COMMON CONVERSATIONAL COMMUNITY PROTOTYPE: SCHOLARLY CONVERSATIONAL ASSISTANT

Krisztian Balog
University of Stavanger
Norway
krisztian.balog@uis.no

Lucie Flekova
Technische Universität Darmstadt
Germany
l.flekova@gmail.com

Matthias Hagen
Martin-Luther-Universität Halle-Wittenberg
Germany
matthias.hagen@informatik.uni-halle.de

Rosie Jones
Spotify
United States
rjones@spotify.com

Martin Potthast
Leipzig University
Germany
martin.potthast@uni-leipzig.de

Filip Radlinski
Google
UK
filiprad@google.com

Mark Sanderson
RMIT University
Australia
mark.sanderson@rmit.edu.au

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

Hamed Zamani
Microsoft
United States
hazamani@microsoft.com

ABSTRACT

This paper discusses the potential for creating academic resources (tools, data, and evaluation approaches) to support research in conversational search, by focusing on realistic information needs and conversational interactions. Specifically, we propose to develop and operate a prototype conversational search system for scholarly activities. This Scholarly Conversational Assistant would serve as a useful tool, a means to create datasets, and a platform for running evaluation challenges by groups across the community. This article results from discussions of a working group at Dagstuhl Seminar 19461 on Conversational Search.

***K*eywords** Conversational search · Conversational recommendation · Evaluation · Benchmark

1 Introduction

Conversational search is a newly emerging research area that aims to provide access to digitally stored information by means of a conversational user interface, that is, a dialogue-based interaction inspired and informed by human communication processes [5, 15, 18]. The major goal of a conversational search system is to effectively retrieve relevant answers to a wide range of questions expressed in natural language, with rich user-system dialogue as a crucial component for understanding the question and refining the answers [1]. The respective dialogue comprises of a sequence of exchanges between one or more users and a conversational search system, which can enable multi-step task completion and recommendation [6]. Several theoretical frameworks that further specify various components and requirements for an effective conversational search system have recently been proposed [2, 14, 16, 17, 19].

It is commonly recognized that only few natural conversational search corpora exist. Rather, corpora are often created through imagined needs (often in task-oriented Wizard-of-Oz studies), are inspired by logs, or come from crawls of community fora. This leads to significant research effort being planned around existing biased data and metrics, rather than data and metrics being constructed to support the most impactful research. While there have been instances of the research community interaction enabling research, such as at ECIR 2019,[1] this is relatively rare. One of our key motivations is to produce a system and corpus that contains and supports real user needs.

Simultaneously, our community has common unsatisfied needs that appear very well suited to conversational search. Some common tasks are performed by researchers repeatedly without providing any community research value in terms

[1] http://ecir2019.org/sociopatterns/

of data and feedback collection, despite being relevant to many published experiments. Examples of these tasks include PC selection or finding interest profiles in EasyChair, or identifying the most relevant sessions in the Whova conference app. The collective time spent (arguably inefficiently) by our community on such tasks may far surpass the cost of creating a system that also supports research progress while providing this *community value*.

2 Proposed Research

We propose to develop and operate a prototype conversational search system (Scholarly Conversational Assistant) that would serve as

- a useful search tool,
- a means to create datasets for further academic research,
- and a platform for running evaluation challenges by groups across the community.

In particular, the Scholarly Conversational Assistant would allow our research community to perform a range of research-related activities. In extensive discussions, we settled on this domain for a number of reasons: (1) The data that is involved (such as papers authored, conferences/talks attended, PC memberships) is generally considered less private. Indeed most such data is already public albeit difficult to search. (2) The system is one that the members of our community would be using ourselves, giving an active knowledgeable participant base, who could contribute improvements and publish papers based on interactions observed. (3) It caters to a broad range of information needs (see below) that are currently not supported well by existing systems. (4) The relevant research groups could avoid competing with commercial providers.

A number of other possible domains were discussed, including movies, music, news, and podcasts. They have a significantly larger potential audience, yet potentially compete with commercial providers. In determining our plan, it became clear that some participants also consider interests in these areas to be highly sensitive or personal. As a critical constraint, privacy of relevant data is key (having impacted, for example, the Living Labs research [10] despite significant effort).

3 Research Challenges

The aim of the Scholarly Conversational Assistant system would be to enable a wide variety of research in conversational search by covering example information needs like:

- "What should I read?" — Find research on a new area of interest.
- "Help me plan my attendance" — Plan what sessions to attend and whom to talk to at a conference. (Conference organizers could also use that information for optimizing room allocations.)
- "Whom should I invite?" — Find conference PC, SPC, session chairs, invite speakers, etc.

Importantly, the system would log all interactions such that classes of information needs that have potential for study may be identified over time. People may evaluate the system by filling out a questionnaire, with the option of free text feedback, after each conversation (and possibly leave comments behind for individual system utterances).

3.1 Connection to Knowledge Graphs

The system would operate on a *personal research graph* (PKG) [3], more specifically, the portion of the PKG that the user wants to share with the system. The PKG could include, among other information:

- Authorship information (which may be connected to a public citation graph),
- Conference committee membership, awards, etc.,
- Talks given anywhere public,
- Attendance of conferences, sessions, etc.,
- (in the private part) Annotations of papers, notes on talks, etc.

3.2 First Steps

The project is ambitious, but we think it can be grown incrementally:

- A starting point would be to get one ore more graduate students to start coding a tool and check it in to GitHub. It is likely that students will be able to build on top of existing infrastructure. In order for this to work, it will be necessary for a research team to own the decisions who (believes they will) get value out of such work. With a prototype system in place, one could establish a shared task at a workshop or conduct a lab study at scale. One might also design a challenge at TREC/CLEF to make use of the skeleton.
- One might alternatively start by collecting evidence that such a system is something the community actually wants. Here, a sample of dialogues or information needs (that one might want to support) could be gathered.

4 Broader Impact

The organization of shared tasks has a long tradition in information retrieval as well as natural language processing and the dialogue community within it. In conversational search, these two communities will collaborate to build search systems that have a natural language interface as well as conversational capabilities. The breadth of potential tasks that are due to this confluence of research fields—as also identified in Dagstuhl seminar 19461—is large. As such, developing common infrastructure and shared tasks would have high value for the community.

In particular, the outcome of shared tasks are typically large corpora and performance measures that, together, form reusable benchmarks. For example, the Cranfield-style evaluation frameworks that were adapted by TREC, or the corpora developed for the CoNLL shared tasks have had a broad impact on their respective communities at large. We expect that a conversational search challenge, too, will help to align and shape the community.

Moreover, by developing specific shared tasks in the form of living labs [9, 10], we see the opportunity to apply early conversational search systems in practice as soon as possible. Here, the application domain of scholarly search, while allowing for a wide range of basic and advanced evaluation setups, may ideally transfer directly into new prototypes to enhance research itself, for instance, impacting the productivity of managing one's personal conferences schedules.

5 Obstacles and Risks

A variety of systems for storing and accessing research publications, reviews and conference attendance already exist. For the Scholarly Conversational Assistant to be successful, it must either be more useful than these, or potentially integrate with them. Some of the existing systems include: dblp, semantic scholar, ACM library, Google scholar, ACL anthology, open review, arXiv, Athena conference chatbot, Citeseer, Arnetminer, and arXivDigest (more on these in related reading).

Risks involved in operationalizing our envisaged conversational search system include:

- *Privacy and data retention rules.* Ideally, the Scholarly Conversational Assistant would allow the logging of user interactions including voice input. For all personal data, the system would require a process for data access, retention and deletion as well as logging, in compliance with local regulations. Even the use of third-party speech recognizers may be sensitive depending on the location of data storage.
- *Opinions != facts in indexing.* Some information that could be collected is likely to be expressed opinions rather than facts (e.g., tweets about papers). Thus, we may want to allow verification of such information before use for search and recommendation, or present it in a separate clearly-marked format with the potential for correction or deletion. Others may wish to combine private information (such as a user's personal opinions about papers), without this information being propagated.
- *Speech recognition.* The use of third-party speech recognizers may be sensitive depending on the location of data storage. In addition, in the Scholarly Conversational Assistant case, the corpus contains many proper names and technical terms. A speech recognizer may require a custom language model integrating this corpus to perform well.
- *Personal Knowledge Graph implementation.* We would need a design that allows both cloud- and client-side storage of personal data. We need to make sure that private parts of the PKG remain private and also that users have full control over what is stored in their PKG. In case an offline dataset is created and shared, there needs to be an agreement in place that ensures that personal data would need to be removed upon request. (It should be noted that there is no way to enforce this, and "unauthorized" access may only be spotted if people publish using that data.)

- *Usage volume.* Low user participation is a concern. Beyond ensuring that the system is useful, other ways to mitigate this could include rewarding (paying) users or incentivizing them through gamification (e.g., at conferences to use the system).
- *Implementation.* The underlying system would require a significant effort to implement. As this would likely be contributions from different practitioners at various stages in their careers over an extended time, the contributors would naturally change. To alleviate some associated risk, a strong modularization would be beneficial, with clear interfaces and documentation. Moreover, the design of the initial prototype should be as simple as possible, with agreement of how the system's continued development is ensured during operation. The live service would also need coordination, for example, of how live experiments are planned and executed.
- *Operation.* Past academic systems have often been deployed on individual servers without redundancy, and potentially lacking resources for scalability. This project would likely wish to consider for this project to identify possible sponsorship from a cloud provider or host institution with significant cluster resources. The hosting decision should likely take into account long-term commitment.
- *Stability and reproducibility.* If used for online challenges where participants submit code that runs live, this would need to be of suitable quality to be widely used. Care would need to be taken in designing common APIs that minimize the risks involved where a component does not behave as expected.

6 Suggested Readings and Resources

In the following, we list a set of resources (data and tools) that might be useful in building such a system.

Software platforms:

- Macaw: A conversational information seeking platform implemented in Python which supports multiple interfaces and modalities [21].
- TIRA Integrated Research Architecture [13] (a modularized platform for shared tasks).

Scientific IR tools:

- ArXivDigest: A personalized scientific literature recommendation framework based on arXiv articles.[2]
- GrapAL: Querying Semantic Scholar's literature graph [4] (web-based tool for exploring scientific literature, e.g., finding experts on a given topic).[3]

Open-source scholarly conversational agents:

- UKP-ATHENA: A scientific conversational agent [12] (early prototype for assisting ACL* conference attendees and answering basic ACL Anthology queries).[4]

Data collections suitable to be incorporated in the Scholarly Conversational Assistant include, but are not limited to:

- Open Research Knowledge Graph[5] (ORKG) [11]: Semantic annotations of scientific publications
- Semantic Scholar: Articles in a broad range of fields
- ACM DL: A subset of computer science articles
- dblp: A clean list of computer science articles
- ACL Anthology: A public collection of ACL* articles
- Open Review: A small subset of conference articles with public reviews
- Other sources include: Google Scholar, Citeseer, Arnetminer, and Conference attendance apps (e.g., Whova)

Other related work:

- Gentile et al. [8]: Recupero: Conference Live: Accessible and Sociable Conference Semantic Data
- Dalton et al. [7]: Vote Goat: Conversational Movie Recommendation
- Wan et al. [20]: Aminer: Search and mining of academic social networks (researcher-centric IR)

[2] https://github.com/iai-group/arxivdigest

[3] https://allenai.github.io/grapal-website/

[4] http://athena.ukp.informatik.tu-darmstadt.de:5002/

[5] http://orkg.org

References

[1] J. Allan, B. Croft, A. Moffat, and M. Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. *SIGIR Forum*, 46(1): 2–32, May 2012. ISSN 0163-5840. doi: 10.1145/2215676.2215678. URL `http://doi.acm.org/10.1145/2215676.2215678`.

[2] L. Azzopardi, M. Dubiel, M. Halvey, and J. Dalton. Conceptualizing agent-human interactions during the conversational search process. In *The Second International Workshop on Conversational Approaches to Information Retrieval*, July 2018. URL `https://strathprints.strath.ac.uk/64619/`.

[3] K. Balog and T. Kenter. Personal knowledge graphs: A research agenda. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, pages 217–220, New York, NY, USA, 2019. ACM. URL `http://doi.acm.org/10.1145/3341981.3344241`.

[4] C. Betts, J. Power, and W. Ammar. Grapal: Querying semantic scholar's literature graph. *arXiv preprint arXiv:1902.05170*, 2019.

[5] B. R. Cowan and L. Clark, editors. *Proceedings of the 1st International Conference on Conversational User Interfaces, CUI 2019, Dublin, Ireland, August 22-23, 2019*, 2019. ACM.

[6] J. S. Culpepper, F. Diaz, and M. D. Smucker. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). *SIGIR Forum*, 52(1):34–90, Aug. 2018. ISSN 0163-5840. doi: 10.1145/3274784.3274788. URL `http://doi.acm.org/10.1145/3274784.3274788`.

[7] J. Dalton, V. Ajayi, and R. Main. Vote goat: Conversational movie recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1285–1288, New York, NY, USA, 2018. ACM. URL `http://doi.acm.org/10.1145/3209978.3210168`.

[8] A. L. Gentile, M. Acosta, L. Costabello, A. G. Nuzzolese, V. Presutti, and D. Reforgiato Recupero. Conference live: Accessible and sociable conference semantic data. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 1007–1012, New York, NY, USA, 2015. ACM. URL `http://doi.acm.org/10.1145/2740908.2742025`.

[9] F. Hopfgartner, A. Hanbury, H. Müller, I. Eggel, K. Balog, T. Brodt, G. V. Cormack, J. Lin, J. Kalpathy-Cramer, N. Kando, M. P. Kato, A. Krithara, T. Gollub, M. Potthast, E. Viegas, and S. Mercer. Evaluation-as-a-service for the computational sciences: Overview and outlook. *J. Data and Information Quality*, 10(4):15:1–15:32, Oct. 2018. URL `http://doi.acm.org/10.1145/3239570`.

[10] F. Hopfgartner, K. Balog, A. Lommatzsch, L. Kelly, B. Kille, A. Schuth, and M. Larson. Continuous evaluation of large-scale information access systems: A case for living labs. In N. Ferro and C. Peters, editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 511–543. Springer, 2019. URL `https://doi.org/10.1007/978-3-030-22948-1_21`.

[11] M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, and S. Auer. Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture*, K-CAP '19, pages 243–246, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-7008-0. doi: 10.1145/3360901.3364435. URL `http://doi.acm.org/10.1145/3360901.3364435`.

[12] M. Mesgar, P. Youssef, L. Li, D. Bierwirth, Y. Li, C. M. Meyer, and I. Gurevych. When is acl's deadline? a scientific conversational agent. *arXiv preprint arXiv:1911.10392*, 2019.

[13] M. Potthast, T. Gollub, M. Wiegmann, and B. Stein. Tira integrated research architecture. In *Information Retrieval Evaluation in a Changing World*, pages 123–160. Springer, 2019.

[14] F. Radlinski and N. Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 117–126, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4677-1. doi: 10.1145/3020165.3020183. URL `http://doi.acm.org/10.1145/3020165.3020183`.

[15] J. R. Trippas. *Spoken Conversational Search: Audio-only Interactive Information Retrieval*. PhD thesis, RMIT University, 2019.

[16] J. R. Trippas, D. Spina, L. Cavedon, and M. Sanderson. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 325–328, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4677-1. doi: 10.1145/3020165.3022144. URL `http://doi.acm.org/10.1145/3020165.3022144`.

[17] J. R. Trippas, D. Spina, P. Thomas, M. Sanderson, H. Joho, and L. Cavedon. Towards a model for spoken conversational search. *Information Processing & Management*, 57(2):102162, 2020.

[18] S. Vakulenko. *Knowledge-based Conversational Search*. PhD thesis, TU Wien, 2019.

[19] S. Vakulenko, K. Revoredo, C. D. Ciccio, and M. de Rijke. QRFA: A data-driven model of information-seeking dialogues. In *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part I*, pages 541–557, 2019.

[20] H. Wan, Y. Zhang, J. Zhang, and J. Tang. Aminer: Search and mining of academic social networks. *Data Intelligence*, 1(1):58–76, 2019.

[21] H. Zamani and N. Craswell. Macaw: An extensible conversational information seeking platform. *arXiv preprint arXiv:1912.08904*, 2019.