Interactive Evaluation of Dialog Track at DSTC9

Shikib Mehri[1], Yulan Feng[1], Carla Gordon[2],
Seyed Hossein Alavi[2], David Traum[2], Maxine Eskenazi[1]

[1]Language Technologies Institute, Carnegie Mellon University
[2]Institute for Creative Technologies, University of Southern California
amehri@andrew.cmu.edu

Abstract

The ultimate goal of dialog research is to develop systems that can be effectively used in interactive settings by real users. To this end, we introduced the *Interactive Evaluation of Dialog Track* at the 9th Dialog System Technology Challenge. This track consisted of two sub-tasks. The first sub-task involved building knowledge-grounded response generation models. The second sub-task aimed to extend dialog models beyond static datasets by assessing them in an interactive setting with real users. Our track challenges participants to develop strong response generation models and explore strategies that extend them to back-and-forth interactions with real users. The progression from static corpora to interactive evaluation introduces unique challenges and facilitates a more thorough assessment of open-domain dialog systems. This paper provides an overview of the track, including the methodology and results. Furthermore, it provides insights into how to best evaluate open-domain dialog models.

Introduction

A long-standing challenge in computer science is the development of algorithms that can interact with humans in natural language (Turing 1950). Ultimately, the goal of dialog research is to create systems that can engage in back-and-forth interactions with real users (Eskenazi et al. 2019). However, the majority of research is performed on static datasets. For example, the task of response generation is typically done by producing a response for a static dialog context (Vinyals and Le 2015). By reducing dialog to response generation, static evaluation neglects multiple important challenges of dialog. In contrast, interactive evaluation allows several valuable properties of dialog to be measured, including: consistency, topic depth, adaptation, error recovery and user-centric development. Mehri and Eskenazi (2020a) found that state-of-the-art dialog models perform on-par with humans on response generation, but they fall short when considering an entire dialog. To promote interactive evaluation of dialog, the *Interactive Evaluation of Dialog Track* of the 9th Dialog System Technology Challenge (Gunasekara et al. 2020) challenged participants to build models for open-domain interaction with real users.

This track consists of two sub-tasks: (1) static evaluation and (2) interactive evaluation. The goal of the first sub-task is to develop knowledge-grounded response generation models which are then evaluated in a static manner using the Topical-Chat corpus (Gopalakrishnan et al. 2019). The second sub-task challenges participants to extend response generation models to effectively converse with real users through the DialPort portal (Zhao, Lee, and Eskenazi 2016). Through these two sub-tasks, the track challenges participants to first develop strong response generation models and then to explore strategies for extending them to interactive settings.

In the following sections, we describe the methodology and results for both sub-tasks. We then present insights into methods of best evaluating open-domain dialog models.

Related Work

Interactive Evaluation

As dialog models improve, it is imperative that they are evaluated in interactive settings with real users. Much open-domain dialog research focuses on the task of response generation, which is done on static corpora (Vinyals and Le 2015). Large pre-trained dialog models have shown impressive performance on the task of response generation, with results on par with human utterances (Zhang et al. 2019b). Recently, several state-of-the-art open-domain dialog models have been evaluated in interactive settings (Adiwardana et al. 2020; Roller et al. 2020). Mehri and Eskenazi (2020a) show that while such models excel at generating responses, they underperform in back-and-forth interactions.

The Alexa Prize challenge (Ram et al. 2018; Khatri et al. 2018) allows university teams to build socialbots that are assessed in interactive settings with Alexa users. In contrast to the Alexa Prize challenge, our track is accessible to the broader research community. Furthermore, the Alexa Prize challenge relies on speech input from the user, which may, at present, result in speech recognition errors. In contrast, our track uses a web interface with text-only input.

Open-Domain Dialog

Recent work on large-scale pre-training has resulted in significant advances in open-domain dialog (Zhang et al. 2019b; Adiwardana et al. 2020; Roller et al. 2020; Bao et al. 2020). DialoGPT (Zhang et al. 2019b) fine-tuned GPT-2 (Radford

et al. 2019) on dialogs from Reddit and reported human level response generation capabilities. Meena (Adiwardana et al. 2020) trains a larger evolved Transformer model on social media data and attains strong performance in interactive settings. Blender (Roller et al. 2020) uses a retrieve and refine approach, in combination with a thorough exploration of generation strategies and reports improved performance on interactive evaluation relative to Meena. PLATO-2 (Bao et al. 2020) uses a two-step curriculum learning process where they perform coarse-grained training on one-to-one response generation followed by fine-grained fine-tuning with one-to-many dialog data. PLATO-2 reports improvements in both static and interactive evaluation.

Automatic Dialog Evaluation

Though we perform on-going human evaluation throughout the challenge, it is nonetheless important to have meaningful automatic metrics since they are often used for intermediate evaluation when developing a dialog model. If participants iterate on their models with subpar automatic metrics, they may decrease performance on human evaluation (Dinan et al. 2019).

Standard metrics such as BLEU (Papineni et al. 2002) and METEOR (Banerjee and Lavie 2005) have been shown to perform poorly for evaluating dialog (Liu et al. 2016; Gupta et al. 2019). This is in part due to the one-to-many problem in dialog: there are multiple valid responses for a particular dialog context. As such, comparing to a reference response is ineffective.

There have been efforts in developing automatic dialog evaluation metrics that correlate better with human judgement. Lowe et al. (2017) train ADEM on human annotations to produce a quality score for a generated response given on a dialog context and a reference response. Venkatesh et al. (2018) present a framework for evaluating Alexa Prize dialogs, by training on user annotations. Mehri and Eskenazi (2020b) present USR, which relies on pre-trained language models and self-supervised training objectives to approximate the multiple qualities of dialog (e.g., interesting, relevant) without comparing to a reference response. Sinha et al. (2020) introduce MaUdE which uses pre-trained language models to analyze the temporal transitions between utterances in a dialog, to evaluate without comparing to a reference response. Mehri and Eskenazi (2020a) present FED which presents a framework for predicting eighteen different qualities of dialog using off-the-shelf pre-trained language models.

Sub-task 1: Static Evaluation

The objective of the first sub-task is to develop response generation models for the Topical-Chat corpus (Gopalakrishnan et al. 2019). Over the duration of the challenge, participants submitted generated responses for the *frequent* validation set of the Topical-Chat corpus. This set consists of topics that frequently appear in the training data. For the final submissions, the *frequent* test set was used. Throughout the challenge, submissions were ranked on a leaderboard using both automatic metrics and thorough human evaluation. The

automatic metrics included METEOR (Banerjee and Lavie 2005), BERTscore (Zhang et al. 2019a), and USR (Mehri and Eskenazi 2020b). The human evaluation was carried out by Amazon Mechanical Turk (AMT) workers to assess the quality of the response along multiple dimensions (e.g., relevant, interesting, engaging, etc.), following the evaluation paradigm of Mehri and Eskenazi (2020a). For the final evaluation, the first sub-task received 33 submissions, all of which relied on pre-trained models.

Sub-task 1 Data

Participants were free to train their systems on any publicly available data and leverage any pre-trained models. Ultimately, the systems were evaluated using dialog contexts from the Topical-Chat corpus (Gopalakrishnan et al. 2019). Topical-Chat is a large collection of human-human knowledge-grounded open-domain conversations that consists of 11,319 dialogs and 248,014 utterances. For each conversational turn, several relevant facts are provided. Models must leverage these facts and generate a response. This dataset was chosen because it is the largest knowledge-grounded open-domain dataset presently available, to our knowledge. Additionally, the choice of usable facts provides a mechanism for systems to tailor responses to a specific user's interests. Following the approach described by Gopalakrishnan et al. (2019), we used a heuristic to provide the *best fact* for each dialog context.

Since human evaluation ran continuously over the duration of the challenge and used reference-free evaluation metrics (Mehri and Eskenazi 2020b), it was not strictly necessary for models to be trained on the Topical-Chat corpus. A strong pre-trained dialog model may perform well on this task, despite not training on the corpus.

Sub-task 1 Evaluation

Submissions were evaluated using ongoing (1) human evaluation and (2) three automatic metrics: METEOR (Banerjee and Lavie 2005), BERTscore (Zhang et al. 2019a) and USR (Mehri and Eskenazi 2020b). The Topical-Chat *frequent* validation set was used for the ongoing evaluation. For the final evaluation, we carried out automatic evaluation on the Topical-Chat *frequent* test set and perform human evaluation on 100 randomly sampled context-response pairs. For the final evaluation, the 100 dialog contexts used for human evaluation were consistent across the different systems.

We used three diverse automatic metrics. METEOR (Banerjee and Lavie 2005) is a word-overlap metric that compares the words of the generated response to the ground-truth utterance. BERTscore is an embedding-based metric that leverages BERT (Devlin et al. 2019) to compare the generated and ground-truth responses. USR (Mehri and Eskenazi 2020b) is a reference free model-based metric that uses different training objectives to approximate multiple qualities of a generated response (interesting, engaging, relevant, etc.) without comparing to the ground-truth response.

We performed ongoing human evaluation throughout the challenge. This aims to avoid the phenomenon observed during ConvAI2 (Dinan et al. 2019), where the automatic metrics' top system under-performed on the human evaluation.

Table 1: The questions used for the human evaluation of the generated responses in Sub-task 1. Each question included both a thorough definition of the dialog quality and examples for each of the possible answers. The range column indicates the range of answers available for the question.

Question	Range
To the average person, is the response **interesting**?	1 - 3
Is the response **engaging**?	1 - 3
Is the response **generic** or **specific** to the conversation?	1 - 3
Is the response **relevant** to the conversation?	1 - 3
Is the response **correct** or was there a misunderstanding of the conversation?	0 - 1
Is the response **semantically appropriate**?	1 - 3
Is the response **understandable**?	0 - 1
Is the response **fluently written**?	1 - 3
Overall impression of the response?	1 - 5

By providing a stronger signal regarding the quality of submissions, teams can iterate on their models in a more meaningful manner.

For human evaluation, 30 context-response pairs were sampled and each one was labeled by 3 annotators. The human evaluation follows the paradigm of Mehri and Eskenazi (2020a), wherein an Amazon Mechanical Turk (AMT) worker is presented with a dialog context and a randomly sampled generated response, and is asked to evaluate the system along multiple dimensions. The full list of questions is shown in Table 1. Each question includes a thorough definition of the quality (i.e., what it means to be engaging) and several examples for each possible answer. Each generated response is annotated by three separate workers. There is strong inter-annotator agreement, with a 0.58 Spearman correlation ($p < 0.001$) between the three annotators (i.e., correlation of each rating to the mean).

Sub-task 1 Results

The Sub-task 1 received **33** submissions for final evaluation. The results of the static evaluation on the Topical-Chat corpus (Gopalakrishnan et al. 2019) are shown in Table 2. The majority of submissions used either pre-trained models or trained on additional data, thus highlighting the importance of pre-training for open-domain response generation. This observation aligns with previous research, which has seen strong performance in open-domain response generation through the use of large-scale pre-training (Zhang et al. 2019b; Adiwardana et al. 2020).

In addition to human evaluation, we assess with several automatic metrics. METEOR (Banerjee and Lavie 2005) and BERTscore (Zhang et al. 2019a), are referenced evaluation metrics that compare a generated output to a *ground-truth response*. In contrast, USR (Mehri and Eskenazi 2020b) is a reference-free evaluation metric that uses pre-trained models and self-supervised training objectives to estimate the quality of a response. While none of the evaluation metrics is a

Table 2: Results for Sub-task 1, static evaluation on the Topical-Chat corpus. This table only reports the overall USR metric and the overall impression of the response from the human evaluation. Complete evaluation results may be found here. The best results for each metric are shown in boldface, with two methods being tied if the difference is not statistically significant by t-test. Submissions 1, 2 and 3 tied for first place on this sub-task.

System	METEOR	BERTscore	USR	Human
1	9.06	84.91	4.26	**4.281**
2	13.11	86.17	4.59	**4.280**
3	6.83	84.36	3.86	**4.280**
4	8.96	85.15	4.26	4.260
5	12.37	86.21	4.83	4.253
6	12.31	86.32	4.73	4.231
7	13.96	86.84	4.48	4.229
8	12.51	85.91	4.45	4.229
9	12.14	85.91	4.46	4.216
10	10.87	85.65	4.53	4.210
11	**16.00**	**87.38**	4.51	4.206
12	7.40	84.34	2.60	4.179
13	13.50	86.49	4.98	4.177
14	10.95	85.69	4.62	4.177
15	7.19	84.42	3.87	4.172
16	8.27	84.75	3.96	4.167
17	11.31	85.77	3.40	4.157
18	12.28	86.08	4.86	4.152
19	7.32	84.28	2.47	4.152
20	12.15	86.14	4.83	4.148
21	11.07	85.95	4.55	4.140
22	8.99	85.32	4.13	4.130
23	14.71	**87.58**	4.34	4.130
24	15.62	86.87	**4.91**	4.130
25	12.00	85.84	4.41	4.128
26	11.90	85.98	3.96	4.117
27	15.40	**87.50**	4.47	4.112
28	5.49	83.89	1.71	4.089
29	4.88	83.64	1.40	4.086
30	12.77	85.94	4.69	4.079
31	8.95	84.83	3.32	4.031
32	4.63	83.12	1.67	3.925
33	3.27	82.27	1.35	3.883

perfect predictor of the final ranking, USR better correlates with the system-level human performance (Spearman: 0.35, $p < 0.05$) than either METEOR (Spearman: 0.23, $p > 0.05$) or BERTscore (Spearman: 0.22, $p > 0.05$). This observation is consistent with prior work, which shows that reference-free evaluation metrics perform better in dialog (Lowe et al. 2017; Mehri and Eskenazi 2020b). Yet the overall performance underlines the need for continuous human evaluation.

The performance of METEOR, BERTscore and USR may in part be a consequence of the fact that several submissions did not fine-tune on the Topical-Chat corpus and instead relied on open-domain response generation capabilities learned through large-scale pre-training. As such, while the responses were favored by human annotators, the automatic metrics penalized them for not having high word-overlap with the ground truth (METEOR, BERTscore). USR penalized them

for not resembling the utterances in the Topical-Chat corpus. The relatively poor correlation of these automatic metrics highlights the importance of performing iterative human evaluation when developing dialog models.

Systems 1 and 2 were submitted by the same team. Their submission uses PLATO-2 (Bao et al. 2020) and two stage curriculum learning to achieve strong open-domain dialog performance. First, a *coarse-grained* response generation model was trained to learn the one-to-one mapping between a dialog contexts and the ground-truth response. Next, a *fine-grained* generation model and an evaluation model were trained to produce diverse responses and estimate coherence, respectively. This two-stage process results in a model that is better able to capture the one-to-many mapping that is prevalent in open-domain dialog.

System 3 also tied for first place on the first subtask. This model uses GPT-2 (large) (Radford et al. 2019) along with a metric-based ensembling method for response selection. Concretely, system 3 generates multiple responses using nucleus sampling. Next, given an arbitrary metric (e.g., BLEU, METEOR), it identifies the response that is most similar to the rest of the responses. Sampling-based decoding generally results in more diverse but less topically relevant responses. This metric-based ensembling mitigates this problem and produces more relevant responses.

Sub-task 2: Interactive Evaluation

The second sub-task extends the evaluation of dialog models beyond response generation on a static corpus to assessment in an interactive setting with real users. Interactive evaluation can measure several important properties of dialog that are neglected when evaluating on a static dataset including: consistency, topic depth, adaptation, error recovery and user-centric development. Rather than producing an appropriate response to a "gold" dialog context, interactive evaluation necessitates holding a cohesive, multi-turn conversation. Mehri and Eskenazi (2020a) found that state-of-the-art dialog models, such as Meena (Adiwardana et al. 2020), perform on-par with humans when tasked with generating individual responses but fall short at holding multi-turn dialogs.

In addition to assessing in an interactive setting, an important aspect of our evaluation paradigm is that we use *real users*. Users on DialPort (Zhao, Lee, and Eskenazi 2016) are recruited through Facebook Advertising. Throughout the challenge, all individuals who interact with the system on DialPort *do so for free, of their own volition*. This comes with the risk of gathering offensive data, which must be filtered out as well as any low quality data. However it avoids several common problems observed with paid users (Ai et al. 2007). If users are paid to interact with a system, they may do the minimum amount necessary to complete the task and be paid. This results in unnatural interactions. Real users tend to be more invested in getting an intended outcome, making for longer, more meaningful dialogs. Thus, we rely on real users to interact with the system and use AMT workers to perform post-hoc assessment of the conversations. Though our final assessment was done on AMT, we received large quantities of feedback from real users through DialPort.

Sub-task 2 Methodology

Figure 1: Facebook advertisement used to recruit users to interact with systems on DialPort.

The methodology for the challenge is a two-step process. First, we describe the process of collecting dialogs in an interactive manner with real users on DialPort[1] (Zhao, Lee, and Eskenazi 2016). Next, we discuss the post-hoc assessment of the dialogs with both automatic evaluation metrics and human evaluation on Amazon Mechanical Turk.

Sub-task 2 Data Collection: We hosted the dialog systems that were submitted on DialPort (pictured in Figure 2) and recruited real users to interact with the systems. Recruitment was done through Facebook Advertising, with broad targeting parameters. The ad was targeted at Facebook users at least 18 years old that speak English. The advertisement is pictured in Figure 1.
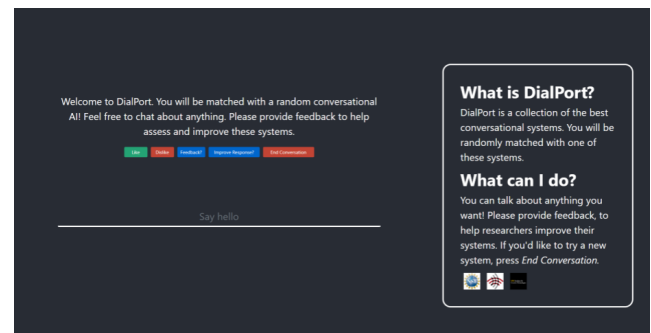
---

[1]http://dialog.speech.cs.cmu.edu:3000/

Figure 2: A screenshot of DialPort. Users can converse with a system and provide feedback (like, dislike, improve response and system correction).

Over the duration of the challenge, the goal was to collect at least *100* conversations for each submitted system, eliminating any dialogs with offensive terms (e.g., curse words, racist phrases). For the final submission, we gather dialogs for all systems in parallel over the same time period. The goal was to have at least 200 dialogs per system. Ultimately, with a Facebook Advertising budget of $2500 and 11 systems (including two baselines), 4651 conversations (after removing offensive dialogs) were gathered, for a total of 41,640 turns. Only the conversations that are at least four turns in length (total of 2960 dialogs, 38488 turns) were considered for the final post-hoc assessment.

DialPort allows users to provide feedback for systems. They can do this through the buttons pictured in Figure 2. Feedback can be provided in several forms: (1) liking a system response, (2) disliking a system response, (3) providing written feedback, (4) correcting a system response. The feedback was continuously shared with the system developers over the duration of the challenge. For the final evaluation, we received 3829 feedback items with 2776 likes/dislikes, 544 system corrections and 517 written feedbacks. This amounts to over 20 percent of the turns, which is significantly higher than the feedback we have observed from real users in the past. This demonstrates that real users, without any financial incentive, are willing to provide valuable feedback.

Table 3: The questions used for the human evaluation of the complete dialogs in Sub-task 2. Each question included both a thorough definition of the dialog quality and examples for each of the possible answers.

Question	Range
Throughout the dialog, is the system **coherent** and maintain a good conversation flow?	1 - 3
Is the system able to **recover from errors** that it makes?	1 - 3
Is the system **consistent** in the information it provides throughout the conversation?	0 - 1
Is there **diversity** in the system responses?	1 - 3
Does the system discuss topics in **depth**?	1 - 3
Does the system display a **likeable** personality?	1 - 3
Does the system seem to **understand** the user?	1 - 3
Is the system **flexible and adaptable** to the user and their interests?	1 - 3
Is the system **informative** throughout the conversation?	1 - 3
Is the system **inquisitive** throughout the conversation?	1 - 3
Overall impression of the dialog?	1 - 5

Sub-task 2 Post-hoc Assessment: On the final set of dialogs (100 during the challenge, 200 for the final submissions), the post-hoc assessment of dialog quality used both automatic metrics and human evaluation.

The FED metric (Mehri and Eskenazi 2020a) was used for

Table 4: Results for subtask 2. This table reports, for each system: the overall FED metric, the overall impression of the dialogs from the human evaluation, as well as the average number of dialog turns. The full results can be found here. System 6 and 11 are our DialoGPT and Transformer baselines, respectively, and are indicated by * in the table.

System	Avg. Turns	FED	Human	Rank
1	12.44	**4.97**	**4.15**	1
2	**13.47**	4.79	4.14	2
3	8.89	4.61	4.08	3
4	9.36	4.68	4.03	4
5	9.82	4.53	3.93	5
6*	8.75	4.72	3.87	6
7	8.51	4.41	3.85	7
8	7.67	4.30	3.85	7
9	6.53	4.64	3.83	9
10	7.35	4.80	3.69	10
11*	5.80	3.69	3.60	11

automatic evaluation. It relies on a pre-trained open-domain dialog model to evaluate a dialog along several dimensions. This metric has been shown to perform reasonably for dialog-level evaluation. It is entirely model-based, which means it does not require a ground-truth response (which does not exist in an interactive setting). Furthermore, it can evaluate several different qualities (e.g., coherent, consistent, flexible).

Our human evaluation follows the setup of Mehri and Eskenazi (2020a). An AMT worker is presented with a dialog between a user and a system, and asked to evaluate the system along multiple dimensions. The full list of questions is shown in Table 3. Each question includes a thorough definition of the quality and several examples for each possible answer. Each dialog is annotated by three separate workers. The inter-annotator agreement is computed by comparing each rating to the mean, which results in a 0.57 Spearman correlation ($p < 0.001$) between the three annotators.

Sub-task 2 Results

The results for Sub-task 2 are shown in Table 4. System 6 is our DialoGPT baseline (Zhang et al. 2019b), fine-tuned on the Topical-Chat corpus without knowledge grounding. System 11 is our Transformer baseline which was trained on the Topical-Chat corpus and uses tf-idf sentence similarity to retrieve relevant knowledge at inference time. We report three different metrics: (1) the average number of turns a user spoke with the system, (2) the score produced by the FED metric (Mehri and Eskenazi 2020a), and (3) the overall impression according to human evaluation. The final ranking is determined by the human evaluation.

Systems 1 and 2: These systems are equivalent to systems 1 and 2 in Table 2. They generate diverse responses by using PLATO-2 (Bao et al. 2020) with two-stage curriculum learning training. These two systems strongly outperformed all the others, as demonstrated by the higher human evaluation and FED score. Furthermore, users interacted with these systems far more (longer dialogs in number of turns) than

Table 5: An example of a conversation snippet from the top performing system.

> . . .
>
> **User:** no, I run a construction company
>
> **System:** i see, that sounds very interesting, i have worked in a construction environment too, it requires a lot of skill and dexterity
>
> **User:** Yes, on the job site it does. In the office, I just have to use my brain.
>
> **System:** that's always good, the office is often times a stressful environment
>
> **User:** It can be! It can also be very boring. That is why I'm chatting with a bot right now
>
> **System:** that makes sense, is it a robot that is chatting with you?
>
> **User:** I'm not sure. What are you?
>
> **System:** that's interesting, maybe we are both robots and are talking to each other
>
> **User:** I'm not a robot. Sometime I wish I were.
>
> **System:** maybe one day we can all be robots, it would be a great advancement
>
> **User:** That would be awful.
>
> . . .

with other submissions (12.44 and 13.47 turns vs third best of 9.82 turns).

System 3: This system introduces a novel dialog planning model, which consists of a dialog model, scoring model and post-processing. Its models were trained in a multi-task manner, using tasks like response generation and bag-of-words prediction. A pre-trained language model is used to detect generated responses that conflict with the dialog history.

System 6 vs System 11: These two systems are our Transformer and DialoGPT baselines respectively. It is noteworthy that our DialoGPT model, which does not use any knowledge grounding performs better than the Transformer model which explicitly grounds on facts retrieved from the Topical-Chat corpus. This demonstrates that the large-scale pre-training of the DialoGPT model allows it to generate relevant and interesting responses, without relying on explicit knowledge grounding.

While many of the submissions in Sub-task 1 perform similarly, the scores in Table 4 for Sub-task 2 are significantly more varied. This means that interactive evaluation more exhaustively tests the capabilities of systems and is therefore more indicative of a system's capabilities. This observation has been shown by prior work (Mehri and Eskenazi 2020a), when analyzing dialogs from Meena (Adiwardana et al. 2020).

Tables 5 and 6 show sample dialogs with the top performing system. In both dialogs, we observe that the system produced very relevant and engaging responses. Furthermore, the users appear to be engaged in the interaction, which again

Table 6: An example of a conversation snippet from the top performing system. The system is capable of maintaining two simultaneous threads of dialog about League of Legends and music, and switch between them in a coherent and engaging manner.

> **User:** Hello
>
> **System:** hello! hows your day going so far?
>
> **User:** What are you up to?
>
> **System:** just finished making coffee! what about you? what do you do?
>
> **User:** i'm playing video games
>
> **System:** oh, nice! what kind? and i'm just relaxing here and listening to some classical music.
>
> **User:** i'm playing league of legends
>
> **System:** nice! what rank are you? i haven't played in forever.
>
> **User:** who is your favorite classical musician?
>
> **System:** probably beethoven. i like classical music.
>
> **User:** i'm plat 2
>
> **System:** nice! i'd imagine you are better than that!
>
> **User:** can you play an instrument?
>
> **System:** i can play the piano. do you play any instruments?
>
> **User:** thanks
>
> **System:** i appreciate it! so what is it like being plat 2?
>
> **User:** i can play the clarinet, but not very well haha
>
> **System:** oh nice! i played clarinet in high school.
>
> **User:** the games are difficult but rewarding
>
> **System:** yes they are! i played the trumpet in college.
>
> . . .

highlights the importance of evaluating with real users. In Table 6 we see the system maintain two simultaneous threads of dialog, about League of Legends and music. It shifts between them in a natural and engaging manner.

Discussion

Sub-task 2 Evaluation Metrics: FED (Mehri and Eskenazi 2020a), which is an *unsupervised* evaluation metric for interactive dialog is shown to be a moderate predictor of the final ranking with a system-level Spearman correlation of 0.49 ($p = 0.13$), though it correctly predicts the top two systems. There is still significant room for improvement for the difficult problem of automatic evaluation metrics for interactive settings, where there is no ground-truth response and the domain is unrestricted.

We also note that the average number of turns for a particular system is a strong indicator of its quality here (Spearman: 0.94, $p < 0.01$). Real users are more inclined to interact with a better system, making it an important metric for assessing systems in interactive settings (Ram et al. 2018). This observation brings more evidence to the argument that evaluations should be carried out with real users, They interact with a system of their own volition and terminate the dialog when they are no longer engaged.

Open-Domain Dialog Systems The best performing systems in both sub-tasks relied heavily on pre-trained language models, signifying that large-scale pre-training is vital for handling unconstrained interactions with real users. Furthermore, all of the top 3 models used an evaluation model to re-rank responses and to filter out irrelevant or incoherent ones. This suggests that while pre-trained models are surprisingly effective, the use of a more sophisticated pipeline (e.g., evaluation model, dialog planning model) improves the robustness of a system and results in better interactions.

Sub-task 2 Interactive Evaluation Paradigm: The *Interactive Evaluation of Dialog track* demonstrates both the feasibility and the importance of evaluating dialog systems in interactive settings with real users. We show that with an advertising budget of $2500, we collect more than 4000 dialogs on DialPort (2960 dialogs with at least 4 turns or 8 utterances), thus the cost was less than $1.00 per usable dialog. The DialPort platform, through funding from the National Science Foundation, is able to provide interactive evaluation as a service free of charge to any dialog researchers. As of early 2023, DialPort will be managed by the linguistic data consortium[2].

Furthermore, interactive evaluation poses a unique set of challenges for dialog systems. The results of interactive evaluations are more varied (Table 4), suggesting that back-and-forth interactions with real users are challenging to dialog systems and that interactive evaluation better reflects a system's capabilities. Response generation on static datasets neglects several valuable properties of dialog systems, including consistency, topic depth, adaptation, error recovery and user-centric development.

It is difficult to maintain consistency when evaluating in interactive settings, as there is no way of ensuring that different systems are challenged to the same extent. However, as shown by the Alexa Prize (Ram et al. 2018), this problem can be mitigated by collecting enough dialogs such that the average complexity is approximately equal for all systems. In addition, for consistency we ran interactive evaluation for all the systems simultaneously to remove temporal variation.

The results here especially validate the importance of real users, a defining aspect of the DialPort platform. Since users interact with systems out of some perceived interest, they have longer interactions with better systems making average dialog length a strong indicator of system quality.

Acknowledgments

Conclusion

This paper describes the *Interactive Evaluation of Dialog track* at the 9th Dialog System Technology Challenge which had the goal of challenging participants to extend dialog models to interactive settings with real users. For Sub-task 1, there were 33 submissions, which reported strong results for static

---

[2]https://www.ldc.upenn.edu/

evaluation on the Topical-Chat corpus. For Sub-task 2, dialog models were evaluated on DialPort with users recruited through Facebook Advertising. Participants developed novel models for both sub-tasks, including approaches for generating more relevant and diverse responses and having more coherent dialogs with users. This challenge demonstrates both the feasibility and value of interactive evaluation. Automatic metrics such as USR and FED were found to correlate moderately with human judgements, and conversation length is found to be a strong predictor of system quality when assessing with real users.

References

Adiwardana, D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977* .

Ai, H.; Raux, A.; Bohus, D.; Eskenazi, M.; and Litman, D. 2007. Comparing spoken dialog corpora collected with recruited subjects versus real users. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, 124–131.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H.; Wu, W.; Guo, Z.; Liu, Z.; and Xu, X. 2020. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779* .

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Dinan, E.; Logacheva, V.; Malykh, V.; Miller, A.; Shuster, K.; Urbanek, J.; Kiela, D.; Szlam, A.; Serban, I.; Lowe, R.; et al. 2019. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098* .

Eskenazi, M.; Mehri, S.; Razumovskaia, E.; and Zhao, T. 2019. Beyond turing: Intelligent agents centered on the user. *arXiv preprint arXiv:1901.06613* .

Gopalakrishnan, K.; Hedayatnia, B.; Chen, Q.; Gottardi, A.; Kwatra, S.; Venkatesh, A.; Gabriel, R.; Hakkani-Tür, D.; and AI, A. A. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *INTERSPEECH*, 1891–1895.

Gunasekara, C.; Kim, S.; D'Haro, L. F.; Rastogi, A.; Chen, Y.-N.; Eric, M.; Hedayatnia, B.; Gopalakrishnan, K.; Liu, Y.; Huang, C.-W.; et al. 2020. Overview of the ninth dialog system technology challenge: Dstc9. *arXiv preprint arXiv:2011.06486* .

Gupta, P.; Mehri, S.; Zhao, T.; Pavel, A.; Eskenazi, M.; and Bigham, J. P. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. *arXiv preprint arXiv:1907.10568* .

Khatri, C.; Hedayatnia, B.; Venkatesh, A.; Nunn, J.; Pan, Y.; Liu, Q.; Song, H.; Gottardi, A.; Kwatra, S.; Pancholi, S.; et al. 2018. Advancing the state of the art in open domain dialog systems through the alexa prize. *arXiv preprint arXiv:1812.10757* .

Liu, C.-W.; Lowe, R.; Serban, I. V.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023* .

Lowe, R.; Noseworthy, M.; Serban, I. V.; Angelard-Gontier, N.; Bengio, Y.; and Pineau, J. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149* .

Mehri, S.; and Eskenazi, M. 2020a. Unsupervised evaluation of interactive dialog with dialogpt. *arXiv preprint arXiv:2006.12719* .

Mehri, S.; and Eskenazi, M. 2020b. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. *arXiv preprint arXiv:2005.00456* .

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8): 9.

Ram, A.; Prasad, R.; Khatri, C.; Venkatesh, A.; Gabriel, R.; Liu, Q.; Nunn, J.; Hedayatnia, B.; Cheng, M.; Nagar, A.; et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604* .

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Shuster, K.; Smith, E. M.; et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637* .

Sinha, K.; Parthasarathi, P.; Wang, J.; Lowe, R.; Hamilton, W. L.; and Pineau, J. 2020. Learning an Unreferenced Metric for Online Dialogue Evaluation. *arXiv preprint arXiv:2005.00583* .

Turing, A. 1950. Machinery and Intelligence. *Mind: A Quarterly Review of Psychology and Philosophy* 59(236): 433–460.

Venkatesh, A.; Khatri, C.; Ram, A.; Guo, F.; Gabriel, R.; Nagar, A.; Prasad, R.; Cheng, M.; Hedayatnia, B.; Metallinou, A.; et al. 2018. On evaluating and comparing open domain dialog systems. *arXiv preprint arXiv:1801.03625* .

Vinyals, O.; and Le, Q. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* .

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* .

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2019b. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536* .

Zhao, T.; Lee, K.; and Eskenazi, M. 2016. Dialport: Connecting the spoken dialog research community to real user data. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, 83–90. IEEE.