Artificial Intelligence and Social Computing, Vol. 72, 2023, 77–88
https://doi.org/10.54941/ahfe1003276

Towards a Proper Evaluation of Automated Conversational Systems

Abraham Sanders[1], Mara Schwartz[1], Albert Chang[1], Shannon Briggs[1], Jonas Braasch[1], Dakuo Wang[2], Mei Si[1], and Tomek Strzalkowski[1]

[1]Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[2]IBM Research, USA

ABSTRACT

Efficient evaluation of dialogue agents is a major problem in conversational AI, with current research still relying largely on human studies for method validation. Recently, there has been a trend toward the use of automatic self-play and bot-bot evaluation as an approximation for human ratings of conversational systems. Such methods promise to alleviate the time and financial costs associated with human evaluation, and current proposed methods show moderate to strong correlation with human judgements. In this study, we further investigate the fitness of end-to-end self-play and bot-bot interaction for dialogue system evaluation. Specifically, we perform a human study to confirm self-play evaluations of a recently proposed agent that implements a GPT-2 based response generator on the Persuasion For Good charity solicitation task. This agent leverages Progression Function (PF) models to predict the evolving acceptability of an ongoing dialogue and uses dialogue rollouts to proactively simulate how candidate responses may impact the future success of the conversation. The agent was evaluated in an automatic self-play setting, using automatic metrics to estimate sentiment and intent to donate in each simulated dialogue. This evaluation indicated that sentiment and intent to donate were higher ($p < 0.05$) across dialogues involving the progression-aware agents with rollouts, compared to a baseline agent with no rollout-based planning mechanism. To validate the use of self-play in this setting, we follow up by conducting a human evaluation of this same agent on a range of factors including convincingness, aggression, competence, confidence, friendliness, and task utility on the same Persuasion For Good solicitation task. Results show that human users agree with previously reported automatic self-play results with respect to agent sentiment, specifically showing improvement in friendliness and confidence in the experimental condition; however, we also discover that for the same agent, humans reported a lower desire to use it in the future compared to the baseline. We perform a qualitative sentiment analysis of participant feedback to explore possible reasons for this, and discuss implications for self-play and bot-bot interaction as a general framework for evaluating conversational systems.

Keywords: Dialogue system evaluation, Dialogue agent, Dialogue planning, Conversational artificial intelligence, Natural Language Processing

INTRODUCTION

Efficient evaluation of dialogue agents is a major problem in conversational AI, with current research still relying largely on human studies for method validation. Recently, there has been a trend toward the use of automatic **self-play** and **bot-bot** evaluation as an approximation for human ratings of conversational systems (shown in Figure 1). Such methods promise to alleviate the time and financial costs associated with interactive human-bot evaluation, and current proposed methods

(see Related Work section) show moderate to strong correlation with human judgements. Our goal in this study is to further confirm the fitness of the general self-play framework for dialogue system evaluation, focusing on both social and goal-oriented aspects. To do so, we investigate if humans interacting with the charity solicitation agent from Sanders et al. (2022) report findings that agree with their earlier self-play results obtained with no human involvement. This agent implements a DialoGPT (Zhang et al., 2020) response generator fine-tuned on the Persuasion For Good (Wang et al., 2019) charity solicitation task dataset, where the agent's goal is to convince its user to make a small donation to an international charity. Sanders et al. evaluate their agent using an end-to-end automatic self-play framework where the agent models its own role and also that of the user while completing dialogues without human participation. Under this framework, the agent is reported to achieve better performance in both sentiment and task-completion when equipped with a progression function (PF) model that allows it to plan ahead via dialogue rollouts (Lewis et al., 2017) when considering different response options.

In this study, we aim to determine if human evaluators also conclude that the rollout-equipped agent is more successful at the solicitation task than the baseline. We conduct a between-group human-subjects experiment in which participants interact with either the rollout-equipped agent or the

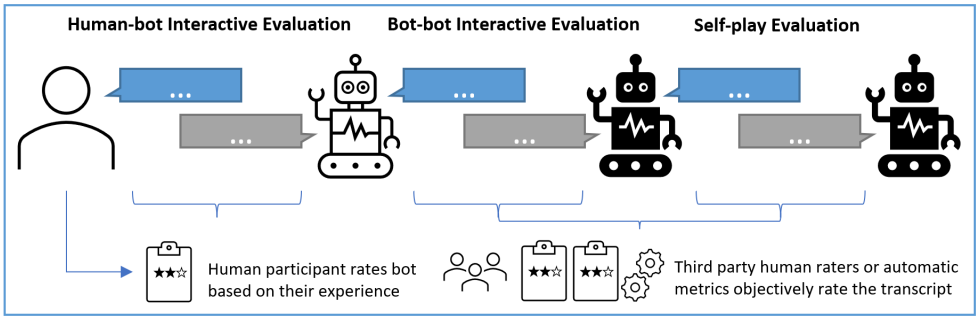

Figure 1: Types of interactive evaluation for dialogue systems. **Left**: in human-bot evaluation, a human participant converses with a bot and then rates their experience by a set of criteria (e.g., fluency, knowledgeability, etc.); **Middle:** in bot-bot evaluation, two bots converse with each other and the transcript is rated by third party humans or automatic metrics that approximate the evaluation criteria; **Right**: self-play evaluation is a special case of bot-bot evaluation where the bot converses with itself instead of another bot.

baseline agent in the original configurations previously used for self-play. Each participant chats with their respective agent, which attempts to solicit a donation to charity from them. The conversation is then followed by a survey in which the participant rates the agent on a variety of social and goal-related competencies (e.g., friendliness, confidence, convincingness, imparted desire to donate). We perform statistical analysis to establish any differences between the groups and discuss how the results relate to the earlier self-play conclusions.

RELATED WORK

Self-play and bot-bot interaction has been previously explored in dialogue settings for reinforcement learning (e.g., Li et al., 2016; Lewis et al., 2017) and evaluation, which is what we will discuss further here. Both tasks rely on self-play and bot-bot interaction approximating human-bot interaction well.

Most relevant to our study are those works that use self-play or bot-bot interaction for evaluation. These approaches fall into two categories: end-to-end and hybrid. End-to-end methods (Ghandeharioun et al., 2019; Deriu & Cieliebak, 2019; Sanders et al., 2022) combine self-play or bot-bot interaction with learned or algorithmic metrics to provide a completely automatic evaluation procedure, while hybrid approaches (Li et al., 2019; Deriu et al., 2020) use them to generate conversations only, which human raters must later evaluate.

We are particularly interested in the viability of end-to-end approaches since these are the most accessible and scalable. Ghandeharioun et al. (2019) combine a set of automatic sentiment, semantic, and engagement metrics into a single learned composite metric, and use it to measure six open-domain agents in self-play. They report a strong correlation ($r > 0.7$, $p < 0.05$) with interactive human evaluation of the same agents across five Likert-scale questions measuring quality, fluency, diversity, contingency, and empathy. Similarly, Deriu & Cieliebak (2019) propose the AutoJudge method which learns a metric to approximate human response quality judgement, which is then applied to automatically rank five open-domain agents in self-play. They report moderate correlation ($r = 0.573$) between AutoJudge ratings and human ratings of the same self-play dialogues collected by crowdsourcing. Sanders et al. (2022) also train a learned metric, the Progression Function (PF), to approximate the "acceptability score" of a dialogue. This is a composite metric which combines social and goal-oriented metrics by the degree to which they correlate with task success. Two agents are ranked in self-play (DialoGPT with and without a PF-based lookahead planning mechanism). They report a moderate correlation ($r = 0.48$) between PF scoring and human judgement at the utterance level on ground-truth dialogues and report a positive effect of the planning mechanism on sentiment and goal completion as measured by self-play, but do not report any interactive human-bot evaluation.

This last part is what we address in this study to further confirm the integrity of self-play for general dialogue system evaluation. Unlike the open-domain chit-chat tasks seen in the other self-play and bot-bot evaluations

referenced here, success at the Persuasion For Good task requires a balance of social skills (e.g., friendliness) and goal-awareness (e.g., persuading the user to donate), and it remains unseen how effectively an automated interactive evaluation can capture (and enforce) the need for such a trade-off of conversational skills.

EVALUATION

To evaluate the effectiveness of the agents in the social and goal-oriented aspects of the Persuasion For Good charity solicitation task, we designed an empirical study to measure users' opinions in response to interacting with the chatbot.

EXPERIMENT DESIGN

We used a between-group design with two groups. Both groups conversed with the DialoGPT agent fine-tuned on Persuasion For Good. Following the self-play setup in Sanders et al. (2022), the experimental group conversed with a version of this agent equipped with the RoBERTa-large-adapted progression model and the 2x2x3 rollout configuration, where the look-ahead planning mechanism considers 2 response candidates, 2 rollouts per candidate, and 3 utterances per rollout (shown in Figure 2). The control group conversed with the same agent minus the progression model and rollouts, which is just vanilla DialoGPT fine-tuned on Persuasion For Good. The subjects did not know which group they were in, and the experimental procedures were exactly the same for them.

Subjects

We recruited 30 subjects from the student population at Rensselaer Polytechnic Institute, including both undergraduate and graduate students. 15 were assigned to the experimental group and 15 to the control group.

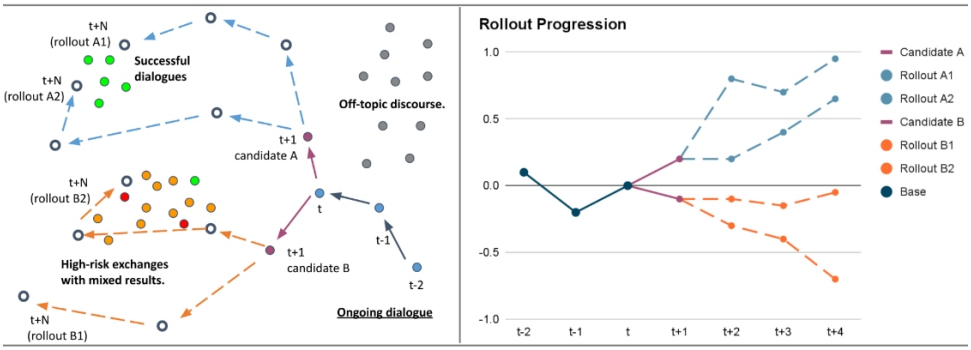

Figure 2: The look-ahead planning mechanism based on dialogue rollouts, from Sanders et al. (2022). **Left:** At each turn the agent generates N simulated turns into the future s times, for each of c candidate responses. Shown here, $c = 2$, $s = 2$, and $N = 3$, which is referred to as a 2x2x3 rollout configuration. **Right:** The candidate leading to the best average simulated future outcome as ranked by a progression (PF) model is selected.

All participants were between 18–30 years old except for one in the experimental group who was between 30–40 years old. Twelve subjects in the control group identified as male and two as female, while nine subjects in the experimental group identified as male, three as female, two as other, and one preferred not to specify. All participants identified as digitally literate on a 5-point scale, with average digital literacy self-rating of 4.733 ($\pm$ 0.458) for the control group and 4.467 ($\pm$ 0.640) for the experimental group. All participants report having past experience with chatbots, also on a 5-point scale, with averages of self-rated experience level 3.867 ($\pm$ 0.915) for the control group and 4.333 ($\pm$ 0.488) for the treatment group. Finally, the most frequent type of past chatbot experience in both groups were personal assistants (e.g., Siri, Alexa) and online customer care chatbots. The average self-reported level of satisfaction with past chatbots, also on the same scale, is 3.467 ($\pm$ 0.834) for the control group and 3.133 ($\pm$ 0.915) for the experimental group. Participants were compensated with a $5 Amazon gift card on completing the study.

METHODS

The study was conducted over the internet. The participants were given a link to a questionnaire without any requirement to login. The first part of the form contained a description of the study and a pre-questionnaire. Participants needed to check a box to indicate their consent to participate and finish the pre-questionnaire to be ready to proceed with the study. The full questionnaires are provided in Tables 4 & 5 in the appendix.

Participants were then directed to a link that would bring them to the chatbot. They were instructed to talk with the chatbot via text for a minimum of 5 turns. A minimum of 5 and maximum of 15 turns were enforced, and no other information about the nature of the task was provided. During the chat, responses were generated by each agent using the same decoding hyperparameters used for self-play in Sanders et al. (2022) (beam sampling with num_beams = 6, top_k = 50, top_p = 0.95, and temperature=$1.5 + 0.002 \times T$, where T is the length of the dialogue history). Participants were free to leave the chat at any time. If a participant did so, they were able to resume their conversation later. When finished interacting with the chatbot, participants were directed back to the survey section to answer additional questions in a post-questionnaire, after which their session would end.

RESULTS AND DISCUSSION

Our experiment yielded 30 vectors of 19 Likert-scale ratings (1-5; 0 indicating non-response) from the post-questionnaire - 15 for the control group and 15 for the experimental group.

Outlier Detection

Due to the small sample size of our study, we wanted to ensure that our results were robust to the presence of outliers in the responses. Instead of trimming

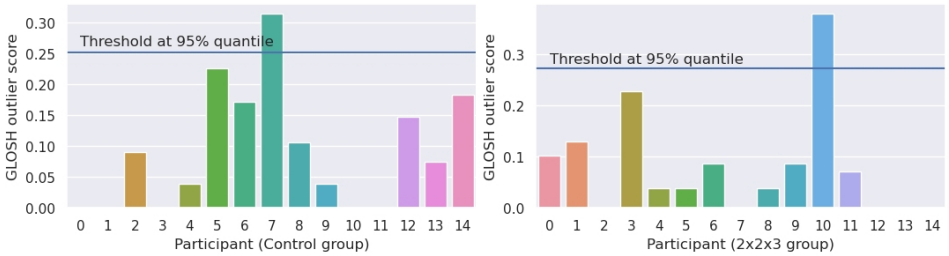

Figure 3: GLOSH outlier score distribution for the control group with no rollouts **(Left)** and experimental group with the 2x2x3 rollout configuration **(Right)**.

outlying responses at the question level, we applied GLOSH (Campello et al., 2015) a multivariate outlier detection algorithm based on HDBSCAN clustering (Campello et al., 2013; McInnes and Healy, 2017) to identify entire questionnaires that were likely to contain outlying responses across all questions. By using this approach, we allow responses at rare extremes for any individual question to be included in our analysis if most other responses in the same questionnaire lie in the typical range for the group. As shown in Figure 3, we compute GLOSH outlier likelihood scores for the 15 post-questionnaire vectors in the control and experimental groups independently and eliminate any questionnaire that falls above the 95% quantile of scores for its group. This process eliminates $N = 1$ subject from the control group and $N = 1$ subject from the experimental group, leaving $N = 14$ remaining subjects in each group for our analysis.

Statistical Analysis

Of the 19 Likert-scale post-survey questions, the first two relate to **goal attainment**, measuring whether the user might donate in the future and felt increased desire to donate as a result of the conversation. The next five questions relate to social aspects of the conversation relating to the **persuader role**, measuring whether the user felt the agent was worthy of using again, and whether it was convincing, pressuring, dishonest, and able to relate to the user's moral beliefs. Then, the following seven questions relate to **general social skills**, measuring whether the user felt the agent was competent, confident, efficient, intelligent, friendly, well intentioned, and trustworthy. Finally, the last five questions relate to **overall conversational quality**, measuring whether the user felt the agent gave fluent, consistent, knowledgeable responses that were high-quality, and whether the user felt the agent was adequately responsive. The response distributions for both the control and experimental groups are shown in Figure 4.

To measure differences in response distribution between groups for each question, we collapse responses from a five-point Likert scale to a three-point scale and perform a one-sided unpaired t-test for each question. This tests whether the experimental group mean response is greater than or less than that of the control group, where the direction depends on whether the question is a "higher-is-better" or a "lower-is-better" metric. Before running each

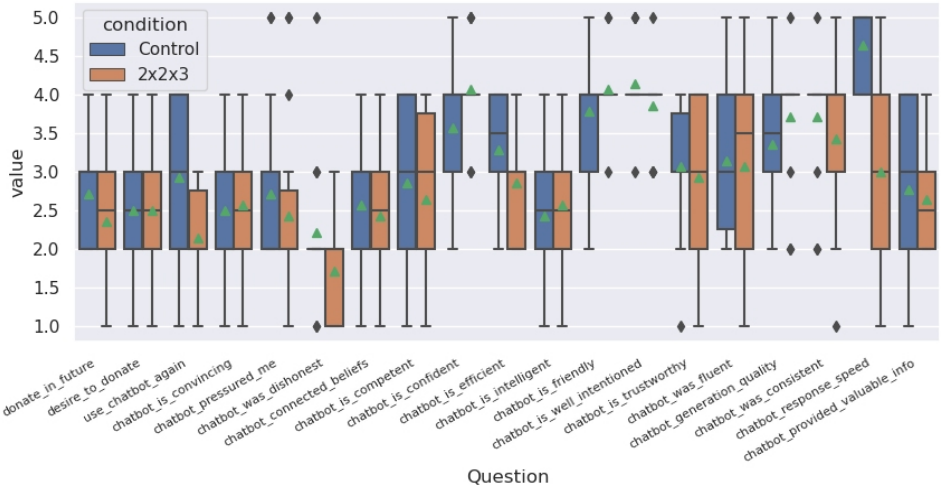

Figure 4: Response distributions for the 19 Likert-scale post-survey questions in the control group with no rollouts **(Blue)** and the experimental group with the 2x2x3 rollout configuration **(Orange).** Diamonds indicate outlying values (beyond the range of $1.5 \times IQR$) and flat boxes indicate where most responses for a question within the group are identical.

t-test, we use an F-test to check for equal variances between groups. For questions where the F-statistic p-value < 0.05 we use the unequal variances t-test and use the equal variances t-test for all others. Results of the t-tests are shown in Table 1, where we see differences for the questions relating to whether the user felt the agent was: (1) worthy of using again, (2) confident, (3) friendly, and (4) responsive. These differences are all significant at $p < 0.05$ except for (2) confident, which is on the threshold at $p = 0.05$.

We observe that users in the experimental group with the 2x2x3 rollout configuration found the agent to be friendlier and more confident, agreeing with the previously reported self-play finding that rollouts led the agent to select responses with higher sentiment. We also see that users in that group found the agent to be less responsive, which is expected due to the additional computation involved in rollouts (5-10 seconds per response). However, we also observe a seemingly contradictory result: users in the experimental group reported less desire to use the agent again, despite rating the agent to be friendlier and more confident.

Table 1. Questions with difference in mean below (*) or at (†) the 5% significance threshold. Test direction (>) indicates experimental group mean is greater than control group mean.

Question	t-test p-val. (direction)
Would you use the chatbot again?	0.015 (<) *
To what extent do you think the chatbot is confident?	0.050 (>) †
To what extent do you think the chatbot is friendly?	0.036 (>) *
Had good response speed?	< 0.01 (<) *

Participant Feedback Analysis

To help understand why users in the experimental group reported less desire to use the agent despite finding it friendlier and more confident, we analyze the opinion feedback that each participant was asked to provide in the post-questionnaire. We look at: (a) what feedback corresponds with the lowest ratings for the question "Would you use the chatbot again?" in each group, and (b) what feedback has the lowest sentiment in each group. To measure sentiment we use a publicly available RoBERTa (Liu et al., 2019) model fine-tuned on the sentiment classification task of the TweetEval (Barbieri et al., 2020) benchmark. Following Sanders et al. (2022) we combine sentiment class probabilities (negative, neutral, positive) for each opinion to a continuous value in the range $[-1, 1]$. We find that the average sentiment of experimental group opinions is lower and varies less than those of the control group (-0.477 ± 0.324 and -0.358 ± 0.517 respectively), which aligns with the lower reported desire to use the agent again. Table 2 shows the opinions in each group with the lowest rating (a score of 1) for the question "Would you use the chatbot again?" and Table 3 shows those for each group with the lowest sentiment.

Table 2. Opinions provided by participants who scored 1/5 for the question "Would you use the chatbot again?" in each group.

Group	Opinion
Control	It really doesn't back down from a "no"
2x2x3	Most of the information was superficial, which makes sense for imitating conversation. I think it tends to repeat itself, and its grammar needs improvement.
2x2x3	The chatbot talked in circles a lot, and its answers to my questions were vaguer than expected.

Table 3. The three opinions with the lowest sentiment scores in each group, along with the score given for the question "Would you use the chatbot again?" (shown as **UA**).

Group	Opinion	sent	UA
Control	it was incredibly shallow.	−0.958	2/5
Control	It's transition into pitching me the charity was forced and it often refused to answer my questions, causing me to lose interest in what it was saying.	−0.867	2/5
Control	Respond not well when I answer "no" to "have you donated before"	−0.806	4/5
2x2x3	Annoying	−0.854	2/5
2x2x3	It gave short responses to what I said but didn't have much to contribute to the conversation, and tried to end the exchange with a "have a good one!" only a few turns after we started talking. If it were a person I would have been kind of annoyed.	−0.803	2/5
2x2x3	I think I have the feeling that the chatbot was trying to force the conversation into the direction of charity and donation, and it was awkward sometimes. Moreover, what says by the chatbot was inconsistent among different turns.	−0.744	2/5

We observe that the three most common themes in these opinions corresponding to the lowest sentiments and desire to use the agent again are: (1) generation quality issues such as repetitive and inconsistent responses (shown in purple); (2) giving unsatisfactory or vague answers to questions about the charity (shown in blue); and (3) being too aggressive in soliciting for donations (shown in red). These themes appear in the opinions shown for both the control and experimental groups.

Discussion & Future Direction

Most of the concerns around generation quality and vagueness are not surprising and can be attributed to use of a small language model (DialoGPT). However, the concerns around the agent being aggressive in solicitation are interesting - in a negotiation dialogue task, Lewis et al. (2017) reported that goal-directed agents equipped with rollouts and/or reinforcement learning negotiate harder and are less likely to settle on a deal, causing human participants in their experiments to walk away without an agreement more often than they did when negotiating with baseline agents. However, they observed the opposite effect when their goal-directed agents engaged in self-play: here, negotiations between goal-directed and baseline ablations without rollouts or reinforcement learning had ***higher*** agreement ratings than negotiations between two instances of the baseline agent.

Similarly, it may be possible in our study that the experimental agent tends to solicit more aggressively than agents in the control group, frustrating participants and turning them off from future use. We see evidence for this trend in the opinion analysis, but failed to find a statistically significant difference between groups in the questions relating to feeling pressured or feeling that the agent was well intentioned, dishonest, or trustworthy. To investigate if this is the case, a larger or perhaps more targeted study could be done in which we ask participants more specific questions about their perception of aggressive tactics.

CONCLUSION

We conducted an interactive human-bot study of a charity solicitation agent and investigated how well the results agree with previously published self-play evaluations of the same agent. Our study found increases in ratings of friendliness and confidence in the experimental group that interacted with the goal-directed version of the agent, aligning with increases in sentiment reported for the same agent in self-play. However, we also found that the same human participants interacting with the experimental agent reported lower desire to use the agent in the future, suggesting a misalignment between the reactions of human users and self-play agents when playing the user role. We followed up with a qualitative sentiment analysis of negative participant opinions, discovering that over-aggression could be a potential factor influencing poor human experiences interacting with the agent, which has been shown in prior literature to have the opposite effect on the success of autonomous agents engaging in self-play. Overall, our results suggest that before applying self-play or bot-bot interaction to evaluate a dialogue system, it is important

to verify alignment between agent and human reaction to violation of behavioural expectation specific to the task and domain at hand. Our analysis code is available at https://github.rpi.edu/LACAI/dialogue-human-eval.

APPENDIX

Table 4. Questions in the pre-questionnaire.

Question	Question Type
What is your age?	Categorical
What is your gender?	Categorical
Which of the following describes your race/ethnicity?	Categorical
What is the highest level of school you have completed or the highest degree you have received?	Categorical
I have a good understanding about Computers/Internet	Likert scale (5 point)
I have experience using an automated chatbot.	Likert scale (5 point)
My previous experience with a chatbot has been good.	Likert scale (5 point)
Where have you encountered a chatbot before?	Categorical

Table 5. Likert-scale questions in the post-questionnaire (5 point).

Question	Question	
Are you going to donate to this charity in the future?	The chatbot is competent.	The chatbot was fluent.
My desire to donate has grown as a result of my dialogue with the chatbot.	The chatbot is confident.	The chatbot generated good quality of text.
Would you use the chatbot again?	The chatbot is efficient.	The chatbot was consistent.
The chatbot is convincing.	The chatbot is intelligent.	The chatbot had good response speed.
The chatbot tried to pressure me.	The chatbot is friendly.	The chatbot provided valuable information.
The chatbot was dishonest.	The chatbot is well-intentioned.	
The chatbot connected to my beliefs about fundamental right and wrong.	The chatbot is trustworthy.	

ACKNOWLEDGMENT

This paper is based upon work supported in part by the United States Air Force under Contract No. FA8750-21-C-0075 and in part by the IBM Corporation under the Artificial Intelligence Research Collaboration Agreement No. W1771793 between IBM and Rensselaer. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of USAF or IBM Corporation.

REFERENCES

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., Neves, L., 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification, in: Findings of the Assoc. for Comp. Linguistics: EMNLP 2020. Assoc. for Comp. Linguistics, Online, pp. 1644–1650.

Campello, R. J. G. B., Moulavi, D., Sander, J., 2013. Density-Based Clustering Based on Hierarchical Density Estimates, in: Pei, J., Tseng, V. S., Cao, L., Motoda, H., Xu, G. (Eds.), Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 160–172.

Campello, R. J. G. B., Moulavi, D., Zimek, A., Sander, J., 2015. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. ACM Trans. Knowl. Discov. Data 10, 1–51.

Deriu, J., Tuggener, D., von Däniken, P., Campos, J. A., Rodrigo, A., Belkacem, T., Soroa, A., Agirre, E., Cieliebak, M., 2020. Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Assoc. for Comp. Linguistics, Online, pp. 3971–3984.

Deriu, J. M., Cieliebak, M., 2019. Towards a Metric for Automated Conversational Dialogue System Evaluation and Improvement, in: Proceedings of the 12th International Conference on Natural Language Generation. Assoc. for Comp. Linguistics, Tokyo, Japan, pp. 432–437.

Ghandeharioun, A., Shen, J. H., Jaques, N., Ferguson, C., Jones, N., Lapedriza, A., Picard, R., 2019. Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems, in: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d', Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc.

Lewis, M., Yarats, D., Dauphin, Y., Parikh, D., Batra, D., 2017. Deal or No Deal? End-to-End Learning of Negotiation Dialogues, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Assoc. for Comp. Linguistics, Copenhagen, Denmark, pp. 2443–2453.

Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M., Gao, J., 2016. Deep Reinforcement Learning for Dialogue Generation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Assoc. for Comp. Linguistics, Austin, Texas, pp. 1192–1202.

Li, M., Weston, J., Roller, S., 2019. ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons, in: Advances in Neural Information Processing Systems, Conversational AI Workshop.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

McInnes, L., Healy, J., 2017. Accelerated Hierarchical Density Based Clustering, in: 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, New Orleans, LA, pp. 33–42.

Sanders, A., Strzalkowski, T., Si, M., Chang, A., Dey, D., Braasch, J., Wang, D., 2022. Towards a Progression-Aware Autonomous Dialogue Agent, in: Proceedings of the 2022 Conference of the North American Chapter of the Assoc. for Comp. Linguistics: Human Language Technologies. Assoc. for Comp. Linguistics, Seattle, United States, pp. 1194–1212.

Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., Yu, Z., 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good, in: Proceedings of the 57th Annual Meeting of the Assoc. for Comp. Linguistics. Florence, Italy, pp. 5635–5649.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., Dolan, B., 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation, in: Proceedings of the 58th Annual Meeting of the Assoc. for Comp. Linguistics: System Demonstrations. Online, pp. 270–278.