The Eighth Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI'24)

Alexander Frummet
Universität Regensburg
Regensburg, Germany
alexander.frummet@ur.de

Andrea Papenmeier
University of Twente
Twente, Netherlands
a.papenmeier@utwente.nl

Maik Fröbe
Friedrich-Schiller-Universität Jena
Jena, Germany
maik.froebe@uni-jena.de

Johannes Kiesel
Bauhaus-Universität Weimar
Weimar, Germany
johannes.kiesel@uni-weimar.de

ABSTRACT

With the emergence of voice assistants and large language models, conversational interaction with information has become part of everyday life. The eighth edition of the search-oriented conversational AI (SCAI) workshop brings together practitioners and researchers from various disciplines to discuss challenges and advances in conversational search systems. This year's edition focuses on evaluations beyond relevance and accuracy and looks at conversational search from the user's perspective. The workshop features a shared task on user-centered evaluation datasets and metrics, challenging participants to develop new and innovative ways to evaluate conversational search systems while accounting for the needs and preferences of users.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; **Presentation of retrieval results**; **Evaluation of retrieval results**.

KEYWORDS

conversational search, conversational information access, information-seeking dialogue, evaluation

ACM Reference Format:
Alexander Frummet, Andrea Papenmeier, Maik Fröbe, and Johannes Kiesel. 2024. The Eighth Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI'24). In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24), March 10–14, 2024, Sheffield, United Kingdom.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3627508.3638310

1 INTRODUCTION

With the emergence of voice assistants and large language models, conversational interaction with information has become part of everyday life. SCAI[1] (Search-oriented Conversational Artificial Intelligence), a full-day workshop, brings together and further grows a community of researchers and practitioners interested in conversational systems for information access. In this eighth edition, we shift the theme towards the user and focus on *"Evaluation and Metrics"*. It is important to develop SCAI systems that are both reliable and accurate and account for the full range of factors of a successful and enjoyable conversation between human and agent. For example, the third Strategic Workshop in Information Retrieval in Lorne (SWIRL) [6] proposed that metrics should go beyond traditional measures and evaluate user-centered aspects such as user engagement, fluency, and usefulness [1].

The workshop will feature informative and interactive elements: In the morning, expert talks and discussion panels provide an entry point for novice participants and allow participants to share and discuss current research directions. In the afternoon, participants of the SCAI Eval 2024 shared task will present their submissions. SCAI Eval 2024[2] allows participants to engage in the topic ahead of the workshop and targets both technical and conceptual aspects of evaluation. The shared task presentations serve as a basis for in-depth group discussions. The workshop outcomes are afterward shared with the community in *SIGIR forum*.

We invite experts and prepare for discussions that approach diverse topics related to conversational search with a human perspective:

- conversational search frameworks
- conversational analyses (e.g., speech acts, turn-taking)
- linguistic personalization in conversational search
- initiative-taking in conversations
- evaluation of conversational aspects beyond accuracy (e.g., fluency, politeness, entertainment)
- user simulation for conversational search
- fairness, transparency, and trust in conversational search
- conversational agents in society

The workshop is open to all participants, academics and industry researchers, and we encourage novices to join. An introductory talk will provide a comprehensive overview of the field, and panel discussions will offer insights from leading experts.

CHIIR '24, March 10–14, 2024, Sheffield, United Kingdom

ACM ISBN 979-8-4007-0434-5/24/03.
https://doi.org/10.1145/3627508.3638310

[1] https://scai.info
[2] https://scai.info/scai-eval-2024/

1.1 Academic Background

SCAI is one of the first and long-standing workshops dedicated to conversational search, with previous editions at ICTIR (2017), EMNLP (2018), WebConf (2019), IJCAI (2019), EMNLP (2020), independently organized (2021) and at SIGIR (2022). Previous editions of this workshop have already demonstrated the breadth and multidisciplinarity inherent in the design and development of conversational search agents. The workshop is relevant to CHIIR since conversational search is a well-established topic at the frontier of information retrieval research and poses new challenges as an interaction paradigm for the field of human-computer interaction. Evaluation in conversational systems research traditionally relies on established metrics, which gauge answer relevance and accuracy relative to a ground truth. These metrics encompass word overlap-based measures like Exact Match, F1, BLEU, and ROUGE, as well as embedding-based metrics such as Semantic Answer Similarity (SAS) [11] and BERTScore [12], which excel at determining the semantic similarity of outputs [11, 12]. Various studies evaluated the usefulness and reliability of both word overlap-based [2, 8, 10] and embedding-based metrics [5, 8, 9]. However, these metrics fail to produce a complete picture for evaluating conversations as they do not measure the quality of an answer from a human perspective. For example, aspects that the CHIIR community focuses on, including fairness, transparency, fluency, ease of understanding, and others, remain unaddressed in conventional evaluation methods. Building on previous CHIIR workshops dedicated to evaluation [3, 4], the SCAI workshop's goal is to expand the scope of evaluation for conversational search systems.

1.2 Expected Key Outcomes

With this workshop, we aim to:

- **connect researchers** working on diverse aspects of conversational search by creating a space for networking and discussion,
- foster **discussions of current challenges** in conversational systems such as testing and evaluating,
- **collect and discuss datasets and metrics** for user-centered evaluation of conversational systems beyond relevance and accuracy in a shared task,
- identify promising **directions for future work** in the field of conversational search, and
- **share the key findings** with the community in a *SIGIR forum* article after the workshop.

1.3 Workshop Structure

The workshop will start with an introductory talk by an expert in the field, followed by two interactive panel discussions with SCAI experts. With the introductory talk, we ensure that novice participants are able to follow the workshop actively. Two panel discussions with invited SCAI experts will focus on the challenges and future directions of testing and evaluating conversational search systems in practice. The audience will have the opportunity to ask questions live, with question cards, or online. One member of the organizing team will take notes during the panel discussions to record the key outcomes in a *SIGIR forum* paper.

The workshop's afternoon session will be dedicated to the shared task and breakout groups. Participants in the shared task will be invited to present their submissions in lightning talks. The three submission categories of the shared task are datasets, aspect definitions, and metrics for search-oriented conversational AI evaluation. The barrier to submitting to the shared task is low – e.g., already published resources are also welcome. The lightning talks introduce everyone to each other and set the stage for the following discussions. As for the QPP++ Workshop at ECIR 2023 [7], we will have participants vote on topics from the workshop and then assign participants in two 30-minute sessions to breakout groups on the most-voted topics. The discussion results are then summarized by the participants in a plenary roundtable discussion. The voting gives participants the opportunity to choose the topics that are most interesting and relevant to them. The afternoon sessions thus aim to be a valuable opportunity for participants to learn from each other, share their insights and ideas, and contribute to the advancement of conversational search system evaluation while making meaningful connections with other experts in the field.

To make the workshop accessible to a wider audience, we plan to host the morning session and the shared task talks as a hybrid event. Those attending online should be able to watch the talks and panel discussions live and ask questions using a chat feature. Previous editions of SCAI have successfully used an online (SCAI'2021) or hybrid (SCAI'2022) setup, and it has been shown that a hybrid format makes the workshop more inclusive and engaging.

1.4 Organizers

Alexander Frummet (Universität Regensburg): Alexander is a final-year PhD student at the University of Regensburg. His research focuses on domain-specific conversational search with publications at CHIIR and TOIS.

Johannes Kiesel (Bauhaus-Universität Weimar): Johannes is a senior researcher at the Webis group. He is now co-organizing his eighth shared task (six as the main organizer). The two shared tasks for which he was the main organizer at SemEval attracted 42 and 39 teams, respectively. He was co-organizer of the Touché workshop at CLEF since 2022 and is its main organizer for 2024. He is paper chair for the CUI conference in 2024.

Maik Fröbe (Friedrich-Schiller-Universität Jena): Maik is a PhD student at the Webis group with research interests in information retrieval. He has co-organized the Touché shared task since 2020 and the SCAI shared task since 2021. He is an active developer of TIRA, which improved the reproducibility of a number of shared tasks and has an archive of more than 500 research prototypes.

Andrea Papenmeier (University of Twente): Andrea is an Assistant Professor at the University of Twente. Her research focuses on user-centered search, including conversational search, and user-centered AI systems with publications at CHI, DIS, CHIIR and in ToCHI.

REFERENCES

[1] Nicholas J. Belkin. 2016. People, Interacting with Information1. *SIGIR Forum* 49, 2 (jan 2016), 13–27. https://doi.org/10.1145/2888422.2888424

[2] Kathrin Blagec, Georg Dorffner, Milad Moradi, and Matthias Samwald. 2020. A critical analysis of metrics used for measuring progress in artificial intelligence. *CoRR* abs/2008.02577 (2020). arXiv:2008.02577 https://arxiv.org/abs/2008.02577

[3] George Buchanan, Dana McKay, Charles L A Clarke, Leif Azzopardi, and Johanne Trippas. 2020. Made to Measure: A Workshop on Human-Centred Metrics for Information Seeking. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (Vancouver BC, Canada) *(CHIIR '20)*. Association for Computing Machinery, New York, NY, USA, 484–487. https://doi.org/10.1145/3343413.3378020

[4] George Robert Buchanan, Dana McKay, and Charles Clarke. 2023. Made to Measure: A Workshop on Human-Centred Metrics for Information Seeking. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (Austin, TX, USA) *(CHIIR '23)*. Association for Computing Machinery, New York, NY, USA, 461–462. https://doi.org/10.1145/3576840.3578301

[5] Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating Question Answering Evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, Hong Kong, China, 119–124. https://doi.org/10.18653/v1/D19-5817

[6] J Shane Culpepper, Fernando Diaz, and Mark D Smucker. 2018. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). In *ACM SIGIR Forum*, Vol. 52. ACM New York, NY, USA, 34–90.

[7] Guglielmo Faggioli, Nicola Ferro, Josiane Mothe, Fiana Raiber, and Maik Fröbe. 2023. Report on the 1st Workshop on Query Performance Prediction and Its Evaluation in New Tasks (QPP++ 2023) at ECIR 2023. (2023).

[8] Zeyang Liu, Ke Zhou, and Max L. Wilson. 2021. Meta-Evaluation of Conversational Search Evaluation Metrics. *ACM Trans. Inf. Syst.* 39, 4, Article 52 (sep 2021), 42 pages. https://doi.org/10.1145/3445029

[9] Farida Mustafazade, Peter Ebbinghaus, and Seth Darren. 2022. Evaluation of Semantic Answer Similarity Metrics. *International Journal on Natural Language Computing* 11 (08 2022), 15. https://doi.org/10.5121/ijnlc.2022.11305

[10] Preksha Nema and Mitesh M. Khapra. 2018. Towards a Better Metric for Evaluating Question Generation Systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3950–3959. https://doi.org/10.18653/v1/D18-1429

[11] Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic Answer Similarity for Evaluating Question Answering Models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 149–157. https://doi.org/10.18653/v1/2021.mrqa-1.15

[12] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SkeHuCVFDr