INFACT: An Online Human Evaluation Framework for Conversational Recommendation

Ahtsham Manzoor*, Dietmar Jannach

University of Klagenfurt, Universitätsstraße 65-67, Klagenfurt am Wörthersee, 9020, Austria

Abstract

Conversational recommender systems (CRS) are interactive agents that support their users in recommendation-related goals through multi-turn conversations. Generally, a CRS can be evaluated in various dimensions. Today's CRS mainly rely on *offline* (computational) measures to assess the performance of their algorithms in comparison to different baselines. However, offline measures can have limitations, for example, when the metrics for comparing a newly generated response with a ground truth do not correlate with human perceptions, because various alternative generated responses might be suitable too in a given dialog situation. Current research on machine learning-based CRS models therefore acknowledges the importance of humans in the evaluation process, knowing that pure offline measures may not be sufficient in evaluating a highly interactive system like a CRS.

In this work, we provide a user-centric evaluation approach to conversational recommendation along with the **INFACT**, an onl**I**ne huma**N** evaluation **F**ramework for convers**A**tional re**C**ommender sys**T**ems, which can be used to assess the suitability of system responses in a given dialog situation. The INFACT framework is prepared to enable the crowdsourcing of the evaluation task, where various CRS can be integrated for comparison. We have successfully applied the INFACT framework for conducting a number of user studies in our previous research. We believe that our study design along with the INFACT framework can be helpful in facilitating user-centric studies in domains such as dialog systems, machine translation, or Q&A. We release the source code of the framework at https://github.com/ahtsham58/INFACT.

Keywords

Conversational recommender systems, evaluation, user-centric studies, dialog systems

1. Introduction

Conversational recommender systems (CRS) support their users in finding items of interest through multi-dialogs, often in natural language [1]. A CRS is generally considered a highly interactive system, where users converse with the agent and seek for recommendations. Due to the highly interactive nature, modern CRS are generally complex and consist of multiple components, see e.g., [1, 2, 3, 4]. Overall, the eventual goal of a CRS is to support non-trivial yet useful interactions with their users [5].

Evaluating the usefulness of CRS in the academic environment is generally challenging and can be both time and resource intensive in particular when humans are involved in the loop, see also [6]. For example, assessing the quality of responses and thereby dialogs is as important as assessing the quality of the underlying *recommendation algorithms*. Mostly the research community relies on *offline* evaluation approaches using historical datasets in order to understand how good an algorithm performs. Such an approach can be appropriate in evaluating the prediction capability of an algorithm, e.g., which item a user will consider to consume or rate highly. However, these evaluation approaches are unable to inform about the quality perceptions from a user's perspective. For example, whether the made recommendations are acceptable to the user or if the made recommendations will assist users in discovering new yet relevant items.

Furthermore, assessing linguistic aspects such as consistency, naturalness or fluency as a proxy for the language quality of the system's responses is challenging in its own. In this context, to assess language quality, researchers mainly apply offline metrics, e.g., distinct N-gram and Perplexity, or they compare the system's responses with their ground truths using the BLEU [7] or NIST [8] scores, see, e.g., [9, 10]. However, these offline metrics do not inform us whether the response is grammatically and semantically complete or if the response is *meaningful* given the previous dialog history. Moreover, in reality, the system might respond to a user's utterance in a meaningful way, but may not match the ground truth [11]. Ultimately, offline linguistic metrics may therefore not fully inform us about the users' quality perceptions in practice.

Current research in CRS acknowledges the importance of humans in the evaluation process, and this also holds

4th Edition of Knowledge-aware and Conversational Recommender Systems (KaRS) Workshop @ RecSys 2022, September 18–23 2023, Seattle, WA, USA.

*Corresponding author.

ahtsham.manzoor@aau.at (A. Manzoor);
dietmar.jannach@aau.at (D. Jannach)
https://ahtsham58.github.io/ (A. Manzoor); https://www.aau.at/en/aics/research-groups/infsys/team/dietmar-jannach/ (D. Jannach)
0000-0001-9418-753 (A. Manzoor); 0000-0002-4698-8507 (D. Jannach)

for most recent "end-to-end" learning approaches, where deep neural network models are trained using recommendation dialogs collected between humans, see, e.g., [12, 13, 14, 15, 16]. In these recent works, we therefore find studies involving humans, and experiments are conducted using various evaluation methodologies. However, such evaluations often have limitations. For example, in [17] human judges were asked to provide a *relative* ranking of the responses by different systems. In case of relative comparison, it remains unfortunately unclear if any of the compared systems are useful at all [18]. Moreover, the scope of such studies seems limited as in many cases there are only a few evaluators involved and sometimes the details regarding the background of the human judges are missing too. Also, often only the language quality of the responses is the main focus of such studies. An assessment if the made recommendations are suitable in an ongoing dialog context is sometimes missing, see, e.g., [14, 19].

In this work, we present a user-centric evaluation approach to CRS that can be used to assess both linguistic and recommendation quality aspects along with the **INFACT**, an onl**I**ne huma**N** evaluation **F**ramework for convers**A**tional re**C**ommender sys**T**ems. We have applied our evaluation approach for a number of studies [11, 20, 21], using the INFACT framework as a basis. To easily involve a larger set of subjects than in earlier studies, the INFACT framework is prepared to support the evaluation task through online crowdworking platforms. We believe that our study design may serve as a blueprint for future human evaluation studies for CRS. It can furthermore be easily extended to evaluate dialogs systems, machine translation, or Q&A tasks. We release the source code of the INFACT framework at https://github.com/ahtsham58/INFACT.

2. Related Work

According to recent surveys on CRS [1, 6], we can generally observe three main dimensions in which a CRS can be evaluated: *(i) effectiveness of task support*, i.e., the ability of the system to support a recommendation-related task, *(ii) efficiency of task support*, i.e., how much effort is required by the user, and *(iii) conversation quality and usability*, which may cover aspects like fluency, naturalness, or the consistency of the system responses. All of these aspects can contribute to the success of a CRS in practice.

From a methodological standpoint, quality measurements are typically either made with the help of computational ("offline") experiments or with studies involving humans in the loop. In offline experiments, system effectiveness is often evaluated in terms of *recommendation quality*, where metrics like *precision* or *recall* are used as proxies. In addition to recommendation quality, offline experiments are common for assessing the *dialog quality*. Specifically, linguistic measures such as *distinct N-gram* and *perplexity* to assess the diversity and fluency of the system-generated responses have been applied in various recent works on CRS, see, e.g., [2, 12, 13, 14]. Similarly, inspired by the machine translation domain, metrics like *BLEU* or *NIST* are applied in several works on CRS, where the system response is compared with a given ground truth in order to estimate the overall quality of the generated responses, see, e.g., [2, 15, 19, 22].

Given the interactive nature of CRS, studies involving humans in the evaluation process are not uncommon in the CRS literature. Such studies mainly assess the quality aspects from a user's perspective. See [23, 24, 6] for a set of relevant quality attributes. Looking at most recent works, different studies were conducted in which human judges were tasked to *rate* or *rank* system responses in various dimensions. For example, in [17], ten evaluators were asked to rank the responses by baseline recommenders and the proposed system in terms of the overall quality. Similarly, in [19], the authors reported a study in which human evaluators had to rate the responses on a scale from 1-5 in terms of *Fluency*, *Consistency*, *Naturalness*, *Persuasiveness*, and *Engagingness*. Similarly, in a recent work [14], five human judges were given the task to rate the responses generated by the proposed system and various baselines on a scale from 1-3 in terms of *Fluency*, *Coherence*, *Informativeness*, and *Interoperability*. Similar examples of such studies can also be found in [13, 12, 15, 25].

Interestingly, such studies mainly focused on *linguistic* aspects of the systems' responses. If the recommendations themselves were considered meaningful—a main aspect in terms of a system's usefulness—was not assessed with the help of human judges but rather evaluated through offline analyses. On the other hand, the generic concept of *meaningfulness* of a response can be used to evaluate both aspects, i.e., language and recommendation quality, see, e.g., [11, 20, 21]. Moreover, in such studies the details about the study setup and background of the evaluators were quite brief and sometimes missing at all. Furthermore, several studies were conducted with a small number of judges, for example, in [12, 13], *three* judges were involved and no information was provided regarding their linguistic expertise.

Given the potential limitations of offline experiments and of user studies with unclear significance, we provide a study design that may serve as a template for scalable human-centric evaluation studies of dialog systems. Next, we explain the experiment design of our evaluation approach and highlight the features that the INFACT framework offers to support the evaluation task through online crowdworking platforms.

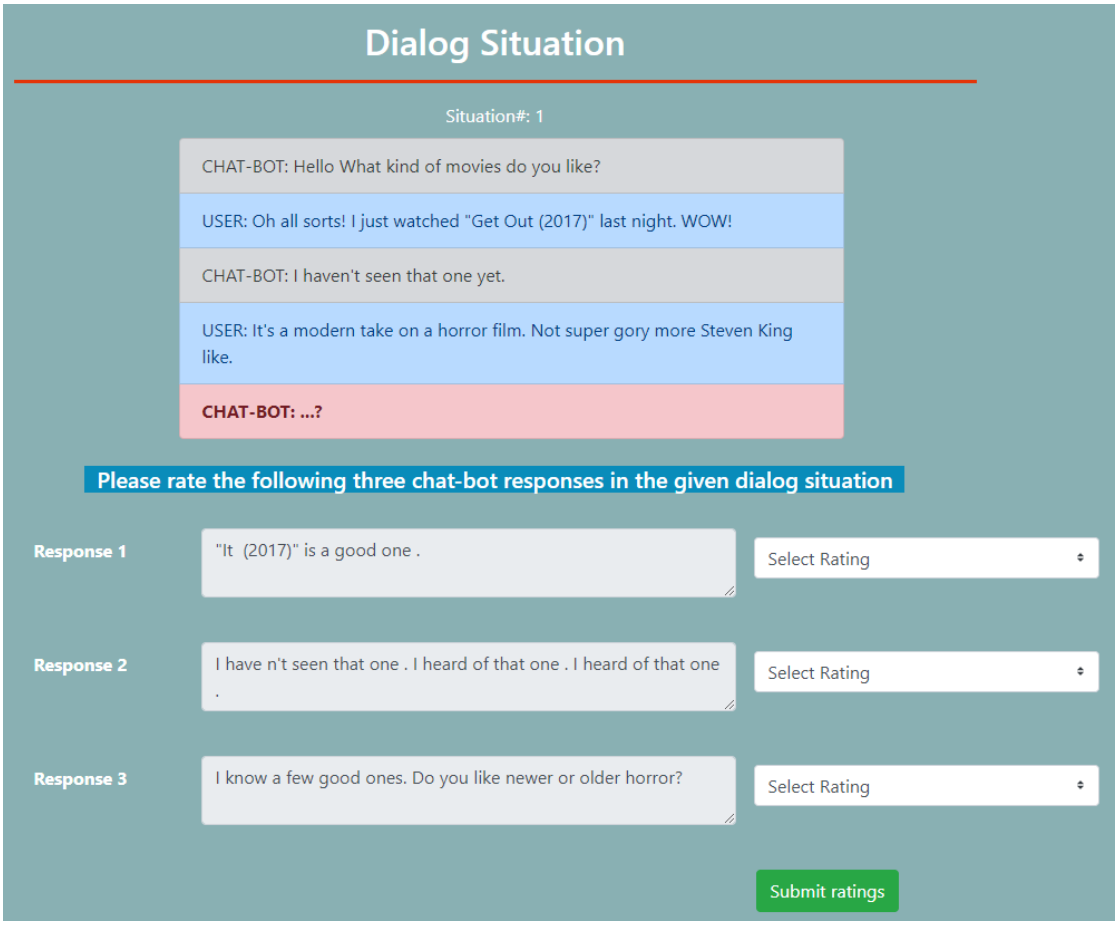

Figure 1: Response rating user interface

3. Experiment Design with INFACT

General Design The INFACT framework is developed based on the concept that user-centric evaluations are vital to assess the effectiveness of highly interactive systems like CRS. Moreover, evaluating such systems requires studies at scale in order to investigate the quality of both linguistic and recommendation aspects in practice. Specifically, in our approach we ask human subjects to assess dialog continuations ("system responses") provided by a CRS given a piece of dialog ("dialog situation") using one or more quality criteria. In our own studies, we use the ReDial dataset [17] consisting of real-world dialogs for such evaluations. We note that various recent current CRS approaches rely on this dataset or similar ones to *generate* or *retrieve* suitable responses, e.g., [15, 19, 26]. To avoid biases and to receive feedback for all stages of the dialog, the dialog situations to be evaluated are selected from such datasets at random [27, 28]. In addition, we assume to have a larger set of participants than many existing studies; in particular we consider crowdsourcing to be helpful.

Such an approach can be instantiated in various ways, depending on the research question(s). For example, an experiment could include one or more algorithms to evaluate for each participant. Similarly, there could be one or more questions regarding the dialog quality and for each dialog continuation. Moreover, each participant can be tasked to assess only one or more dialog situations and the feedback scale could be different too.

When deciding on these specifics, it is important to keep the cognitive load and the overall workload for the study participants in mind. Moreover, the specific design also determines how many participants are required to achieve a sufficient number of human judgements. In our own experiments, as discussed later, we decided to ask participants to assess exactly three different dialog

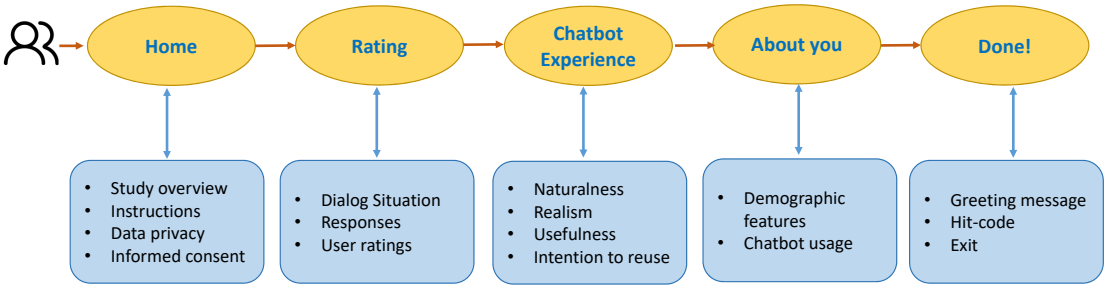

Figure 2: INFACT Workflow Diagram

continuations, and they had to assess ten such dialog situations. As a result, we obtain multiple assessments from each participant, which helps to keep the number of participants low. However, using such a design, it is important to check for intraclass (per user) correlations in the statistical analyses.

Arbitrary dialog datasets can be used with our framework, as long as they follow the format used in the framework, which is currently based on the ReDial dataset, as mentioned. Finally, arbitrary post-task questionnaire elements can be introduced, and the INFACT framework implements a number of typically required functionalities, e.g., for persistently storing the feedback into a database.

A Specific Implementation In our experiments [11, 20, 21], each participant is presented with dialog situations that always start from the first utterance and end with a *user* utterance. Below the dialog situation, as shown in Figure 1, we show three responses to the last user utterance by three different CRS under comparison. To highlight and differentiate item recommendations from regular language words in the dialog situation and responses, we enclosed item names, e.g., in this case movie titles, in double quotes[1].

In our experiments, the only question for the study participants was to *independently* assess (or rate) the quality of each response in terms of the *meaningfulness* of the responses in the given dialog context. To obtain fine-grained assessments, we use a 5-point scale labeled from 'Entirely meaningless' to 'Perfectly meaningful', which can be modified depending on the research question. To avoid any sequential rating bias by the evaluators, the order of showing responses to the user is randomized.

On the landing page, we provide specific instructions to the evaluators about how to judge the *meaningfulness* of responses. For example, a response by the specific system should be logical continuation of the provided dialog situation. In case an item recommendation(s) is included in the response, it has to match to the user's stated interest and preferences. If the system response does not include a movie recommendation, e.g., chit-chat sentence, the participants are supposed to rate the *meaningfulness* of the response as a reply to the user's last utterance while considering also the context of the ongoing dialog. Overall, human judges are supposed to provide ratings based on their *subjective* quality perceptions.

From the linguistic perspective, a deep discussion of the concept '*meaningfulness*' is provided in [29], where the author makes a distinction between 'grammaticality' and 'meaningfulness'. In our study design, instead of challenging participants with complex linguistic concepts or considerations, we provide examples and guidelines when a response can be considered meaningful or meaningless.[2]

In this way, each evaluator assessed *ten* such dialog situations. However, this is a configurable parameter that can changed based on the experiment design. On submission, we store the rating scores, including the dialog situation, corresponding responses, and the overall time it took for the evaluator to rate, in a NoSQL cloud database. We relied on a NoSQL database as it offers flexible, affordable, and scalable database management.

In order to check if participants are attentive during the study, one of the ten dialog situations that appear in random order is used as an *attention check*. Specifically, in one of the three responses for this dialog situation, we asked the participants to select a particular rating from the given scale. The attention check was considered to be failed whenever a study participant did not select the required score. In this case, we completely discard all data from such unreliable crowdworkers[3]. Furthermore, apart

[1] A detailed description about how to prepare the evaluation data is explained in the online repository.

[2] The term "meaningfulness" is also used in the context of a human evaluation in [9]. Differently from our work, the term "meaningful" is used in [9] to summarize other evaluation dimensions in an informal way. We note that our framework can easily be configured to collect annotator feedback on several dimensions, e.g., fluency, coherence, or informativeness as in [14].

[3] We provide the Python script to automatically parse the study data

Table 1
Example Questions from a Post-Task Questionnaire

	Questions
Q1	I found the presented dialogues natural.
Q2	The presented dialogue situations look realistic.
Q3	I could imagine that such dialogues also happen between humans.
Q4	Considering only the best responses found in each dialogue, I would find the chat-bot useful.
Q5	Considering only the best responses found in each dialogue, I would probably use such a movie recommendation chat-bot in the future.

Table 2
Example questions for Participant Demographics

Demographic Feature	Scale
Gender	Male Female Other
Age	18-25 25-30 30-35 35-45 45-70
English fluency level	Beginner Intermediate Fluent Advanced
Education level	High school or less Bachelor's Master's Doctorate Other
Frequency of watching movies	Everyday Several times a week Once in a week Once every few weeks Less frequent
Ever interacted with a chat-bot	Yes No
Ever interacted with a chat-bot for getting movie recommendations	Yes No

from the explicit attention check, the INFACT framework is equipped with various *implicit* checks to deal with the potentially unreliable crowdworkers such as time interval (in seconds) for each individual event, overall study completion time, etc.

In our specific experiments, no particular training or expertise is required by the crowdworkers to participate. To fulfill the task, the crowdworkers were asked for their subjective assessment regarding the generated system responses in terms of their meaningfulness. In case participants should answer more complex questions, e.g., regarding fluency or interpretability as in [14], appropriate measures must be taken to ensure that the crowdworkers are able to fulfill the task reliably, e.g., by providing more instructional material or by requiring certain skills.

After the submission of ratings for ten such dialog situations, a post-task questionnaire is shown to the participants, where we collect general feedback regarding the quality of dialogs, demographics, and general remarks or suggestions. Table 1 shows parts of a questionnaire for dialog quality, asking, for example, if the shown dialogs feel natural, realistic, or useful. An example list of demographic questions is shown in Table 2. These questionnaires can be modified depending on the research question(s). The overall workflow of our experiment design is visualized in Figure 2.

To enable and support studies through crowdworking platforms like Amazon Mechanical Turk, Prolific, etc., we used a pre-implemented feature like hit-code generation. Our assumption is that a large number of human judges are needed for the evaluation, hence the INFACT framework is prepared accordingly. Technically, the INFACT framework is a web-based application developed using the Django framework in Python 3.0, and Bootstrap 4.4.

4. Conclusion

Research on conversational recommender systems (CRS) has attracted increased attention in recent years. The most recent proposals on CRS, and in particular ones that follow an end-to-end learning paradigm, mainly rely on computational measures in order to demonstrate the effectiveness of their systems in comparison to different baselines. However, the aspects that contribute to the success or failure of a CRS may not be fully assessed without involving humans in the evaluations process.

In this work we provide a user-centric evaluation approach for CRS, which can be used to investigate both recommendation and linguistic quality aspects of system responses in a given dialog. Since the scope of several studies reported in the context of recent CRS seems limited, we propose an online evaluation tool which can be used to perform human evaluations at scale with the help of crowdworkers. Due to the modular and flexible nature of the architecture underlying the INFACT framework, it can be modified and adapted in the context of a similar study design. For example, replacing the scale or metric requires only a few modifications on a single template page. Ultimately, we hope that our user-centric evaluation approach can be considered as a template to facilitate the design of similar studies in domains like dialog systems, Q&A or machine translation.

stored in the JSON format on the cloud.

References

[1] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, ACM Computing Surveys 54 (2021) 1–36.

[2] K. Chen, S. Sun, Knowledge-based conversational recommender systems enhanced by dialogue policy learning, in: IJCKG '21, 2021, pp. 10–18.

[3] A. Rana, D. Bridge, Navigation-by-preference: A new conversational recommender with preference-based feedback, in: IUI '20, 2020, p. 155–165.

[4] D. Jannach, ADVISOR SUITE – A knowledge-based sales advisory system, in: ECAI '04, 2004, pp. 720–724.

[5] D. Jannach, L. Chen, Conversational Recommendation: A Grand AI Challenge, AI Magazine 43 (2022).

[6] D. Jannach, Evaluating conversational recommender systems, Artificial Intelligence Review forthcoming (2022).

[7] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: A method for automatic evaluation of machine translation, in: ACl '02, 2002, p. 311–318.

[8] G. Doddington, Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, in: HLTR '02, 2002, pp. 138–145.

[9] L. Nie, W. Wang, R. Hong, M. Wang, Q. Tian, Multimodal dialog system: Generating responses via adaptive decoders, in: MM '19, 2019, pp. 1098–1106.

[10] Q. Chen, J. Lin, Y. Zhang, H. Yang, J. Zhou, J. Tang, Towards knowledge-based personalized product description generation in e-commerce, in: KDD '19, 2019, pp. 3040–3050.

[11] A. Manzoor, D. Jannach, Towards retrieval-based conversational recommendation, Information Systems (2022) 102083.

[12] Q. Chen, J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, J. Tang, Towards knowledge-based recommender dialog system, in: EMNLP-IJCNLP '19, 2019, pp. 1803–1813.

[13] K. Zhou, W. X. Zhao, S. Bian, Y. Zhou, J.-R. Wen, J. Yu, Improving conversational recommender systems via knowledge graph based semantic fusion, in: KDD '20, 2020, pp. 1006–1014.

[14] J. Zhou, B. Wang, R. He, Y. Hou, CRFR: Improving conversational recommender systems via flexible fragments reasoning on knowledge graphs, in: EMNLP '21, 2021, pp. 4324–4334.

[15] K. Zhou, Y. Zhou, W. X. Zhao, X. Wang, J.-R. Wen, Towards topic-guided conversational recommender system, in: ICCL '20, 2020, pp. 4128–4139.

[16] J. Zou, E. Kanoulas, P. Ren, Z. Ren, A. Sun, C. Long, Improving conversational recommender systems via transformer-based sequential modelling, in: SIGIR '22, 2022, pp. 2319–2324.

[17] R. Li, S. E. Kahou, H. Schulz, V. Michalski, L. Charlin, C. Pal, Towards deep conversational recommendations, in: NIPS '18, 2018, pp. 9725–9735.

[18] D. Jannach, A. Manzoor, End-to-end learning for conversational recommendation: A long way to go?, in: IntRS Workshop at RecSys '20, Online, 2020.

[19] S. A. Hayati, D. Kang, Q. Zhu, W. Shi, Z. Yu, INSPIRED: Toward sociable recommendation dialog systems, in: EMNLP '20, 2020.

[20] A. Manzoor, D. Jannach, Conversational recommendation based on end-to-end learning: How far are we?, Computers in Human Behavior Reports (2021) 100139.

[21] A. Manzoor, D. Jannach, Generation-based vs. retrieval-based conversational recommendation: A user-centric comparison, in: RecSys '21, 2021.

[22] A. Bartl, G. Spanakis, A retrieval-based dialogue system utilizing utterance and context embeddings, in: ICMLA '17, 2017, pp. 1120–1125.

[23] P. Pu, L. Chen, R. Hu, A user-centric evaluation framework for recommender systems, in: RecSys '11, 2011, pp. 157–164.

[24] Y. Jin, L. Chen, W. Cai, P. Pu, Key qualities of conversational recommender systems: From users' perspective, in: HAI '21, 2021, pp. 93–102.

[25] F. Pecune, S. Murali, V. Tsai, Y. Matsuyama, J. Cassell, A model of social explanations for a conversational movie recommendation system, in: HAI '19, 2019, p. 135–143.

[26] D. Kang, A. Balakrishnan, P. Shah, P. Crook, Y.-L. Boureau, J. Weston, Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue, in: EMNLP-IJCNLP '19, 2019, pp. 1951–1961.

[27] W. Cai, L. Chen, Towards a taxonomy of user feedback intents for conversational recommendations, in: RecSys' 19 Late-Breaking Results, 2019, pp. 572–573.

[28] S. Lyu, A. Rana, S. Sanner, M. R. Bouadjenek, A workflow analysis of context-driven conversational recommendation, in: WWW '21, 2021, pp. 866–877.

[29] Y. Wilks, Decidability and natural language, Mind (1971) 497–520.