Factors in crowdsourcing for evaluation of complex dialogue systems

Annalena Bea Aicher, Stefan Hillmann, Isabel Feustel, Thilo Michael, Sebastian Möller, Wolfgang Minker

Angaben zur Veröffentlichung / Publication details:

Aicher, Annalena Bea, Stefan Hillmann, Isabel Feustel, Thilo Michael, Sebastian Möller, and Wolfgang Minker. 2024. "Factors in crowdsourcing for evaluation of complex dialogue systems." In 13th International Workshop on Spoken Dialogue Systems Technology (IWSDS 2023), Los Angeles, CA, USA, February 21-24, 2023, arXiv:2411.11137. arXiv. https://doi.org/10.48550/arXiv.2411.11137.

Factors in Crowdsourcing for Evaluation of Complex Dialogue Systems

Annalena Aicher, Stefan Hillmann, Isabel Feustel, Thilo Michael, Sebastian Möller and Wolfgang Minker

Abstract In the last decade, crowdsourcing has become a popular method for conducting quantitative empirical studies in human-machine interaction. The remote work on a given task in crowdworking settings suits the character of typical speech/language-based interactive systems for instance with regard to argumentative conversations and information retrieval. Thus, crowdworking promises a valuable opportunity to study and evaluate the usability and user experience of real humans in interactions with such interactive systems. In contrast to physical attendance in laboratory studies, crowdsourcing studies offer much more flexible and easier access to large numbers of heterogeneous participants with a specific background, e.g., native speakers or domain expertise. On the other hand, the experimental and environmental conditions as well as the participant's compliance and reliability (at least better monitoring of the latter) are much better controllable in a laboratory. This paper seeks to present a (self-)critical examination of crowdsourcing-based studies in the context of complex (spoken) dialogue systems. It describes and discusses observed issues in crowdsourcing studies involving complex tasks and suggests solutions to improve and ensure the quality of the study results. Thereby, our work contributes to a better understanding and what needs to be considered when designing and evaluating studies with crowdworkers for complex dialogue systems.

Annalena Aicher
Ulm University, e-mail: annalena.aicher@uni-ulm.de

Stefan Hillman
TU Berlin, e-mail: stefan.hillmann@tu-berlin.de

Isabel Feustel
Ulm University, e-mail: isabel.feustel@uni-ulm.de

Thilo Michael
TU Berlin, e-mail: thilo.michael@tu-berlin.de

Sebastian Möller e-mail: sebastian.moeller@tu-berlin.de · Wolfgang Minker
Ulm University, e-mail: wolfgang.minker@uni-ulm.de

1 Introduction

One crucial step in the development of dialogue systems is their evaluation. This evaluation is challenging, as the definition of what constitutes a high-quality dialogue is not always clear and often depends on the specific application, domain, task, and user group [4, 16]. Even if a definition is assumed, it is not always clear how it can be measured. For example, if we assume that a high-quality dialogue system is defined by its ability to respond with an appropriate utterance, it is not clear how to measure appropriateness or what appropriateness means for a particular system. Furthermore, one could ask users whether responses were appropriate, but as we will discuss below, user feedback is not always reliable for a variety of reasons. Depending on the ability and purpose of a dialogue system it is unclear how to measure appropriateness or what appropriateness means. Furthermore, the evaluation of dialogue systems is very cost- and time-intensive. This is especially true when the evaluation is carried out through user studies, which compensate users for their participation [4]. Therefore, quite a lot of efforts are made, aimed at automating the evaluation, or at least automating certain aspects of the evaluation [19, 20, 21, 22]. But still, as automated metrics do not necessarily capture all aspects of the system's quality, a human evaluation is performed, which usually asks about the naturalness and quality of the generated utterances and flow of dialogue [6, 17].

One possible way to lower these costs and enable a more flexible evaluation procedure is to perform such human user studies via crowdsourcing. Crowdsourcing engages the help of large numbers of people in tasks, activities, or projects, usually via the internet [5]. The application areas of crowdsourcing are very diverse and range from health research areas to the field of market and customer analysis.

Most studies on crowdsourcing tend to focus on its use of software, technology, online platforms, or its application [12, 13]. Still, as Bassi et al. [9] point out, there is a need for further exploration to understand how best to use crowdsourcing for research. This is underpinned by numerous references which describe gaps in the research related to crowdsourcing, including a lack of decision aids to assist researchers using crowdsourcing, and best practice guidelines [10, 7, 11].

Thus, in the herein presented work, we aim to provide a critical examination and discussion of our experience with crowdworking studies in the context of complex (argumentative) dialogue systems and propose some guidelines for researchers who undertake studies with crowdworkers involving complex tasks in the interaction with dialogue systems.

The remainder of the paper is as follows. Section 2 gives an overview of related relevant work and literature. After briefly outlining the essential background information on our crowdsourcing study settings, we discuss the anomalies and problems we encountered in Section 3. We introduce recommendations for respective guidelines in Section 4 and close with a conclusion and a brief discussion of future work in Section 5.

2 Related Work

As crowdsourcing has broad application prospects and significant business value, many studies have been conducted in recent years. The existing literature reviews on crowdsourcing range from general overviews (e.g., [12, 18]) to specific research areas like [13]. Jiang et al. [8] characterize crowdsourcing as suitable for tasks "that are trivial for humans but difficult for computers, such as classification tasks". They describe the advantages of crowdsourcing including faster completion speed, lower costs, higher accuracy, and completion of tasks that computers cannot perform. Thereby crowdsourcing engages the help of large numbers of people in tasks, activities, or projects; usually via the internet [5].

Although a few surveys have attempted to present a more general review of crowdsourcing, they solely introduce definitions of crowdsourcing and/or review typical crowdsourcing systems [8]. A highly cited example of a crowdsourcing platform according to [12] is Amazon's Mechanical Turk (AMT). An overview of AMT as an academic research platform is provided K. B. Sheehan [11] and its strengths (e.g. quick data collection, global respondent pool, relatively low costs, easy handling) and weaknesses (e.g. validity, reliability, ethics) for research are discussed. Fort et al. [37] highlight issues raised by AMT, especially with regard to ethical concerns. Furthermore, Goodman et al. [24] compare AMT participants with other participant samples (community/student) on a set of personality dimensions[1] and classic decision-making biases (i.e., framing effects, the conjunction fallacy, and outcome bias). They recommend that researchers should use AMT taking into account certain conditions (integration of screening questions etc.). Manuvinakurike et al. [14] present a case study in which 200 remote participants were recruited to participate in a fast-paced image matching game via AMT and the authors discuss encountered technical challenges of the study.

Beyond technical challenges also potential problems and issues are discussed among some crowdsourcing-related publications. For instance, Thebault-Spieker et al. [15] discuss how mobile crowdsourcing tasks, situated in the physical world (e.g., checking street signs, running household errands), are influenced by geographical implications (suburbs, etc.). Nevo et al. [3] state that most crowdsourcing experts found that "stand-alone tasks" and tasks with a "clear definition" are best suited for crowdsourcing. Furthermore, they claim that crowdsourcing projects require specific domain knowledge and a "well-developed problem statement". Nevertheless, Noel-Storr et al. [5] stress that there is evidence that a "crowd" of non-specialists is capable of tasks like identifying quantitative studies by topic-based citation screening (i.e., assessing titles and abstracts). Thus, there seems to be a dependency on the expertise, skills, and knowledge of crowdworkers and the complexity and difficulty of the task. This is underpinned by our findings as well and further discussed in Section 3. Furthermore Shumeli et al. [38] draw attention to the lack of ethical con-

[1] Gosling et al. [36] developed a Ten-Item Personality Inventory (TIPI) which is a 10-item short inventory to assess the big-five dimensions (openness to experience, emotional stability, extraversion, agreeableness, and conscientiousness).

siderations related to the various tasks performed by workers, including labeling, evaluation, and production.

Another crucial aspect we also have strongly experienced in our crowdsourcing studies is described by Palogiannidi [1]. She claims that crowdsourcing's main advantage—the great diversity of people groups that participate — is also its main disadvantage. They refer to the supposed "tendency of humans to cheat, whenever it is possible, especially if anonymity is preserved" (ibid. 30). Furthermore cheating was detected in a spoken dialogue system's related crowdsourcing task (assessment of text-to-speech synthesis) by Buchholz and Latorre [29].

To counteract these disadvantages as far as possible and design tasks that will yield a higher quality of the workers' responses Palogiannidi [1] stresses the importance to quantify senses such as freedom, politeness, specificity, etc during the design process of crowdsourcing tasks. On the one hand, the selection of the crowdworkers can be adjusted, especially with regard to their expertise and suitability for a task. Liao et al. [28] proposed an approach that aims to recommend a group of suitable workers through worker personality analysis and community classification. On the other hand, it makes sense to define guidelines and best practice recommendations to conduct crowdsourcing studies, which is also our main objective in this paper. For instance, Sabou et al. [23] discuss crowdsourcing methods for corpus acquisition and propose a set of best practice guidelines based on their experiences. Furthermore, Ramírez et al. [30] examined the current state of reporting of crowdsourcing experiments on the basis of 171 experiments and propose a checklist for crowdsourcing experiments.

With regard to (spoken) dialogue systems for instance Li et al. [33] introduce LEGOEval, an open-source Python-based toolkit that allows researchers to easily develop human evaluation tasks for dialogue systems on AMT in a LEGO plug-and-play fashion. In [35] Jurcicek et al. discuss a framework for evaluating dialogue systems using Amazon Mechanical Turk for recruiting a large group of users and conclude that crowdsourcing provides an effective method of rapidly testing spoken dialogue systems at a modest cost. Furthermore, Huynh et al. [34] stress the importance of the clarity of instructions, examples, fair payment, and low quality to be considered when creating Human Intelligence Tasks [2] so that the data gathered is of the highest quality possible. But the existing literature concerning very advanced and complex tasks, i.e. real-time interactions over a lot of turns with a complex (spoken) dialogue system is very scarce.

According to Buettner [10] the majority of the crowdsourcing papers that discuss complex and creative tasks are related to idea generation, competition, and evaluation by the crowd. Thus in this paper, we aim to close this gap and report our positive and negative experiences with crowdsourcing-based user studies considering complex dialogue systems. To the best of our knowledge, so far there does not exist a respective discussion of crowdsourcing studies for complex (argumentative) dialogue systems nor a suggestion for guidelines to perform such studies.

[2] AMT terminology for tasks like image labeling, semantic labeling and audio transcription.

3 Studies

In this section, we will describe the study settings and evaluate our findings of three different user studies conducted via crowdsourcing. Each study was performed with a different type of dialogue system, indicating that the tasks and their respective complexity were strongly dependent on the respective dialogue system.

3.1 Study 1: Comparison of Modalities in Argumentative Dialogue Systems

The aim of this study was to analyze the impact of different input/output (I/O) modalities (I: drop-down menu / O: text vs. I: speech / O: speech) on the evaluation of an argumentative dialogue system (ADS) and user interest.

3.1.1 Study Settings

Our user study [32] was conducted online via the crowdsourcing platform *Crowdee*[3] with participants from the UK, US, and Australia in the period 12-29th November 2021. All 292 crowdworkers were non-experts without a topic-specific background. The crowdworkers could access the study via their *Crowdee* job board, which showed the title "Use our BEA dialogue system and provide feedback!", the language option English and the estimated task duration. After the job was selected the task description: "to listen to enough arguments to build a well-founded opinion [4] on the topic "Marriage is an outdated institution"[5] (at least 10 arguments)" was displayed. If the participants confirmed their willingness to proceed and accepted the data protection regulation guidelines they were redirected to the study introduction page. There, a short text and demo video (menu: 1:32min, speech: 2:02min) explained how to interact with the assigned version of the dialogue system (either menu-/GUI-based or speech-based input modality). Particular attention was paid to the speech interaction system, where the users had to express their requests in their own words, whereas, in the menu system, they had to click on a response option. Therefore, the speech system represented a significantly more demanding and complex setting as the user had to understand the arguments, how to interact with the system, and formulate the respective requests. Thus, in the speech system participants get an instruction how about the actions they can perform and how to express

[3] https://www.crowdee.com/

[4] BEA is a cooperative argumentative dialogue system that has neither persuasion nor its own (ethically critical) agenda. It merely presents the users with arguments on a certain topic on their demand and gives them the opportunity to express their opinion.

[5] This sample debate serves as knowledge base for the arguments and is taken from the *Debatabase* (https://idebate.org/debatabase, last accessed 19th November 2021).

them. Furthermore, users of the speech system could ask for assistance during the interaction if they were unsure about what to do next. In response, the system displayed all possible interaction options on the graphical user interface of the dialogue system. One of the main objectives of this study was to estimate and analyze changes in the user's content-related interest. Therefore, the users were asked to rate their interest in the currently presented aspect/argument as described in detail in Aicher et al. [32]. In case the users lost interest in participating/continuing with the task, it was repeatedly stressed that the end of the interaction could be chosen freely as soon as at least ten arguments were heard. The minimum amount of arguments was chosen to ensure that enough data could be collected. The first 139 participants interacted with our ADS via drop-down menu input, and the other 153 via speech. The latter user group was strongly advised to use headsets to reduce background noise. Each participant completed the study online without any supervision. Before and after the conversation with the system the participants had to rate statements (meta questions regarding the interaction, e.g. user's perception and impression of the system etc.) on a five-point Likert scale. A part of these questions was taken from a questionnaire according to ITU-T Recommendation P.851[6] [31]. Furthermore, these questionnaires contained three control questions, to identify participants who did not fill out the questionnaire seriously or carefully, e.g., by randomly selecting options on the Likert scale [24]. In particular, Aker et al. [2] stress that real tasks require high precision objective questions such as the maths questions in combination with the free text design can be used to filter out non-compliant or "unprecise workers". Retrospectively, it became clear that establishing some demographic facts of the participants (gender, age,..) would have been helpful for completeness and for the interpretation of the results, which should therefore be included in further studies.

The evaluation of these control questions and considering additional feedback revealed that the data of 90 participants showed anomalies. Their data were excluded according to previously defined exclusion criteria: Contradictory answers in control questions in the questionnaire, taking less than 30 seconds to read through the introduction and watch the introduction videos, taking less than 120 seconds to answer the 40+15 questions in the final questionnaire and feedback indicating that problems occurred during the interaction or participants reported that they did not know what to do. This led to a total number of data records of 202 participants (menu: 104, speech: 98) which were included in the analysis.

3.1.2 Evaluation

On average the participants interacted with the ADS for 31:45 minutes (menu: 27:57, speech: 35:34). This significant difference can be explained by the fact that the spoken interaction (speaking and hearing) inherently takes longer than clicking on an option in a drop-down menu and reading the response. This also displays one of the disadvantages of the menu system as users do not have to attentively read the

[6] Such questionnaires can be used to evaluate the quality of speech-based services.

presented argument but can just move on by clicking the next move, whereas in the spoken interaction, the user is forced to listen to the ADS.

But even though the average time the users of the menu system interacted with the ADS is lower, the number of provided arguments is significantly higher in the menu system. 9.6%/17.3% of the menu/speech system participants quit the conversation after hearing the minimum number of arguments (in total: 13.4%). Most of the participants heard between 20-30 arguments of 72 available arguments, which indicates that even though the given task was quite subjective, the crowdworkers interacted with the system longer than they had to[7]. These findings underpin the ones of Nevo et al. [3], who described a problem with the strict duration of specific steps, which is unlikely to occur if the participants are not bound to any stepwise time limits during the interaction.

Regarding all aspects of the evaluation questionnaire, the menu system outperformed the speech system significantly. This was particularly observed in aspects that rated if the system provided the expected information or errors in the interaction which occurred. To a certain extent, this points to a lack of processing of the user utterances and can either be explained by errors in the automated speech recognition (ASR) or the natural language understanding (NLU) module of the ADS. By checking the dialogue logs of the interactions with users in the speech system, we found that about 15% of all spoken utterances were recognized erroneously by the ASR. Even though the NLU module matched about a third of these erroneously recognized utterances to the correct (i.e., intended) action, the impact is still significant. The errors of the ASR might partly be explained by the participants not wearing headsets during the interaction which led to additional noise. In order to avoid or mitigate problems due to insufficient speech or audio quality one could either switch to typed input or in order to preserve the speech modality and ensure suitable environmental conditions by conducting the study in the lab on site. It is also possible to check the audio quality on the user's device, but this is technically complex (i.e. error-prone), detects problems *after* the user has started the task and prolongs the overall task.

Still, there are some observations that strongly suggest that this outcome was also influenced by the unsupervised introduction to the task[8]. We noticed inconsistencies in the user behavior, as users repeated their request multiple times, without reacting to the system's answer to choose another action. This indicates that in contrast to the menu system, where the users were always only displayed the available option, the speech users had to figure out what actions they can perform and how to suitably formulate it. Even though the system's design incorporated a "Help" button, as well as the "available" options action, only 15% of the speech users used them. This can be explained by the fact that only 35% of the users spend enough time on the introduction website to read through the explanation and watch the according video properly.

[7] The users could take as much time as they wanted for each response, as well as for the overall interaction.

[8] Unfortunately the introduction could be skipped rather easily

Unfortunately, we did not include any test questions before the interaction to check if the speech participants thoroughly read/viewed the interaction instructions. Thus the speech system displayed a far more complex task whereas the menu system just displayed the actions the user had to choose from. Even though the latter system presented less flexibility and was perceived as static and not natural, the speech system was perceived as too complicated and unreliable. This is underpinned by feedback that was received from the crowdsourcing platform *Crowdee* after the study, e.g., stating that "It was not possible to do what I wanted to do. I repeated myself many times"/"I was stuck in the argument and could not get back"/"The idea is cool, still, sometimes things got messy when I could not proceed to the argument I wanted to go to.". Therefore, it might useful to conduct a double-staged study, which ensures that all participants passing the first stage have well-understood how the interaction with the ADS works. Another option would be to conduct on-site, where the participants can be supervised and the environmental conditions are controlled.

A further indicator of the higher complexity of the speech system can be seen in the number of study abortions, as it needed seven times more participants to achieve the targeted number of participants.

In general, the results clearly show that the input/output modalities and respective difficulties/problems decrease the rating of the general impression of the system, even in aspects that have no relation to the former. For example, the incremental approach to present arguments, the sufficiency of different options, or the conclusiveness of arguments, which depend on the content but not on the modality, are rated in the speech system significantly worse than in the menu system. Therefore, it is crucial to solve the identified issues and take precautions, e.g., in the design of crowdsourcing studies to avoid them.

3.2 Study 2: Evaluation of a Pipeline for Argumentative Dialogues

In order to introduce arguments into dialogue systems in a meaningful way, a suitable data structure for the dialogue flow is required. In our previous works [27] a tree structure that allows access to arguments in dialogue form was developed. The tree represents both pro and con arguments to a statement and their relation to each other. Until now, these tree structures were created manually requiring a lot of effort. Therefore, we [27] developed a pipeline to feed arguments from an argument mining search engine into a tree structure suitable for argumentative dialogues. These structures were then used in a agent-agent dialogue system, where each agent had either a pro or con position on the discussed topic. In order to test the pipeline, several studies were conducted, which were mainly intended to check the (logical) coherence of the arguments and the overall dialogue.

3.2.1 Study Settings

In order to measure coherence, three categories were used: Comprehension, Reference, and Polarity. Each category was converted into a yes/no question that directly assessed the utterance properties that were affected by the retrieval pipeline [27]. The resulting questions are as follows:

- Comprehensible: Do you understand what the speaker wants to say?
- Reference: Does the utterance address its reference?
- Polarity: Does the utterance contradict the speaker's position?

The web interface allowed participants to rate utterances of the generated dialogues with regard to the mentioned categories. Before, an introduction to the study was provided containing a sample dialogue with an exemplary rating and respective explanations. Each participant had to evaluate three dialogues. Afterwards, a questionnaire on the comprehensibility of the categories had to be answered, as well as a text field for further comments on the study. The dialogues for the study were created from argument structures for seven different topics with two different configurations of the pipeline and one annotated structure. A total of 30 dialogues with equal amounts of utterances were created which were divided into 10 groups. Each group consisted of seven participants, such that a total of 70 participants was needed. The study was conducted via the crowdsourcing platform *Clickworker*[9].

3.2.2 Evaluation

A total of 10122 ratings for all three categories were collected with the study. Since the ratings of the participants diverged strongly, the three most agreeing users of each group were chosen for the evaluation. Results were found to be highly subjective - with regard to the direct comparison to each other and the impact of the pipeline configuration.

Considering the questionnaire responses in terms of the comprehensibility of each category, minor problems could be perceived. Unfortunately, 29 participants rated at least one category as incomprehensible. The majority of the ratings are neutral or positive, nevertheless, comprehension problems seemed to have been present, especially with regard to the polarity question. This was also confirmed in the test stage of the study. A few test participants were asked to test the study extensively. Although they were partially experts with knowledge of the argumentation domain, they found it difficult to evaluate the study correctly in all points. This strongly indicates the high complexity of the task even for experts. Furthermore, determining the coherence of arguments is a challenging and difficult task.

We assume that the participants of the "real" study without expert knowledge may have needed a supervised introduction to fully understand the categories and their usage. Nevertheless, it is not clear whether this would have led to a better result

[9] https://www.clickworker.com

than the crowdsourcing study. Since this study primarily consisted of a questionnaire setting rather than a real conversational interaction, the use of crowdsourcing seems still adequate.

3.3 Study 3: Data Acquisition for Indoor-Navigation Dialogue System

Further insights into crowdsourcing-related issues are provided by our data collection study [26]. The purpose of this study was to collect route descriptions using different communication styles for an indoor navigation dialogue system. Therefore, we conducted a crowdsourcing study, which showed a video with a route, which the study participants were then asked to describe.

3.3.1 Study Settings

We chose three routes in a public university building and filmed videos while walking along the former. All of the routes contained different points of interest and route elements like the cafeteria, an elevator, or stairs. The start and end points of the routes were linked such that all routes could be taken one after the other. In other words, route one ends where route two starts, and the end point of route two is the starting point of route three. In order to perform the study, a web application containing a few demographic questions (age, gender, and country of residence) and a video description tool were implemented. The tool consists of two parts: the video and the description box. Within the description box, the users had to enter their route descriptions while watching the video. While typing, the video stopped automatically and when pressing enter to save the input, which was then added to a list, the video continued. All saved items could still be edited or deleted. To prevent misuse, a minimum amount of items was required for the video to proceed in combination with a timer which secured that the video had been watched completely. The study was conducted via the crowdsourcing platform *Clickworker* allowing for English native speakers. We aimed to have 100 participants, 50 from the US and 50 from the UK (each with 25 females/males).

Prior to starting the study, three users from Germany checked and tested the system with regard to smooth operation, potential errors, comprehensibility of the task, etc.

3.3.2 Evaluation

Overall we had 97 participants, 53 from the UK (28 female, 25 male) and 44 participants from the US (18 male, 25 female, 1 prefer not to tell). Unfortunately, some of the participants misused our study tool and did not fill in the correct data. Moreover,

some participants did not finish their descriptions. The distribution of participants with respect to the quality of their data is described in the following:

- 74 participants entered correct data that could be used after minimal changes (e.g., merging some items due to unnecessarily created items).
- 8 participants provided correct information but did not finish the description. We assume that the participants did not see a need to continue the study after the timer expired. Nevertheless, this data could still be used as it is basically correct although incomplete.
- 15 participants entered invalid data (e.g., wrong route descriptions like left instead of right) or data that required a lot of post-processing and correction.
- 3 participants received the compensation for crowdworkers without even entering data. We assume that the code for the payment was passed by other users.

When evaluating the study data, we encountered two main challenges. The first is that there were many attempts to abuse the study setup, despite some precautions (e.g., time-based buttons to proceed according to the video length). Nevertheless, 18% of the study data had to be excluded (invalid/no data). In general, it was very time-consuming to filter out incorrect data records, since the entire route had to be traced in order to find an incorrect description. Second, all statements were written in chat style and not necessarily in text for spoken language, therefore the statements had to be partially expanded to fit our purpose. The adjustments of the statements in order to be able to use them for our dialog system were therefore indispensable. Nevertheless, as we aimed for a study with English native speakers a crowdsourcing study was suitable for this task. Moreover, it was crucial to access a large number of heterogeneous participants (only possible via crowdsourcing) in order to capture different communication styles. Thus, we already expected additional effort for data post-processing, since the extraction of the communication styles had to be done manually.

To conclude, we found that not only to summarize the different study results presented here and to derive respective general recommendations, but also to generally guarantee comparability and reproducibility of the results of crowdsourcing studies, a certain standardization, e.g. using checklists [30], would be very helpful.

4 Recommendations

Based upon our experiences and analysis of the study results described in Section 3 we propose the following recommendations for researchers using crowdsourcing to evaluate dialogue systems tailored particularly for complex tasks.

- Incorporate measures to ensure that the necessary instructions such as the introduction text and instructional videos can not be skipped, e.g., respective timers, test questions that can only be answered with careful consideration, etc. .
- Provide various options for assistance, which are easy to notice. Offer proactive assistance and more explanations, if the interaction is paused for a certain period

of time or the user obviously does not know what to do (performing actions that are not possible, repeating an action without taking the system's reaction into account).
- Double-staged study setting: Incorporating simulated test requests/actions in a first step, which emulate the possible actions the user can perform in the real interaction. Thus, it can be checked whether the participant understood the assignment and is only allowed to proceed to the real interaction if executed correctly.
- Record and consider demographic data and differences, especially with regard to argumentative dialogue systems and controversial topics.
- Include screening questions that gauge attention and language comprehension and can be used to filter out random ticking [24] and not conscientious participants.
- Usage of checklists [30] when reporting crowdstudies to ensure reproducibility and comparability of results (post self-control).
- Always use unique, one-time redeemable payment codes for participants compensation.

5 Conclusion and Future Work

In this work, we have described and analyzed three crowdsourcing studies of three different dialogue systems and have identified issues related to crowdsourcing. Based upon this we propose recommendations for the design of crowdsourcing studie with dialogue systems, in particular, if the latter involve more complex task. Furthermore, if the complexity exceeds a certain level, we recommend to perform user studies under supervision of experimenter on site who can assist accordingly.

In future work, we will incorporate the recommendations in Section 4 in our crowdsourcing study setup and analyze whether this solves or reduces the described issues, respectively. Furthermore, we plan a rerun of Study 1 (see *A.* in Section 3) in the lab to further investigate the influence of the study environment (lab vs. crowdsourcing) on the study evaluation results. Especially, we are interested in how the interaction, the perception of the systems and respective user ratings change under direct supervision.

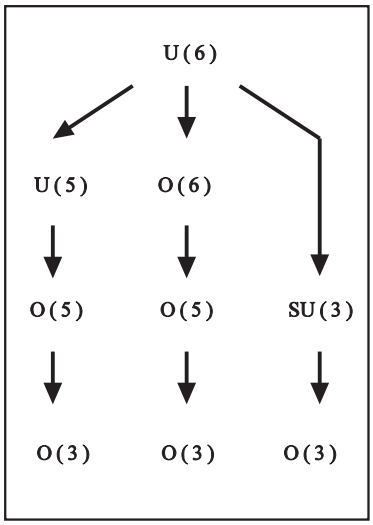
Acknowledgment

This work has been funded by the DFG within the project "BEA - Building Engaging Argumentation", Grant no. 313723125, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

References

1. E. Palogiannidi, "Using crowdsourcing for grammar induction with application to spoken dialogue systems", Thesis, Technical University of Crete, Electronic and Computer Engineering, June 2013.
2. A. Aker, M. El-Haj, M-D. Albakour and U. Kruschwitz, "Assessing crowdsourcing quality through objective tasks", Proc. of 8th Int. Conf. on Language Resources and Evaluation (LREC'12), pp. 1456–1461, 2012.
3. D. Nevo, J. Kotlarsky and S. Nevo, "New Capabilities: Can IT Service Providers Leverage Crowdsourcing?", Proc. 33rd Int. Conf. on Information Systems, Orlando, 2013.
4. J. Deriu, Á. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre and M. Cieliebak, "Survey on evaluation methods for dialogue systems. Artificial Intelligence Review 54, pp. 755–810, 2020.. DOI: 10.1007/s10462-020-09866-x.
5. A. H. Noel-Storr, P. Redmond, G. Lamé, E. G. Liberati, S. Kelly, L. Miller, G. Dooley, A. Paterson and J. Burt, "Crowdsourcing citation-screening in a mixed-studies systematic review: a feasibility study", BMC Med. Res. Meth. 21 (88), 2021. https://doi.org/10.1186/s12874-021-01271-4
6. O. Dušek, J. Novikova and V. Rieser, "Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge", Computer Speech & Language, 59, pp. 123–156, 2020.
7. E. Law, K. Z. Gajos, A. Wiggins, M. L. Gray and A. Williams, "Crowdsourcing as a Tool for Research: Implications of Uncertainty", Proc. of the 2017 ACM Conf. on Computer Supported Cooperative Work and Social Computing, ACM, pp. 1544–1561, New York, USA, 2017. DOI: 10.1145/2998181.2998197.
8. J. Jiang, B. An, Y. Jiang, D. Lin, Z. Bu, J. Cao and Z. Hao, "Understanding crowdsourcing systems from a multiagent perspective and approach" ACM Transactions on Autonomous and Adaptive Systems (TAAS), 13(2), pp.1–32, 2018.
9. H. Bassi, L. Misener and A. M. Johnson, "Crowdsourcing for Research: Perspectives From a Delphi Panel", SAGE Open, 10(4), 2020. DOI: 10.1177/2158244020980751.
10. R. Buettner, "A Systematic Literature Review of Crowdsourcing Research from a Human Resource Management Perspective", Proc. 48th Hawaii Int. Conf. on System Sciences, pp. 4609–4618, 2015. DOI: 10.1109/HICSS.2015.549.
11. K. B. Sheehan, "Crowdsourcing research: Data collection with Amazon's Mechanical Turk", Communication Monographs, 85(1), pp. 140–156, 2018. DOI: 10.1080/03637751.2017.1342043.
12. M. Hossain, I. Kauranen, "Crowdsourcing: A comprehensive literature review", Strategic Outsourcing: An Int. J., 8(1), pp. 2–22, 2015. DOI: 10.1108/SO-12-2014-0029.
13. P. Crequit, G. Mansouri, M. Benchoufi, A. Vivot, P. Ravaud, "Mapping of crowdsourcing in health: Systematic review", J. of Medical Internet Research, 20 (5), e187, 2018. DOI: 10.2196/jmir.9330.
14. R: Manuvinakurike, M. Paetzel and D. DeVault, "Reducing the cost of dialogue system training and evaluation with online, crowd-sourced dialogue data collection", Proc.s of SEMDIAL 2105, pp. 113–121, 2015.
15. J. Thebault-Spieker, L. G. Terveen and B. Hecht, "Avoiding the South Side and the Suburbs: The Geography of Mobile Crowdsourcing Markets", Proc. of the 18th ACM Conf. on Computer Supported Cooperative Work & Social Computing, ACM, New York, USA, pp. 265–275, 2015. DOI: 10.1145/2675133.2675278.
16. ITU-T, "Subjective Quality Evaluation of Telephone Services Based on Spoken dialogue Systems", ITU-T Rec. P.851, International Telecommunication Union, Geneva, 2003.
17. S. Möller, "Quality of Telephone-Based Spoken dialogue Systems", Kluwer, Boston, 2005.
18. M.-C. Yuen, I. King and K. -S. Leung, "A Survey of Crowdsourcing Systems", Proc. 3rd Int. Conf. on Privacy, Security, Risk and Trust, 3rd Int. Conf. on Social Computing, pp. 766–773, IEEE, 2011. DOI: 10.1109/PASSAT/SocialCom.2011.203.

19. K.-P. Engelbrecht, “Estimating Spoken dialogue System Quality with User Models”, Springer, Berlin, Heidelberg, 2011.
20. S. Hillmann, “Simulation-Based Usability Evaluation of Spoken and Multimodal dialogue Systems”, Springer, Cham, 2017.
21. S. Mehri and M. Eskenazi, “USR: An Unsupervised and Reference Free Evaluation Metric for dialogue Generation”, Proc. of 58th Annual Meeting of the ACL, pp. 681–707, ACL, 2020. DOI: 10.18653/v1/2020.acl-main.64.
22. S. Mehri and M. Eskenazi. “Unsupervised Evaluation of Interactive dialogue with dialoguePT”, Proc. of SIGDIAL 2020, pp. 225–235, ACL, 2020.
23. M. Sabou, K. Bontcheva, L. Derczynski and A. Scharl. “Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines”, Proc. 9th Int. Conf. on Language Resources and Evaluation (LREC’14), pp. 859–866, ELRA, Reykjavik, Iceland, 2014.
24. J. K. Goodman, C. E. Cryder A. and Cheema, “Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples”. J. of Behavioral Decision Making, 26 (3), pp. 213–224, 2013.
25. F. Jurcıcek, S. Keizer, M. Gašic, F. Mairesse, B. Thomson, K. Yu and S. Young, “Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk”, Proc. Interspeech, ISCA, 2011.
26. J. Miehle, I. Feustel, W. Minker and S. Ultes, “A Script Knowledge Based dialogue System for Indoor Navigation”, In Conversational dialogue Systems for the Next Decade, pp. 379–385, Springer, Singapore, 2012.
27. N. Rach, C. Schindler, I. Feustel, J. Daxenberger, W. Minker, and S. Ultes. From Argument Search to Argumentative dialogue: A Topic-independent Approach to Argument Acquisition for dialogue Systems. In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 368-379), (2021, July).
28. Z. Liao, X. Xu, X. Fan, Y. Zhang and S. Yu, “GRBMC: An effective crowdsourcing recommendation for workers groups”, Expert Systems with Applications, 179, 115039, 2021.
29. S. Buchholz and J. Latorre, “Crowdsourcing preference tests, and how to detect cheating”, Proc. Interspeech 2011, ISCA, Florence, Italy, 2011.
30. Jorge Ramírez, Burcu Sayin, Marcos Baez, Fabio Casati, Luca Cernuzzi, Boualem Benatallah, and Gianluca Demartini. 2021. On the State of Reporting in Crowdsourcing Experiments and a Checklist to Aid Current Practices. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 387 (October 2021), 34 pages. DOI:https://doi.org/10.1145/3479531
31. Möller, S., 2003. Subjective quality evaluation of telephone services based on spoken dialogue systems. ITU-T Recommendation, p.851.
32. Annalena Aicher, Nadine Gerstenlauer, Wolfgang Minker, and Stefan Ultes, 2022. User Interest Modelling in Argumentative Dialogue Systems. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, pages 127–136 (ELRA).
33. Li, Yu, et al. ”Legoeval: An open-source toolkit for dialogue system evaluation via crowdsourcing.” arXiv preprint arXiv:2105.01992 (2021).
34. Huynh, Jessica, et al. ”DialCrowd 2.0: A Quality-Focused Dialog System Crowdsourcing Toolkit.” Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022.
35. Jurcicek, Filip, et al. ”Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk.” Proceedings of INTERSPEECH. Vol. 11. 2011.
36. Gosling, S. D., Rentfrow, P. J., and Swann Jr., W. B. (2003). A very brief measure of the big-five personality domains. Journal of Research in Personality, 37, p. 504–528.
37. Karen Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Last Words: Amazon Mechanical Turk: Gold Mine or CoalMine? Computational Linguistics, 37(2):413–420.
38. Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3758–3769, Online. Association for Computational Linguistics


U(6)
U(5)
O(6)
O(5)
O(5)
SU(3)
O(3)
O(3)
O(3)