Report from Dagstuhl Perspectives Workshop 24352

Conversational Agents: A Framework for Evaluation (CAFE)

Christine Bauer*[1], Li Chen*[2], Nicola Ferro*[3], and Norbert Fuhr*[4]

1 **Paris Lodron University Salzburg, AT.** christine.bauer@plus.ac.at
2 **Hong Kong Baptist University, HK.** lichen@comp.hkbu.edu.hk
3 **University of Padua, IT.** nicola.ferro@unipd.it
4 **Universität Duisburg-Essen, DE.** norbert.fuhr@uni-due.de

Abstract

This report documents the program and the outcomes of the Dagstuhl Perspectives Workshop 24352, "Conversational Agents: A Framework for Evaluation (CAFE)", which brought together 22 distinguished researchers and practitioners from 12 countries. In this workshop, a new framework for the evaluation of conversational information access systems was developed, consisting of six major components: 1) goals of the system's stakeholders, 2) user tasks to be studied in the evaluation, 3) aspects of the users carrying out the tasks, 4) evaluation criteria to be considered, 5) evaluation methodology to be applied, and 6) measures for the quantitative criteria chosen. An evaluation design begins with identifying the stakeholders, whose goals determine the criteria. Tasks and evaluation methodology should be chosen according to these decisions.

Seminar August 25–30, 2024 – https://www.dagstuhl.de/24352
2012 ACM Subject Classification Information systems → Information retrieval; Information systems → Recommender systems; Computing methodologies → Natural language processing
Keywords and phrases Conversational Agents, Evaluation, Information Access, Dagstuhl Perspectives Workshop
Digital Object Identifier 10.4230/DagRep.14.8.53

1 Executive Summary

Christine Bauer
Li Chen
Nicola Ferro
Norbert Fuhr

In this Dagstuhl Perspectives Workshop, a general model for the evaluation of CONversational Information ACcess (CONIAC) systems was developed: Conversational Agents Framework for Evaluation (CAFE).

The framework starts from the assumption that a CONIAC system will be able to (i) *interact* with users more naturally and seamlessly, (ii) guide a user through the process of *refining* and *clarifying* their needs, (iii) *aid decision-making* by making *personalized recommendations and information* while being able to *explain* them, and (iv) *generate*, *retrieve* and *summarize* relevant information.

* Editor / Organizer

Conversational Agents: A Framework for Evaluation (CAFE), *Dagstuhl Reports*, Vol. 14, Issue 8, pp. 53–58
Editors: Christine Bauer, Li Chen, Nicola Ferro, and Norbert Fuhr

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

CAFE distinguishes six major elements of an evaluation design:

- **Stakeholder goals.** Stakeholders of a CONIAC system may have diverse goals that might or might not be directly accessible to system designers or evaluators and must often be implicitly inferred in evaluation. CONIAC systems might also have multiple goals ranging from end users having (in-)direct information needs, to platforms deploying CONIAC systems interested in content usage, user engagement, impression generation, and user retention, to name a few.
- **Tasks.** CONIAC involves tasks characterized by an information need (which may be specific or rather vague), human involvement, goal orientation, and mixed initiative between the user and the system. While some tasks and information needs may benefit from introducing a conversationally competent system, others may not, depending on the complexity of the task or need.
- **User aspects.** When developing an evaluation framework for CONIAC systems, it is crucial to consider user-specific aspects, such as preferences, specialized needs, expertise types, and background characteristics, which may make conversational systems more beneficial than non-conversational alternatives.
- **Criteria.** The scope of evaluation can range from single-turn interactions to entire conversations and long-term system usage, each requiring different criteria for assessment. Additionally, the temporal dimension, which examines how the system's performance changes over time, is a critical factor that can intersect with both stationary and dynamic properties. Criteria may be system-centric, user-centric, or both. The former regard hardware and software aspects like e. g. efficiency, accuracy, comprehensiveness, and verifiability. For the latter, we can distinguish between conversation-oriented (like e. g. adaptability, coherence, fluency), content-oriented (like e. g. continuance, controllability, perceived accuracy, understandability), and consequences-oriented measures (like e. g. addiction, benevolence, decision quality, confidence, trust).
- **Methodology.** In addition to the standard distinction of user-focused and system-focused methodologies, our evaluation framework categorizes evaluation methodologies also according to the employed time model – a dimension especially relevant for CONIAC. This dimension ranges from stationary methodologies like single-interaction experiments to methodologies like controlled lab studies that allow for continuous measurements such as physiological ones.
- **Measures.** Finally, we allow for measures that typically focus on the system's ability to provide accurate, relevant, and timely information during interactions. Measures include objective measures of effectiveness and subjective notions such as perceived effectiveness or user satisfaction (e. g., self-reported satisfaction). By incorporating both objective as well as subjective (self-reported) measures, evaluators can better understand the system's strengths and areas for improvement.

When designing an evaluation, the first step is to identify the stakeholders and their goals that need to be addressed. Based on the goals, the user tasks to be studied in the evaluation have to be defined, as well as the user aspects to be considered. The central element of an evaluation are the criteria to be focused on, which can be determined by the stakeholder goals. The chosen criteria restrict the range of possible evaluation methods (e. g. any user-centric criterion requires the involvement of actual users in the evaluation procedure). Finally, an appropriate measure has to be defined for any quantitative criterion.

2 Table of Contents

3 Overview of Talks

3.1 Conversational Search in 2019

Avishek Anand (TU Delft, NL, neil.hurley@ucd.ie)

In this talk we reflect on the results and insights from the last Dagstuhl Seminar in 2019 on conversational search (`https://www.dagstuhl.de/19461`). There are multiple definitions of conversational search systems or CSS and we looked at the Dagstuhl Typology. We also reflected on some of the challenges of the evaluation of CSS systems. Finally, we discussed about some potential open problems and challenges in the era of LLMs.

3.2 Preferences are Constructive: How to Build and Evaluate Better Conversational Interfaces that Really Give Guidance (with LLMs)?

Martijn C. Willemsen (TU Eindhoven, NL & JADS – 's-Hertogenbosch, NL, M.C.Willemsen@tue.nl)
Bart Knijnenburg (Clemson University, US, bartk@clemson.edu)

Recommender systems build user models to be able to predict users' preferences. However, preferences are volatile and often constructed while in the process of making decisions. In this talk we discuss ways in which recommender systems can go beyond just automatically providing recommendations to learn preferences better via active preference elicitation, interactive recommender systems and conversational interfaces such as critiquing-based recommender systems. We then discuss how such decision guidance should guide future developments of modern conversational agents including LLMs, and we discuss some of the pitfalls such as the persuasive nature and anthropomorphism that might users to over-trust such systems.

3.3 Conversational Recommenders: Reflecting on the Good, the Bad, and the Unknown

Maria Soledad Pera (TU Delft, NL, M.S.Pera@tudelft.nl)

In this talk, we take a somewhat provocative (or rather biased) approach to examining the differences (or lack thereof) between search and recommendation, and exploring what insights can be gained from research in these areas, particularly in terms of evaluation. We then provide a brief overview of studies on conversational recommenders published since the early 2000s, emphasizing evaluation perspectives. Finally, we discuss the challenges of evaluating conversational recommenders throughout different stages of the recommendation-generation process, including the choice of objectives to assess (simultaneously), the "right" metrics to use, data limitations, and how LLMs might increase the complexity of the evaluation process.

3.4 The Challenges and Opportunities in Evaluating Generative Information Retrieval

Mark Sanderson (RMIT University – Melbourne, AU, mark.sanderson@rmit.edu.au)

Evaluation has long been an important part of information retrieval research. Over decades of research, well established methodologies have been created and refined that for years have provided reliable relatively low cost benchmarks for assessing the effectiveness of retrieval systems. With the rise of generative AI and the explosion of interest in Retrieval Augmented Generation (RAG), evaluation is having to be rethought. In this talk, I will speculate on what might be solutions to evaluating RAG systems as well as highlighting some of the opportunities that are opening up. As important as it is to evaluate the new generative retrieval systems it is also important to recognize the traditional information retrieval has not yet gone away. However the way that these systems are being evaluated is undergoing a revolution. I will detail the transformation that is currently taking place in evaluation research. Here I will highlight some of the work that we've been doing at RMIT University as part of the exciting, though controversial, new research directions that generative AI is enabling.

Acknowledgements

We thank Schloss Dagstuhl for hosting us.

Participants

- Avishek Anand
TU Delft – Delft, NL
- Christine Bauer
Paris Lodron University – Salzburg, AT
- Timo Breuer
TH Köln, DE
- Li Chen
Hong Kong Baptist University, HK
- Guglielmo Faggioli
University of Padua, IT
- Nicola Ferro
University of Padua – Padova, IT
- Ophir Frieder
Georgetwon University – Washington, DC, US
- Norbert Fuhr
Universität Duisburg-Essen – Duisburg, DE
- Hideo Joho
University of Tsukuba – Ibaraki, JP
- Jussi Karlgren
Silo AI – Helsinki, FI
- Johannes Kiesel
Bauhaus-Universität Weimar, DE
- Bart Knijnenburg
Clemson University, SC, US
- Lien Michiels
imec-SMIT, Vrije Universiteit Brussel, BE & University of Antwerp, BE
- Andrea Papenmeier
University of Twente – Enschede, NL
- Maria Soledad Pera
TU Delft, NL
- Aldo Lipani
University College London, UK
- Mark Sanderson
RMIT University – Melbourne, AU
- Scott Sanner
University of Toronto, CA
- Benno Stein
Bauhaus-Universität Weimar, DE
- Johanne Trippas
RMIT University – Melbourne, AU
- Karin Verspoor
RMIT University – Melbourne, AU
- Martijn C. Willemsen
TU Eindhoven, NL & Jheronimus Academy of Data Science – 's-Hertogenbosch, NL