots. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 28 (Jan. 2020), 369–379. https://doi.org/10.1109/TASLP.2019.2955290

[14] 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 *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312979

[15] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and trends in Information Retrieval* 3, 1—2 (2009), 1–224.

[16] Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2017. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. arXiv:cs.CL/1603.08023

[17] Alistair Moffat, Falk Scholer, and Paul Thomas. 2012. Models and Metrics: IR Evaluation as a User Process. In *Proceedings of the Seventeenth Australasian Document Computing Symposium (ADCS '12)*. Association for Computing Machinery, New York, NY, USA, 47–54. https://doi.org/10.1145/2407085.2407092

[18] Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. Introducing MANtIS: a novel multi-domain information seeking dialogues dataset. *arXiv preprint arXiv:1912.04639* (2019).

[19] Gustavo Penha and C. Hauff. 2020. Challenges in the Evaluation of Conversational Search Systems. In *Converse@KDD*.

[20] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7 (March 2019), 249–266. https://doi.org/10.1162/tacl_a_00266

[21] Andrew Rutherford. 2001. *Introducing ANOVA and ANCOVA: a GLM approach*. Sage.

[22] T. Sakai. 2020. On Fuhr's Guideline for IR Evaluation. *SIGIR Forum* 54, 1 (June 2020), p14:1–p14:8.

[23] 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 *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 267–275.

[24] John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics* (1949), 99–114.

[25] 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 *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 496–505.

[26] Rui Yan. 2018. " Chitty-Chitty-Chat Bot": Deep Learning for Conversational AI.. In *IJCAI*, Vol. 18. 5520–5526.

[27] Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. 2016. Strategy and policy learning for non-task-oriented conversational systems. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*. 404–412.