Report on the 6th Workshop on Search-Oriented Conversational AI (SCAI 2021)

Svitlana Vakulenko
University of Amsterdam
s.vakulenko@uva.nl

Ondřej Dušek
Charles University
odusek@ufal.mff.cuni.cz

Leigh Clark
Swansea University
l.m.h.clark@swansea.ac.uk

Abstract

The 6th edition of the Search-Oriented Conversational AI workshop (SCAI 2021) was organised as a discussion platform on conversational AI for intelligent information access. The workshop was designed to be multidisciplinary, bringing together researchers and practitioners across the fields of natural language processing (NLP), information retrieval (IR), machine learning (ML) and human-computer interaction (HCI). The workshop included four sessions featuring invited talks, a separate poster session, and a session discussing the results of a shared task on conversational question answering (SCAI-QReCC).

Date: 8 October, 2021.

Website: https://scai.info.

1 Introduction

SCAI is an already established venue with a long-standing tradition of building bridges and integrating expertise from diverse research communities. It was previously organised in the context of ICTIR 2017, EMNLP 2018, TheWebConf 2019, IJCAI 2019 and EMNLP 2020.[1]

For the 6th iteration of SCAI, we focused on discussion between researchers from the different disciplines. It was therefore organised as a strictly non-archival venue, as an opportunity to present papers accepted to other venues in an interdisciplinary meeting focused on search-based conversational AI. The workshop was organised as a standalone online event, with free participation sponsored by Bloomberg Engineering, Google and Hugging Face.

The goal of SCAI is to design systems that allow for more convenient information access by means of a conversational user interface (CUI). A conversational search system should support an open-domain, information-seeking dialogue in natural language that allows people to retrieve and discover novel information relevant to their needs.

[1] To access the materials of the previous five editions of SCAI, see https://scai-workshop.github.io.

Further development of conversational search systems requires closer integration and better information exchange between multiple diverse communities. These include the fields of Dialogue Systems, Information Retrieval and Conversational User Interfaces. SCAI aims to bring together researchers interested in informing the design of a new generation of systems for conversational information access.

SCAI focused on the following topics in the design of conversational search systems: theoretical understanding and empirical analysis of information-seeking dialogues, properties of a mixed-initiative interaction, and modeling conversational contexts. Evaluation was a specific focus of the workshop, including question answering and summarisation metrics, Wizard-of-Oz experiments, user simulation for dialogues, measuring learning outcomes of an information-seeking dialogue, faithfulness and provenance of the dialogue responses. We also solicited papers on applications of conversational search: information-seeking dialogues for personalised education, healthcare, entertainment and knowledge-intensive work.

Participants were invited to apply to present their work at SCAI'21 (or nominate someone else's work) by filling out an on-line form.[2] The majority of the speakers were attracted by reaching out directly to the authors of recent papers published at major venues in relevant fields.

2 Paper Sessions

The workshop included four main presentation sessions with invited talks. After every talk the session chair asked 1-2 follow-up questions, mainly for clarifications. Following the last talk, the session chair moderated a short panel discussion between all the speakers of the session, aimed at discussing relations between the talks, identifying main challenges, opportunities for synergies and directions for future work.

The poster session featuring 14 posters was scheduled between the main sessions around the lunch break. We utilised the StreamYard platform for the main sessions, which offered a convenient setup for a moderated discussion. All sessions, except for the poster session, were live-streamed on YouTube, with an opportunity for the attendees to ask questions via the YouTube live chat functionality.

The Gather.town platform was used for the poster session to enable direct interaction between the workshop participants and to encourage spontaneous group discussions. The last session was dedicated to summarising the results of the shared task on conversational QA (SCAI-QReCC), which was co-located with the workshop.

2.1 Conversational Search Evaluation

The first presentation session consisted of four 10-minute talks, which were carefully selected and revolve around the topic of conversational search system evaluation. Evangelos Kanoulas chaired the session and held the panel discussion with all of the session speakers in the end. The recording is available online.[3]

[2] https://forms.gle/HumH6Z8LuEeGp2UaA

[3] https://www.youtube.com/watch?v=YYQ-fUH5V0c

The invited talks in this session included:

- Zeyang Liu (University of Nottingham): *Meta-evaluation of Conversational Search Evaluation Metrics* (TOIS 21) [Liu et al., 2021b]
- Aldo Lipani (University College London): *How Am I Doing?: Evaluating Conversational Search Systems Offline* (TOIS 21) [Lipani et al., 2021]
- Alexandre Salle (Federal University of Rio Grande do Sul): *Studying the Effectiveness of Conversational Search Refinement through User Simulation* (ECIR 21) [Salle et al., 2021]
- Hsien-Chin Lin (Heinrich-Heine-Universität Düsseldorf) *Domain-independent User Simulation with Transformers for Task-oriented Dialogue Systems* (SigDial 21) [Lin et al., 2021]

The first talk by Zeyang Liu introduced the results of a meta-evaluation study for existing CS metrics. The metrics were evaluated with respect to their predictive power and other aspects. The main argument is that the popular metrics borrowed from machine translation and summarisation have inherent limitations: they do not easily extend to multi-turn dialogues and are not able to account for the vast number of valid responses, which is standard in dialogue situations.

Aldo Lipani's second talk proposed a user simulation approach that models sub-topic transitions and is suitable for off-line system evaluation. They proposed an evaluation metric called Expected Conversation Satisfaction, which iteratively measures answer relevance for the questions sampled from the given subtopics, discounted by dialogue length.

The third talk was delivered by Alexandre Salle in which the author presented their paper on simulating for evaluation of refinement performance, i.e. a conversational search system producing clarification questions. Their setup allows to adjust the levels of cooperativeness and patience of users to quantify the impact on the system performance.

The last talk of this session by Hsien-Chin Lin covered a Transformer-based user simulation approach for task-oriented dialogue systems. The proposed approach aims at domain independence by using the representation of dialogue slots, which is abstracted away from any domain-specific attributes.

In the following panel discussion, limitations of the existing user simulations were discussed: the need for more complex models of user intents, the difference between written and spoken language, and the need for modeling recovery from mistakes.

2.2 User Satisfaction & Dialogue Breakdown

This session comprised of three invited talks that revolved around the topics of dialogue breakdown detection and modeling user satisfaction with a dialogue. The session was chaired by Stefan Ultes, who also lead a short panel discussion following the presentations. The session recording is available here.[4]

The invited talks in this session included:

- Shuo Zhang (Bloomberg): *Simulating User Satisfaction for the Evaluation of Task-oriented Dialogue Systems* (SIGIR 21) [Sun et al., 2021]

[4] https://www.youtube.com/watch?v=gfbhexaOXZk

- Leon-Paul Schaub (Universite Paris-Saclay): *Defining And Detecting Inconsistent System Behavior in Task-oriented Dialogues* (TALN 21) [Schaub et al., 2021]
- Katsuhide Fujita (Tokyo University of Agriculture and Technology): *Dialogue Act-based Breakdown Detection in Negotiation Dialogues* (EACL 21) [Yamaguchi et al., 2021]

Shuo Zhang started the session by describing their user satisfaction annotation efforts for task-oriented and conversational recommendation dialogues. They present a new open-source dataset of User Satisfaction Simulation (USS),[5] where the turns are annotated on a 5-graded scale, alongside the explanations of these scores, and report on their experimental results predicting user satisfaction.

In the second talk of this session, Léon-Paul Schaub explained their annotation and classification approach for detecting dialogue inconsistencies automatically. They annotated a human-machine dialogue dataset (DSTC2) with eight types of inconsistencies they identified, such as bad API calls and repetitions. The GPT2-based approach to predicting dialogue inconsistencies showed the best accuracy results on this dataset.

Finally, Katsuhide Fujita introduced a new crowd-sourced dataset of negotiation dialogues in English simulating a job interview scenario.[6] They also explained an approach for dialogue breakdown detection, which is more effective on this dataset than the state-of-the-art approaches previously proposed for task-oriented dialogues. Their approach utilises dialogue act labels and a dialogue flow model with transitions between those labels.

The subsequent panel discussion first tried to locate the main connection points between the three proposed approaches, then discussed the applicability of the methods to different dialogue types (task-specific, open-domain, negotiation dialogues etc.) and the related annotation difficulties. Finally panel members discussed covered the discrepancy between the system-side and the user-side viewpoints for measuring dialogue satisfaction. All speakers agreed that dialogue breakdowns and inconsistencies are likely to correlate with user satisfaction and perceived dialogue success, and that transferring their approaches for task-oriented dialogues into more open-domain settings is clearly a hard-to-achieve goal. Shuo Zhang also emphasised that it is important to model variance in user perceptions and expectations, which conveniently led us to the next session discussing the main challenges in dialogue personalisation.

2.3 Dialogue Personalisation

This session consisted of four 10-minute talks on the topics of extracting user preferences from dialogue, user modeling and personalisation. The session was chaired by Verena Rieser. A recording is available here.[7] The invited talks in this session included:

- Sergey Volokhin (Emory University): *You Sound Like Someone Who Watches Drama Movies: Towards Predicting Movie Preferences from Conversational Interactions* (NAACL 21) [Volokhin et al., 2021]
- Andrew Yates (University of Amsterdam): *You Get What You Chat: Using Conversations to personalise Search-Based Recommendations* (ECIR 21) [Torbati et al., 2021]

[5] https://github.com/sunnweiwei/user-satisfaction-simulation
[6] https://github.com/gucci-j/negotiation-breakdown-detection
[7] https://www.youtube.com/watch?v=UABdKsvEmNc

- Marco Polignano (University of Bari): *MyrrorBot: A Digital Assistant Based on Holistic User Models for personalised Access to Online Services* (TOIS 21) [Musto et al., 2021]
- Lucie Flek (University of Marburg): *Towards User-Centric Text-to-Text Generation: A Survey* (TSD 21) [Yang and Flek, 2021]

Sergey Volokhin kicked off the session and presented their approach to modeling user preferences from dialogue and evaluating it on the task of predicting user rating for the next movie mentioned in the dialogue. The authors extended the original CCPE dataset with sentiment labels and manually annotated user ratings based on dialogue content.[8] They also linked the movie mentions to the Rotten Tomato movie IDs to collect the ratings made by the professional critics and general audience as an external data for a collaborative filtering algorithm.

Andrew Yates presented their approach to personalising search queries by re-ranking the initial retrieval results using information from chats and questionnaires. The authors conducted a study with 14 participants to produce the dataset for their experiments, in which the users were employed to interact in pairs given a specific topic, fill out the questionnaires and assess retrieval results.[9] This information was subsequently used to create user-specific language models for query expansion.

Marco Polignano followed up by describing their approach to user modeling for a personalised chatbot, MyrrorBot, which is a digital assistant designed to improve access to a range of web services for tasks across various domains, such as music, food, news and fitness. It also provides an interface that allows users to manage the dimensions of personalisation that the chatbot has access to, such as age range, etc. This information can be collected across different social media profiles that the user already has on the Web.

Finally, Lucie Flek provided a broader overview of the existing personalised dialogue generation approaches and their evaluation, describe challenges and opportunities for future work. In their survey and positional work on user-centric generation, the authors described important differences in the existing viewpoints on the personalisation task itself, i.e., what should be personalised. There are important challenges not only inherent in the difficulty of implementation and evaluation of personalisation algorithms on the large scale, but also with respect to the ethical considerations, such as privacy and transparency. The authors also highlighted that more research is needed to assess practical implications and impact of personalisation on the user experience and decision making.

The panel discussion touched upon similarities and differences in the personalisation approaches presented in this session and the possibilities to combine them. All participants agreed that evaluation of personalisation is hard. It should also include a variety of different aspects beyond accuracy, tracking user attention and account for the learning curve in using the system. More standards are still needed to provide high-quality evaluation guidelines and means for an adequate comparison between systems. Finally, ethical concerns start from the input features that are used for different personalisation algorithms – they may include merely named entities mentions, but also protected attributes such as user location and other demographic information. To ensure fairness, the personalisation systems should provide means to audit the user models,

[8] https://github.com/sergey-volokhin/conversational-movies

[9] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/personalised-search-and-recommendation

making the users more aware of their own internal representations within the system, and include elements of serendipity to allow for discovery and user profile extension maintained by the system.

2.4 Question Answering Evaluation

This session included four 10-minute talks on the topic of evaluating question answering system performance. The session and the following panel discussion was chaired by Shayne Longpre. A recording is available here.[10]

The invited talks in this session included:

- Abhilasha Ravichander (Carnegie Mellon University): *NoiseQA: Challenge Set Evaluation for User-Centric Question Answering* (EACL 21) [Ravichander et al., 2021]
- Linqing Liu (University College London): *Challenges in Generalization in Open Domain Question Answering* [Liu et al., 2021a]
- Jifan Chen (University of Texas at Austin): *Can NLI Models Verify QA Systems' Predictions?* (EMNLP 21) [Chen et al., 2021]
- Vaibhav Adlakha (McGill University): *TopiOCQA: Open-domain Conversational Question Answering with Topic Switching* [Adlakha et al., 2021]

Abhilasha Ravichander introduced the NoiseQA dataset that extends QA evaluation to samples with errors originating from automated speech recognition, machine translation and keyboard typos.[11] This approach can be also used to augment existing datasets with synthetic examples generated by perturbing the original questions. The results demonstrate the gap in model performance (BERT and RoBERTa) that was not apparent when evaluated on the standard QA benchmarks. The authors also suggest two mitigating strategies that help to correct such errors.

Linqing Liu presented the results of their study on the model robustness for factoid open-domain QA, which showed the overlap between the patterns found in the test and train sets of the QA benchmarks that essentially evaluate model's memorisation rather than generalisation abilities. They also study in more detail the effect of novel entities on the QA performance and decompose questions into atoms using semantic role labeling to find similar patterns. The results show anticorrelation with entity frequency and cascading errors from the retrieval stage.

Jifan Chen explained their approach designed to verify answer predictions using natural language inference, i.e., to find truthful answers that can entail the hypothesis statements derived from the original questions. The authors employ T5-based models to convert questions into hypothesis statement and decontextualise sentences by automated rewriting. Their results show effectiveness of the proposed pipeline, which allows to mitigate errors and also improve confidence in the predicted answers.

Vaibhav Adlakha introduced a new dataset for open-domain conversational QA, TopiOCQA.[12] The main novel property of TopiOCQA is transitions between several topics within the same conversation that require retrieval of multiple documents, which is typical in an information-seeking scenario. The authors developed an interface with Wikipedia hyperlinks that allows switching between related topics to collect conversational questions.

[10] https://www.youtube.com/watch?v=NaKYNjEARi8
[11] https://noiseqa.github.io
[12] https://mcgill-nlp.github.io/topiocqa/

The panel discussion reviewed possible directions for future work, including the need to further improve model robustness, but also to move the evaluation focus towards orthogonal dimensions beyond accuracy. These involve the usability of QA systems in practice, collaboration with domain experts and finding ways to bridge the knowledge gap when negotiating terminology, meaning and information need, and providing supplementary resources to better inform the user. The panel also highlighted the lack of spoken QA benchmarks that could provide for a much more realistic scenario mirroring the setup of a voice-based assistant. Vaibhav Adlakha suggested to make QA benchmarking into a continuous process by iteratively collecting feedback and further expanding existing datasets to make them more diverse, by including sources beyond Wikipedia, and allowing to evaluate progress in generalisation on out-domain examples. Other challenges, such as efficiency and privacy, should be considered to make QA applications more robust and usable in practice.

3 Poster Session

The poster session included 14 presentations related to QA and Dialogue. The posters selected for this session were mostly placed within just one of our target fields, but were relevant to researchers from the other fields. A short listing of the presented paper topics follows.

Papers presented as posters on dialogue-related topics:

- Francesca Alloatti: *Conversation Analysis, Repair Sequences and Human Computer Interaction* [Alloatti et al., 2021] – using conversational analysis to evaluate dialogue systems as well as to correct their behavior on-the-fly.
- Anna Liednikova: *Gathering Information and Engaging the User ComBot: A Task-Based, Serendipitous Dialogue Model for Patient-Doctor Interactions* [Liednikova et al., 2021] – ensemble dialogue systems for clinical studies, seeking to increase patient engagement by social and follow-up questions.
- Abhishek Kaushik: *A Conceptual Framework for Implicit Evaluation of Conversational Search Interfaces* [Kaushik and Jones, 2021] – a human evaluation framework for conversational search, with multiple criteria related to user experience.
- Gustavo Penha: *On the Calibration and Uncertainty of Neural Learning to Rank Models for Conversational Search* [Penha and Hauff, 2021] – chatbot response rankers explicitly trained to produce a probability distribution and model uncertainty.
- Vahid Sadiri Javadi: *Generating Opinionated Sales Negotiations* – synthetic search/purchase dialogue generation based on mining Amazon reviews.
- Alexandros Papangelis: *Generative Conversational Networks* [Papangelis et al., 2021] – generating additional training data for dialogue tasks using pretrained language models, optimized by reinforcement learning with rewards from target task performance
- Christian Geishauser: *What Does The User Want? Information Gain for Hierarchical Dialogue Policy Optimization* [Geishauser et al., 2021] – a reward function for reinforcement learning, focusing on providing information to the user.
- Vojtech Hudecek: *Discovering Dialogue Slots with Weak Supervision* [Hudeček et al., 2021] – a novel approach for finding relevant dialogue slots using generic semantic parsers and entity recognizers.

Papers presented as posters on QA-related topics:

- Kalpesh Krishna: *Hurdles to Progress in Long-form Question Answering* [Krishna et al., 2021] – comments and suggestions regarding the shortcomings of the current state-of-the-art in long-form QA, including lack of answer grounding, train-test overlap in benchmarks, and unreliable evaluation.
- Timo Möller: *Semantic Answer Similarity for Evaluating Question Answering Models* [Risch et al., 2021] – a new automatic metric for QA based on embedding similarity, designed to capture answer paraphrases.
- Rishiraj Saha Roy: *UNIQORN: Unified Question Answering over RDF Knowledge Graphs and Natural Language Text* [Pramanik et al., 2021] – a QA system capable of operating with both textual and RDF structured inputs, building a knowledge graph from text on-the-fly.
- Magdalena Kaiser: *Reinforcement Learning from Reformulations in Conversational Question Answering over Knowledge Graphs* [Kaiser et al., 2021] – reducing annotation needs by using question reformulations as negative feedback and follow-up questions as positive feedback in a reinforcement learning setup for conversational QA.
- Zhen Jia: *Complex Temporal Question Answering on Knowledge Graphs* [Jia et al., 2021] – a system for answering temporal questions with multiple entities, predicates, or temporal conditions, using graph filtering, BERT embeddings, and graph convolutional networks.
- Farnaz Ghassemi: *Zero-Shot Clinical Questionnaire Filling From Human-Machine Interactions* [Ghassemi Toudeshki et al., 2021] – automatically filling in medical questionnaires based on conversations with chatbots, using QA, textual inference, and text classification.

4 The Shared Task on Conversational Question Answering

In this session, we presented the results of our shared task on Conversational QA[13] and discussed the main challenges and presented our QA evaluation approach. The recordings of this session are available here.[14]

Organising a shared task in the context of the workshop is a long-standing tradition at SCAI. This year we organised a new shared task on Open-Domain Conversational Question Answering. The goal of this task was to establish a reliable QA benchmark and encourage collaboration between communities working on dialogue response generation, passage retrieval and reading comprehension.

To participate in the shared task, the teams had to register using an online form[15] which granted them access to the TIRA platform, allowing them to submit results appearing on the official leaderboard[16] and interact with other participants and organisers through the TIRA forum. While the organisers explicitly encouraged code submissions and TIRA provides the required functionality (virtual machines), none of the participating teams made use of this functionality.

In conversational QA, the task is to answer a series of contextually-dependent questions as they may occur in a natural human-to-human conversation. We also made it possible to partic-

[13] https://scai.info/scai-qrecc
[14] https://www.youtube.com/watch?v=1pqq1Lsl8ag
[15] https://forms.gle/JFBXZXPtWPqbtLhu8
[16] https://www.tira.io/task/scai-qrecc

ipate with non-conversational approaches in the conversational task by using baseline question rewrites [Vakulenko et al., 2021].

The challenge uses the QReCC dataset introduced by Anantha et al. [2021] for evaluation, which contains 14K conversations with 81K question-answer pairs and 54M passages. The passage collection was constructed by processing 10M web pages from the Common Crawl and the Wayback Machine.

We used the submitted runs to pull answers submitted by the participating systems to extend the coverage of the ground truth annotations and improve the evaluation metrics. The primary evaluation metrics are F1 and EM performance on the QA task. We also report MRR and ROUGE for passage retrieval and question rewriting subtasks to provide better comparison in performance over the intermediate steps. More details about the shared task are available under this URL.[17]

All participating teams were invited to submit a short notebook paper detailing their approach. We received 30 runs, in total, submitted by 4 teams. The organisers also contributed 3 baseline runs:

- basic – rewritten question were used as answers;
- simple – the answers were extracted using a simple question word-overlap heuristic from the top passages retrieved by BM25;
- GPT3 – results obtained from running the GPT3 model on the original conversational questions via OpenAI API.

Goncalo Raposo represented team *Rachael* and described their approach using T5-based model for question rewriting fine-tuned on CANARD dataset,[18] a BM25 model from Pyserini for passage retrieval given the previously rewritten question, and a PEGASUS model for answer generation.

Clement Pasti presented the runs submitted by team *Torch*. They implemented question rewriting with a GPT2-based model, initial passage retrieval with BM25 further reranked by a BERT-based model, and an answer generation T5 model.

Shashank Gupta described the approach pioneered by team *Ultron*, which included a BART-based question rewriting model, and an answer generation end-to-end RAG-based. The best performing model used a BM25-based filtering approach.

In a panel discussion, participants discussed the main challenges they faced while working on this shared task. Several teams tried to apply dense passage retrieval but did not succeed due to the large collection size. There are a few ways that could allow to scale such approaches by exploiting redundancy in data, and partitioning the collection into smaller sub-collections or specialised clusters that can be processed independently or on-demand, e.g., by means of question classification.

Notably, all of the teams developed generative rather than extractive QA models. While generative models are more complex, especially in terms of evaluation, they provide sufficient flexibility allowing them to adapt and produce more human-like, grammatical and compact answers.

We extended the initial set of evaluation metrics with several more recently introduced approaches: POS Score, SAS, BERT and KPQA. Moreover, we made a sample of QA pairs where the answers submitted by different models were distinct but scored high by SAS at the same time against our reference answers. Those QA samples were verified by two crowd-workers in-

[17] https://scai.info/scai-qrecc

[18] https://huggingface.co/castorini/t5-base-canard

dependently to check if the provided system response answers the given question, i.e., whether the human annotators consider the provided answers plausible. Evaluating truthfulness of the submitted answers is more challenging since it requires grounding in the passage collection. Such relations are often not clear since several top passages are often used to generate an answer. Since the passages are rather long, verifying them by untrained annotators may be especially prone to errors. These considerations motivated us to restrict the manual evaluation to the answer plausibility as the first phase.

As a result of both automated and manual evaluation, *Rachael* appeared as a clear winner. 93% of *Rachael*'s answers that were sampled for our manual evaluation were marked correct. Diego Cicarelli from Bloomberg Engineering announced the winners and the prizes for all the presenting teams.

5 Discussion & Conclusion

5.1 Challenges & Opportunities of Virtual Formats

This is the second year where the SCAI workshop was forced into a virtual format. We successfully used the situation to expand participation beyond a single conference by providing free access to anyone interested in the topic of conversational search. We also changed the format of the workshop by making all talks invitation-only. This approach succeeding in attracting speakers who had recently published their work at high-impact venues.

Another innovation this year was running a panel discussion between all speakers at the end of every session. This is something usually seen for more senior researchers. This allowed for insightful discussions between more junior and senior researchers alike. Through these different perspectives, discussions combined more hands-on insights and concrete findings, alongside strategic and long-term goals for the research disciplines.

The nature of the virtual sessions resulted in fewer interactions between speakers from different sessions. This speaks to the ongoing challenges in facilitating interactive discussions between participants when not sharing a physical environment. Speakers did refer to the talking points of previous sessions in their talks, ensuring some continuation was enabled throughout the workshop. A return to physical or a hybrid format should further encourage

5.2 Engaging with Other Research Communities

When inviting speakers to this workshop, we surveyed papers published this year in high-impact venues related to Natural Language Processing (NLP), Information Retrieval (IR), Artificial Intelligence (AI) and Human-Computer Interaction (HCI). Notably, we struggled to find speakers working on the HCI-related research questions relevant to the workshop. This was mentioned several times in the panel discussions highlighting the lack of user-centered research that could help to inform design of our systems.

Another commonly highlighted challenge is the need to take into account the differences of interactions when considering speech as an input modality as opposed to text. For future iterations of SCAI, we will continue reach out and connect with researchers working on user-centered approaches and engaging in further dialogue with members of relevant speech research fields.

Acknowledgements

We would like to thank our sponsors, the steering committee, all speakers and attendees for their support. Ondřej Dušek was supported by Charles University grant PRIMUS/19/SCI/10.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. Topiocqa: Open-domain conversational question answeringwith topic switching, 2021.

Francesca Alloatti, Luigi Di Caro, and Alessio Bosca. Conversation analysis, repair sequences and human computer interaction. In *Proceedings of DEEP-DIAL*, 2021. URL https://hdl.handle.net/2318/1795726.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 520–534, 2021. URL https://doi.org/10.18653/v1/2021.naacl-main.44.

Jifan Chen, Eunsol Choi, and Greg Durrett. Can NLI models verify QA systems' predictions? In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3841–3854, 2021. URL https://aclanthology.org/2021.findings-emnlp.324.

Christian Geishauser, Songbo Hu, Hsien-chin Lin, Nurul Lubis, Michael Heck, Shutong Feng, Carel van Niekerk, and Milica Gašić. What does the user want? information gain for hierarchical dialogue policy optimisation. In *Proceedings of ASRU*, 2021.

Farnaz Ghassemi Toudeshki, Philippe Jolivet, Alexandre Durand-Salmon, and Anna Liednikova. Zero-shot clinical questionnaire filling from human-machine interactions. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 51–62, November 2021. URL https://aclanthology.org/2021.mrqa-1.5.

Vojtěch Hudeček, Ondřej Dušek, and Zhou Yu. Discovering dialogue slots with weak supervision. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2430–2442, August 2021. doi: 10.18653/v1/2021.acl-long.189. URL https://aclanthology.org/2021.acl-long.189.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. Complex temporal question answering on knowledge graphs. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 792–802, 2021. doi: 10.1145/3459637.3482416. URL https://doi.org/10.1145/3459637.3482416.

Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. Reinforcement learning from reformulations in conversational question answering over knowledge graphs. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 459–469, 2021. doi: 10.1145/3404835.3462859. URL https://doi.org/10.1145/3404835.3462859.

Abhishek Kaushik and Gareth J. F. Jones. A conceptual framework for implicit evaluation of conversational search interfaces. *CoRR*, abs/2104.03940, 2021. URL https://arxiv.org/abs/2104.03940.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, June 2021. doi: 10.18653/v1/2021.naacl-main.393. URL https://aclanthology.org/2021.naacl-main.393.

Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. Gathering information and engaging the user ComBot: A task-based, serendipitous dialog model for patient-doctor interactions. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 21–29, June 2021. doi: 10.18653/v1/2021.nlpmc-1.3. URL https://aclanthology.org/2021.nlpmc-1.3.

Hsien-chin Lin, Nurul Lubis, Songbo Hu, Carel van Niekerk, Christian Geishauser, Michael Heck, Shutong Feng, and Milica Gasic. Domain-independent user simulation with transformers for task-oriented dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 445–456, July 2021. URL https://aclanthology.org/2021.sigdial-1.47.

Aldo Lipani, Ben Carterette, and Emine Yilmaz. How am i doing?: Evaluating conversational search systems offline. *ACM Trans. Inf. Syst.*, 39(4), August 2021. ISSN 1046-8188. doi: 10.1145/3451160. URL https://doi.org/10.1145/3451160.

Linqing Liu, Patrick Lewis, Sebastian Riedel, and Pontus Stenetorp. Challenges in generalization in open domain question answering, 2021a.

Zeyang Liu, Ke Zhou, and Max L. Wilson. Meta-evaluation of conversational search evaluation metrics. *ACM Trans. Inf. Syst.*, 39(4), August 2021b. ISSN 1046-8188. doi: 10.1145/3445029. URL https://doi.org/10.1145/3445029.

Cataldo Musto, Fedelucio Narducci, Marco Polignano, Marco De Gemmis, Pasquale Lops, and Giovanni Semeraro. Myrrorbot: A digital assistant based on holistic user models for personalized access to online services. *ACM Trans. Inf. Syst.*, 39(4), August 2021. ISSN 1046-8188. doi: 10.1145/3447679. URL https://doi.org/10.1145/3447679.

Alexandros Papangelis, Karthik Gopalakrishnan, Aishwarya Padmakumar, Seokhwan Kim, Gokhan Tur, and Dilek Hakkani-Tur. Generative conversational networks. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 111–120, July 2021. URL https://aclanthology.org/2021.sigdial-1.12.

Gustavo Penha and Claudia Hauff. On the calibration and uncertainty of neural learning to rank models for conversational search. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 160–170, April 2021. doi: 10.18653/v1/2021.eacl-main.12. URL https://aclanthology.org/2021.eacl-main.12.

Soumajit Pramanik, Jesujoba Alabi, Rishiraj Saha Roy, and Gerhard Weikum. UNIQORN: unified question answering over RDF knowledge graphs and natural language text. *CoRR*, abs/2108.08614, 2021. URL https://arxiv.org/abs/2108.08614.

Abhilasha Ravichander, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard H. Hovy, and Alan W. Black. Noiseqa: Challenge set evaluation for user-centric question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2976–2992, 2021. URL https://aclanthology.org/2021.eacl-main.259/.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157, November 2021. URL https://aclanthology.org/2021.mrqa-1.15.

Alexandre Salle, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein. Studying the effectiveness of conversational search refinement through user simulation. In *Proceedings of ECIR*, pages 587–602, 2021.

Léon-Paul Schaub, Vojtech Hudecek, Daniel Stancl, Ondrej Dusek, and Patrick Paroubek. Defining and detecting inconsistent system behavior in task-oriented dialogues. In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 142–152, 6 2021. URL https://aclanthology.org/2021.jeptalnrecital-taln.13.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2499–2506, 2021. ISBN 9781450380379. doi: 10.1145/3404835.3463241. URL https://doi.org/10.1145/3404835.3463241.

Ghazaleh H. Torbati, Andrew Yates, and Gerhard Weikum. You get what you chat: Using conversations to personalize search-based recommendations. In *Advances in Information Retrieval*, pages 207–223, 2021. ISBN 978-3-030-72113-8.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. Question rewriting for conversational question answering. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, pages 355–363, 2021. URL https://doi.org/10.1145/3437963.3441748.

Sergey Volokhin, Joyce Ho, Oleg Rokhlenko, and Eugene Agichtein. You sound like someone who watches drama movies: Towards predicting movie preferences from conversational interactions.

In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3091–3096, June 2021. doi: 10.18653/v1/2021.naacl-main.246. URL https://aclanthology.org/2021.naacl-main.246.

Atsuki Yamaguchi, Kosui Iwasa, and Katsuhide Fujita. Dialogue act-based breakdown detection in negotiation dialogues. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 745–757, April 2021. doi: 10.18653/v1/2021.eacl-main.63. URL https://aclanthology.org/2021.eacl-main.63.

Diyi Yang and Lucie Flek. Towards user-centric text-to-text generation: A survey. In *Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings*, volume 12848 of *Lecture Notes in Computer Science*, pages 3–22, 2021. URL https://doi.org/10.1007/978-3-030-83527-9_1.