Overview of the Eighth Dialog System Technology Challenge: DSTC8

Kim, Seokhwan; Galley, Michel; Gunasekara, Chulaka; Lee, Sungjin; Atkinson, Adam; Peng, Baolin; Schulz, Hannes; Gao, Jianfeng; Li, Jinchao; Adada, Mahmoud; Huang, Minlie; Lastras, Luis; Kummerfeld, Jonathan K.; Lasecki, Walter S.; Hori, Chiori; Cherian, Anoop; Marks, Tim K.; Rastogi, Abhinav; Zang, Xiaoxue; Sunkara, Srinivas; Gupta, Raghav

TR2021-064 June 04, 2021

Abstract

This paper introduces the Eighth Dialog System Technology Challenge. In line with recent challenges, the eighth edition focuses on applying end-to-end dialog technologies in a pragmatic way for multi-domain task-completion, noetic response selection, audio visual scene-aware dialog, and schema-guided dialog state tracking tasks. This paper describes the task definition, provided datasets, baselines and evaluation set-up for each track. We also summarize the results of the submitted systems to highlight the overall trends of the state-of-the-art technologies for the tasks.

IEEE/ACM Transactions on Audio, Speech, and Language Processing

Overview of the Eighth Dialog System Technology Challenge: DSTC8

Seokhwan Kim, Michel Galley, Chulaka Gunasekara, Sungjin Lee, Adam Atkinson, Baolin Peng, Hannes Schulz, Jianfeng Gao, Jinchao Li, Mahmoud Adada, Minlie Huang, Luis Lastras, Jonathan K. Kummerfeld, Walter S. Lasecki, Chiori Hori, Anoop Cherian, Tim K. Marks, Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta

***Abstract*—This paper introduces the Eighth Dialog System Technology Challenge. In line with recent challenges, the eighth edition focuses on applying end-to-end dialog technologies in a pragmatic way for multi-domain task-completion, noetic response selection, audio visual scene-aware dialog, and schema-guided dialog state tracking tasks. This paper describes the task definition, provided datasets, baselines and evaluation set-up for each track. We also summarize the results of the submitted systems to highlight the overall trends of the state-of-the-art technologies for the tasks.**

I. Introduction

The Dialog System Technology Challenge (DSTC) is an ongoing series of research competitions for dialog systems. To accelerate the development of new dialog technologies, the DSTCs have provided common testbeds for various research problems. The earlier Dialog State Tracking Challenges [1], [2], [3] focused on developing a single component for dialog state tracking on goal-oriented human-machine conversations. Then, DSTC4 [4] and DSTC5 [5] introduced human-human conversations and started to offer multiple tasks not only for dialog state tracking, but also for other components in dialog systems as the pilot tasks. From the sixth challenge [6], the DSTC has rebranded itself as “Dialog System Technology Challenge” and organized multiple main tracks in parallel to address a wider variety of dialog related problems. Most recently, DSTC7 [7], [8] focused on developing end-to-end dialog technologies for the following three tracks: noetic response selection [9], [10], grounded response generation [11], and audio visual scene aware dialog [12].

For the eighth DSTC, we received seven track proposals and went through a formal peer review process focusing on each task’s potential for (a) broad interest from the research community, (b) practical impact of the task outcomes, and (c) continuity from the previous challenges. Finally, we ended up with the four main tracks including two newly introduced tasks and two follow-up tasks of DSTC7. Multi-domain task-completion track (Section II) addresses the end-to-end response generation problems in multi-domain task completion and cross-domain adaptation scenarios. NOESIS II (Section III) explores a response selection task extending the first NOESIS track in DSTC7 and offers two additional subtasks for identifying task success and disentangling conversations. Audio visual scene-aware dialog track (Section IV) is another follow-up track of DSTC7 which aims to generate dialog responses using multi-modal information given in an input video. Schema-guided dialog state tracking track (Section V) revisits dialog state tracking problems in a practical setting associated with a large number of services/APIs required to build virtual assistants in practice.

More than 280 participants were registered in one or several of the tracks; finally 70 teams submitted their final results; and 37 scientific papers were presented in the DSTC8 workshop which was held on February 8, 2020 collocated with the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20). The remainder of this paper describes the details of each track.

II. Multi-Domain Task-Completion Track

This track offers two tasks to foster progress in two important aspects of dialog systems: dialog complexity and scaling to new domains. One task is the end-to-end task-oriented dialog task aiming to solve the complexity of building end-to-end dialog systems that span over multiple sub-domains to accomplish complex user goals. The other is the fast domain adaptation task to address the domain adaptation problem by investigating how a dialog system trained on a large corpus can be adapted to a new domain given a smaller in-domain corpus.

A. Task 1: End-to-end multi-domain dialog system

Previous work in dialog research communities mainly focuses on individual components in a dialog system and pushes forward the performance of each component. However, the improvement of individual components does not necessarily boost the entire system performance [13], [14]. The metrics used for an individual component might not be significant for an end-to-end system, and the propagation of error down the pipeline is likely to mitigate the component-wise improvement. With these concerns, recently researchers have taken efforts to create end-to-end approaches [15], [16], but it is hard to compare them with conventional methods given the efforts and complexity to combine individual models in conventional approaches.

To address these concerns, we provide ConvLab[1] [13], a multi-domain end-to-end dialog system platform covering a range of state-of-the-art models, to reduce the efforts of building

[1] https://github.com/ConvLab/ConvLab

and evaluating end-to-end dialog systems. Based on ConvLab, participants of the task are to build a dialog system that takes natural language as input, tracks dialog states during the conversation, interacts with a task-specific knowledge base, and generates natural language response as output. There is no restriction on system architectures, and participants are encouraged to explore various approaches ranging from conventional pipeline systems to end-to-end neural approaches.

1) Data: In this task, we consider MultiWOZ [17] dataset, a dialog corpus collected from conversations over multiple domains under the tourist information desk setting. It consists of 10,438 dialogues and covers 7 domains, including *Attraction, Hospital, Police, Hotel, Restaurant, Taxi*, and *Train*. We enhanced the dataset with additional annotation for user dialog acts, which is missing in the original dataset, and included it in ConvLab.

2) Evaluation and Results: Two evaluation metrics are offered in this task:

Simulator-based evaluation: The end-to-end user simulator for automatic evaluation is constructed by combining an agenda-based user simulator [18], a rule-based natural language generation (NLG) model and a multi-intent language understanding (MILU) model, all of which have been implemented in ConvLab. The evaluation metrics employed include success rate, average reward, and number of turns for each dialog. We also report precision, recall, and F1 score for slot prediction.

Crowdworker-based human evaluation: With simulator-based automatic evaluation, we filter out low-quality submissions and send the remaining systems to Amazon Mechanic Turk for human evaluation. Crowd-workers communicate with the system via natural language, judge the system and provide ratings based on language understanding correctness, response appropriateness on a 5 point Likert-scale. Extra metrics including success rate and number of turns are also reported.

Twelve teams participated in this task. Table I lists the results for both human evaluation and simulator-based evaluation. A component-wise system with BERT-based NLU model [19], elaborated rule-based dialog policy and dialog state tracker achieves the best success rate of 88.80% in simulator-based evaluation. However, there are discrepancies between human evaluation and simulator-based evaluation. The best system in the human evaluation is based on fine-tuning GPT-2 [20]. It predicts dialog states, system actions, and responses in an end-to-end fashion, and achieves a success rate of 68.32%.

B. Task 2: Fast Adaptation Task

Neural dialog response generators require very large datasets to learn to output consistent and grammatically correct sentences [21], [22], [23]. This makes it extremely hard to scale out the system to new domains with limited in-domain data, for example, when modeling user responses for a task-oriented chatbot on a narrow domain. With this challenge, our goal is to investigate whether sample complexity can decrease with time, *i.e.*, if a dialog system that was trained on a large corpus can learn to converse about a new domain given a much smaller in-domain corpus.

1) Data: We provide two dialog datasets, in which each dialog belongs to exactly one domain.

Reddit Dataset We constructed a corpus of over five million dialogs from Reddit submissions and comments spanning one year of data. Content is selected from a curated list of one thousand subreddits using a methodology similar to the DSTC7 sentence generation task [11]. We provide pre-processing code for Reddit data so that all participants work on the same corpus.

Goal-Oriented Corpus MetaLWOz We collected 37 884 goal-oriented dialogs via crowd-sourcing using a *Wizard of Oz* scheme. These dialogs span 47 domains (*e.g.* bus schedule, alarm setting, banking) and are particularly suited for meta-learning dialog models. For each dialog, we paired two crowd-workers, one had the role of being a bot, and the other one was the user. We defined 227 tasks distributed over the domains. Note that all entities were invented by the crowd-workers (for instance, the address of a bus stop) and the goal of this challenge is to predict convincing *user* utterances.

2) Evaluation and Results: We evaluate responses by the domain-adapted dialog model using two metrics:

Automatic metrics: For each incomplete test dialogue, a set of 128 complete single-domain MultiWOZ [17] dialogs is provided to the model, which is then asked to respond to the incomplete test dialog. Intents and slot values correctly detected by the baseline NLU (cf. Sec. II-A) in the response serve as an indicator that the domain adaptation was successful. We report intent F1 as well as intent+slot F1.

Human evaluation: The model is given a small set of complete dialogs from a held-out MetaLWOz domain, and is then asked to predict a response to an incomplete dialog from the same domain. Three human annotators were asked to judge the appropriateness, informativeness and utility of the responses [11] *given the MetaLWOz task*, i.e. whether the simulated user tries to complete the task. Crowd-workers submit pairwise binary preference judgements given dialog context and metric. Pairs are picked using *Multisort* [24] and per dialog/metric rankings are aggregated using Copeland's method [25]. We use bootstrapping [26] over dialog contexts to assess ranking robustness and found it to be stable. Inter-annotator agreement [27], [28] is at $\kappa = 0.29$. No method outperformed the ground truth.

As a baseline, we provided a retrieval model that relies on FastText embeddings [29] of SentencePiece tokens [30] and only takes into account the given in-domain dialogs. The track received four submissions, all of which surpassed baseline performance on automatic evaluation. As in Task 1 (Sec. II-A2), we find differences in ranking between human and automatic evaluation.

The two best teams use a Transformer [31] (TeamB) or BiLSTM-based [32] (TeamA) base model that is fine-tuned on the in-domain dialogs. The BiLSTM-based model is additionally fine-tuned on dynamically sampled Reddit dialogs, while the Transformer model additionally ranks both the observed in-domain dialog responses and the generated

TABLE I: Task 1 Evaluation Results

	Human Evaluation				Simulator-based Evaluation						
Team	Succ. %	Under.	Appr.	Turns	Succ. %	Reward	Turns	Prec.	Rec.	F1	Book %
Baseline	56.45	3.10	3.56	17.54	63.40	30.41	7.67	0.72	0.83	0.75	86.37
$\text{Best}_{\text{human}}$	68.32	4.15	4.29	19.51	79.40	49.69	7.59	0.80	0.89	0.83	87.02
$\text{Best}_{\text{simulator}}$	65.81	3.54	3.63	15.48	88.80	61.56	7.00	0.92	0.96	0.93	93.75

$\text{Best}_{\text{human}}$: The best team for human evaluation. $\text{Best}_{\text{simulator}}$: The best team for simulator-based evaluation.
Metrics: Succ.: success rate, Under.: understanding score, Appr.: appropriateness score, Prec./Rec.: precision/recall of slots prediction.

response using next sentence classification. Team C first fine-tuned GPT-2 on the MetaLWOz training corpus, then fine-tuned it further on the support sets of the MetaLWOz and MultiWOZ test sets. Team D trained a Bi-LSTM encoder and attentional LSTM decoder on both Reddit and MetaLWOz training corpora, without any fine-tuning to the test sets.

III. NOESIS II: Predicting Responses Track

This track is a follow-up to DSTC 7 Track 1, 'NOESIS: Noetic End-to-End Response Selection Challenge' [7], where the next-utterance selection problem in two-party dialogues was considered in two domains. This task extends the previous challenge in three ways: (1) conversations with more than two participants; (2) predicting whether a dialogue has solved the problem or not; (3) handling multiple simultaneous conversations in the same communication channel. Each of these adds an important aspect of real-world conversations.

A. Task definition

The primary task of focus is the next-utterance selection. In this problem, each example consists of a partial dialogue and a set of options for what the next utterance is in the dialogue. Participants must rank the potential messages plus the possibility that the true next message is not in the set. We followed the configuration from DSTC 7 track 1, with one hundred options for the next message. In 20% of cases, the true next message is not in the set. Participants were permitted to use the provided external knowledge sources in their systems.

1) Supplementary task 1: In-Channel Selection: The first supplementary task was a variant of the main task in which the conversation context was not a prefix of a single conversation, but instead a section of chat from the raw Ubuntu IRC channel. The raw chat often contained multiple conversations, including cases where speakers participate in multiple conversations simultaneously. To reduce ambiguity about which conversation the next message is part of, we provided the identity of the speaker.

2) Supplementary task 2: Task Completion Success: The second supplementary task considered identification of task success in the Advising data. Specifically, we provided a partial conversation and participants had to identify utterances that indicated the student had accepted or rejected the advisor's suggestion. Cases were also included in which no utterance accepting or rejecting the suggestion was present.

3) Supplementary task 3: Dialogue Disentanglement: The final supplementary task considered the process of extracting conversations from chat logs. We provided sections of the logs as input and requested sets of messages as output, where each set corresponded to a conversation.

The detailed task description is shown at the github page[2].

B. Data

As in DSTC 7 track 1, two sources of data were considered. Both are task oriented, but one is much broader in scope and has more data (Ubuntu) while the other is smaller and more focused (Advising).

1) Ubuntu: A new set of disentangled Ubuntu IRC dialogues was provided for this challenge based on recent work [33]. These are derived from the raw Ubuntu logs directly, not from any prior corpus. The dataset consists of multi-party conversations extracted from the Ubuntu IRC channel.[3] A typical dialogue starts with a question that was asked by one participant, and then other participants respond with either an answer or follow-up questions that then lead to a back-and-forth conversation. In this challenge, the context of each dialogue contains at least three messages between the participants. The next turn in the conversation is guaranteed to be from one of the participants who has spoken so far.

For the first supplementary task, we use raw samples from the channel, with pre-processing for speaker identification. For the third supplementary task, we used data from [33], without pre-processing.

The test data for each task was chosen so that it did not overlap with any other sets. For example, the test data for the main task came from a portion of the IRC log that was not used for training or testing in any other subtask. This was done to avoid information leakage across tasks and data.

We randomly split the conversations into training, development, and test sets. The development set had 4,827 conversations, the test set had 5,529 conversations, and the training set had the rest. For the first supplementary task there were 112,262 instances for training, 9,565 for development, and 9,027 for testing. For the third supplementary task, we use the training, development and test split from [33], which contains 67,463 messages for training, 2,500 for development, and 5,000 for testing.

2) Advising: This dataset contains two party dialogues that simulate a discussion between a student and an academic advisor. The purpose of the dialogues is to guide the student to pick courses that fit not only their curriculum, but also personal preferences about time, difficulty, areas of interest, etc. The conversations used are the same as those used in

[2] https://github.com/dstc8-track2/NOESIS-II

[3] https://irclogs.ubuntu.com/

TABLE II: Fast Adaptation Task Evaluation Results

	Automatic Evaluation		Human Evaluation	
Submission	Intent F1	Intent & Slot F1	Mean Bootstrap Rank	Final Rank
Baseline	0.52	0.27	3.97	4
TeamA	**0.79**	**0.60**	3.03	3
TeamB	0.64	0.48	**1.01**	**1**
TeamC	0.61	0.42	1.99	2
TeamD	0.55	0.42	5.00	5

Time	Speaker	Message
12:30	s_0	how can i boost microphone volume? The volume is toooooo low
12:30	s_1	s_0 , look for a microphone boost in alsamixer
12:30	s_2	s_0 : type 'alsamixer' into terminal
12:31	s_0	how the heck do i use alsamixer? :P what is microphone ?
12:32	s_0	how do i choose volume on input s_2 ?
12:33	s_2	s_0 : arrow keys up and down
12:33	s_0	s_2 , yes i understand that. But wich one of those things am i supposed to choose ?
12:33	s_2	s_0 : you wanted input, right?
12:34	s_0	s_2 , yes. But i there is no way i can turn that up. :S
12:34	s_2	s_0 : press tab to go over to capture, then turn it up
12:34	s_0	aha :) thanks

Speaker	Message
Student	Hello!
Advisor	Hi!
Student	I am currently trying to figure out what courses to take next semester.
Student	Could you suggest any?
Advisor	Let me see. Give me a minute to go over your transcript
Advisor	Can you tell me what your preferences are?
Student	Of course! I am interested in Computer Science, video game design is something that has always been interesting for me.
Advisor	Eecs 280 I should a prerequisite for most computer science classes, including game design
Student	Okay yeah I will take that course. Do you know of any other prerequisites for game design?
Advisor	Eecs 281 is also necessary, and unfortunately you can't take both 280 and 281 in the same semester.
Advisor	You should take Eecs 203 as that is also a prerequisite for most Eecs classes
Student	Okay thanks for the info! Are both EECS 203 and EECS 280 project based?
Advisor	280 is all project based and 203 is not, but don't let that fool you. Many students say 203 is harder than 280
Student	Oh wow okay so do you think that taking them both in the same semester will be manageable?
Advisor	If you have a good grasp of probability and combinations it I should perfectly manageable

Fig. 1: Examples of data in NOESIS II track: new dialogues from Ubuntu (top) and prior dialogues from Advising (bottom).

Property	Advising	Ubuntu
Dialogues	700	496,469
Average Number of Speakers	2	2.6
Utterances / Dialogue	18.6	7.2
Tokens / Utterance	9.8	11.4
Utterances / Unique utterances	4.4	1.2
Tokens / Unique tokens	10.5	44.1

TABLE III: Comparison of the two data sources (based on training, development, and test data). Tokens are identified by splitting on whitespace.

DSTC 7 task 1 [7]. They were collected by having students at the University of Michigan act as the two roles using provided personas. Structured information in the form of a database of course information was provided, as well as the personas (though at test time only information available to the advisor was provided, i.e. not the explicit student preferences). The data also includes paraphrases of the sentences and of the target responses.

The training, development, and test sets were the same as in DSTC 7 track 1. The development and test sets are based on 100 raw conversations, each paraphrased five times and then cut off at different points. The training set is based on 500 conversations, also paraphrased five times, but then remixed many times. For the second supplementary task, we use the same data split. Instances were annotated by one of the authors.

Examples conversations from the two datasets are shown in Figure 1 and and Table III shows stats about the datasets. A training set with answers was provided to participants to use as they wished. For the evaluation period, inputs for the test set were provided. The answers for the test set were not released until after the challenge was complete.

C. Evaluation and Results

A range of metrics were considered to evaluate the submissions. The main task and the first supplementary task followed DSTC 7 track 1, using the mean of (a) Recall@10 and (b) mean reciprocal rank (MRR). The second supplementary task used accuracy, measured as whether each example was correct or not. The final supplementary task used precision, recall, and F-score over complete conversations and several clustering metrics (Variation of Information [34], Adjusted Rand Index, and Adjusted Mutual Information). These treat each message as an item and conversations as clusters.

We provided baselines for task 1 and task 4. The task 1 baseline was a slightly modified form of the encoder-decoder baseline provided in track one of DSTC 7. The task 4 baseline was from [33]: a feedforward neural network that uses GloVe embeddings and structural features to represent the conversation.

D. Discussion

In this section we discuss the results obtained by the participants in all tasks for the challenge.

1) Main Task: While many participants used a version of BERT[19] as their main model, there was still a broad range of results. This indicates the importance of elements like the loss definition, data augmentation, and text segmentation, in achieving strong results. The team ranked 1st for this task used RoBERTa[35] with a technique to augment the training data and a binary classification at the last layer of the network. One clear trend was a switch from the ESIM model [36] used by participants in DSTC 7 to BERT and RoBERTa.

Performance varied significantly on the main task, with the best teams scoring far higher than the reference baseline provided. As in DSTC 7, the Advising data proved harder. Unlike in DSTC 7, the best approach varied across the datasets, with the best approach on Ubuntu coming second on Advising and the best approach on Advising coming 5th on Ubuntu.

One aspect of the challenge was identifying when no true answer was present. Seven teams did better on cases with no answer and ten teams did better on cases with an answer. Table IV shows the scores of top 3 teams on the main task.

2) Supplementary task 1: In-Channel Selection: As expected, shifting to the more realistic setting of the raw channel led to lower performance. This suggests that the complications introduced in the raw setting are real, but surmountable. Table V shows the scores of the 3 top performed teams on this task.

3) Supplementary task 2: Task Completion Success: All three teams that attempted this new task performed well in general and the results are shown in Table VI. Task success could be a good signal for training dialogue systems with reinforcement learning, and so these results are an encouraging sign that automated training via interaction with people may be feasible (with success detection as the reward).

Two teams had almost exactly the same results (3 and 15). Investigating these further, we found several differences in the patterns of errors in their output. Team 15 tended to predict “No Decision Yet” more often, achieving higher recall and lower precision that team 3 on that category. The trend was reversed for “Accept”, with team 3 predicting it more often and achieving higher recall and lower precision. For identifying “Reject”, the results were extremely similar.

4) Supplementary task 3: Dialogue Disentanglement: Only one team attempted this supplementary task, but they achieved strong performance (shown in Table VII, improving over the baseline by 7.8 F_1. This is still far from perfect performance, indicating that this problem remains an open challenge.

IV. Audio Visual Scene-Aware Dialog Track

Spoken dialog systems on the market are still missing one important piece of technology: natural and context-aware human-machine interaction, in which machines understand the surrounding scene from the human perspective and are able to share their understanding with humans using natural language. The goal of building an automated system to converse with humans about surrounding scenes via natural dialog is a challenging research problem that lies at the intersection of natural language processing, computer vision, and audio processing. To advance research into multimodal reasoning-based dialog generation, we developed the Audio Visual Scene-Aware Dialog (AVSD) dataset and held the AVSD challenge in DSTC7. The DSTC7 winning system of the challenge applied hierarchical attention mechanisms to combine text and visual information, yielding a relative improvement of 22% in the human rating of the output of the winning system vs. that of the baseline system. This large improvement suggested that there is perhaps significantly more potential in store for advancing this new research area. Toward this end, we proposed a second edition of our AVSD challenge in DSTC8.

A. AVSD Task definition

This track is a follow-up of the AVSD track from DSTC7 to evaluate using on a new set of previously unseen test questions from the AVSD dataset. In this track, the system must generate a sentences as a response to a user question, in the context of a given dialog about a video. The target of both VQA and VisDial was *sentence selection* based on information retrieval. For real-world applications, however, spoken dialog systems cannot simply select from a small set of predetermined sentences. Instead, they need to immediately output only one response to a user input, and the quality of the 1-best hypothesis needs to be evaluated precisely. For this reason, in this track we focus on *sentence generation* rather than sentence selection. In this track, the system's task (illustrated in Figure 2) is to use a dialog history (the previous rounds of questions and answers in a dialog) and (optionally) a brief video script (also referred to as the caption), plus (in one version of the task) the visual and audio information from the input video, to answer a follow-up question about the video. The detailed task description is shown at the github page of DSTC8 AVSD[4].

[4] https://github.com/dialogtekgeek/DSTC8-AVSD

TABLE IV: Results of the top 3 performers in Track 2 - main task (subtask 1)

Ubuntu					Advising				
Team	R@1	R@5	R@10	MRR	Team	R@1	R@5	R@10	MRR
Baseline	0.212	0.421	0.565	0.325	Baseline	0.222	0.493	0.622	0.355
15	**0.761**	**0.958**	**0.979**	**0.848**	17	**0.564**	**0.806**	**0.878**	**0.677**
12	0.719	0.948	0.976	0.819	15	0.306	0.632	0.762	0.455
5	0.663	0.943	0.974	0.786	13	0.254	0.560	0.690	0.401

	Recall@				
Team	1	5	10	MRR	Score
3	0.505	0.755	0.834	0.621	0.727
13	0.596	0.847	0.904	0.707	0.806
15	**0.706**	**0.916**	**0.957**	**0.799**	**0.878**

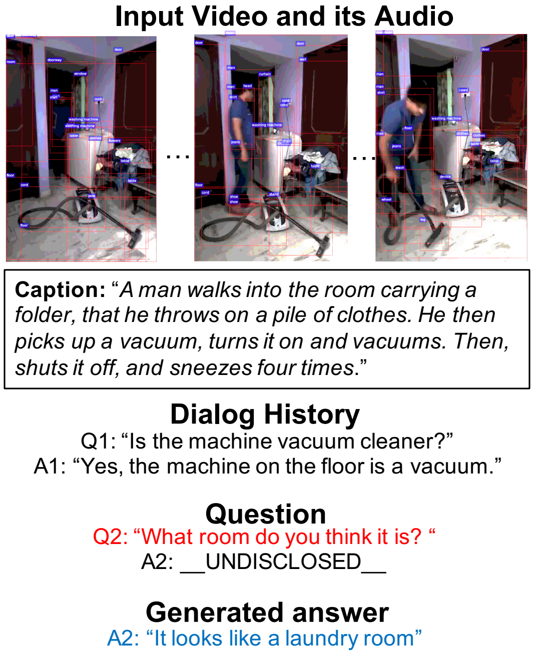
TABLE V: Results for the in-channel next-utterance selection task (Ubuntu).

Team	Exact Match	Precision	Recall	F1
3	80.0	**83.2**	**80.2**	**81.7**
13	66.2	70.7	66.2	68.4
15	**80.8**	**83.2**	**80.2**	**81.7**

TABLE VI: Results for the advising success task (Advising).

Team	P	R	F	VI	Rand	AMI
Baseline	36.3	39.7	38.0	0.915	0.650	0.837
3	**44.3**	**49.6**	**46.8**	**0.933**	**0.752**	**0.865**

TABLE VII: Results for the conversation disentanglement task (Ubuntu).



Caption: “*A man walks into the room carrying a folder, that he throws on a pile of clothes. He then picks up a vacuum, turns it on and vacuums. Then, shuts it off, and sneezes four times.*”

Q1: “Is the machine vacuum cleaner?”
A1: “Yes, the machine on the floor is a vacuum.”

Q2: “What room do you think it is? “
A2: __UNDISCLOSED__

A2: “It looks like a laundry room”

Fig. 2: A sample dialog for the AVSD challenge data set. Given a video clip (including audio), its caption, dialog history, and a follow-up question, the AVSD generation task aims to generate an answer to the question in natural language form.

B. Data and Baseline System

We collected (in [12]) text-based dialogs about short videos from the Charades dataset[5] [37], which consists of untrimmed multi-action videos along with a brief script for each video. The data collection paradigm for dialogs was introduced in [38]. We extracted the test data for DSTC8 from the collected data. In our audio visual scene-aware dialog setup, two parties had a discussion about events in a video. One of the two parties played the role of an answerer who had already watched the video and read the video script. The answerer answered questions asked by their counterpart, the questioner. The questioner was not allowed to watch the video but was able to see three frames of the video (the first, middle, and last frames) as static images. The two parties had 10 rounds of Q and A, in which the questioner asked about the events that happened in the video. At the end, the questioner summarized the events in the video as a video description. This downstream task incentivized the questioner to collect useful information for the video description. Table VIII shows the data statistics.

We provided the same baseline system in both DSTC7 and DSTC8. This baseline system and an additional submitted system featuring encoder-decoder models using multimodal fusion are described in [39]. The baseline system utilizes two state-of-the-art feature encoders, which are described in more detail below, to capture the information from the video: I3D [40] for visual information, and Audio Set VGGish [41] for audio. The VGGish model was trained to predict an ontology of more than 600 audio event classes from only the audio tracks of 2 million human-labeled 10-second YouTube video soundtracks [41]. The I3D features [40] are state-of-the-art spatiotemporal features that were developed for action recognition. The I3D model inflates the 2D filters and pooling kernels from the Inception V3 network along their temporal dimension, building 3D spatiotemporal ones. We pre-computed these features for all of the videos in the dataset (including the test videos), and we made them available to all challenge participants. Detailed results from all models on the DSTC7

TABLE VIII: The dialog data for the AVSD track. The test videos for this challenge were selected from the official test data of the Charades dataset.

	training	validation	DSTC7 test	DSTC8 test
# of dialogs	7,659	1,787	1,710	1,710
# of turns	153,180	35,740	13,490	18,810
# of words	1,450,754	339,006	110,252	178,619

[5] http://allenai.org/plato/charades/

challenge, including additional techniques and data set details, were reported in [42].

C. Evaluation

In both DSTC7 and DSTC8, each automatically generated answer was evaluated by comparing with 6 human-generated ground-truth answers: the answer from the original dialog plus 5 subsequently collected answers. We used the MS COCO evaluation tool[6] for objective evaluation of system outputs. The supported metrics include word-overlap-based metrics such as BLEU, METEOR, ROUGE_L, and CIDEr. We also collected human ratings of the responses of each system using a 5-point Likert Scale, where humans rated system responses given a dialog context as: 5 (very good), 4 (good), 3 (acceptable), 2 (poor), or 1 (very poor).

D. Outcomes from DSTC7

In the AVSD Challenge at DSTC7, Most systems employed an LSTM, Bidirectional LSTM, or GRU encoder/decoder. Some systems used hierarchical and attention frameworks. Furthermore, several additional techniques were introduced to improve the response quality, such as Maximum Mutual Information (MMI) and Episodic Memory Module [42]. The best system applied hierarchical attention mechanisms to combine text and visual information, yielding a relative improvement of 22% in human ratings compared to the baseline system. The language models trained using the text information alone (without video or audio) also performed strongly despite the lack of multimodal information. After the AVSD Challenge at DSTC7, [38] also reported results on the AVSD dataset (although instead of evaluating on the sentence generation task as in the AVSD challenge, that paper evaluated performance on a the task of sentence selection).

In general, results on the AVSD dataset show that including dialog history provided a large boost to performance as compared to only providing the one question to be answered. This makes sense, as dialogs are self-referential; in the AVSD dataset, 55% of the questions contain co-reference words such as *her*, *they*, and *it*. Such questions strongly depend on the prior rounds of dialog. Systems using text information alone (questions, answers, and dialog history) performed quite well. Nonetheless, adding the audio and video as inputs improved systems' performance further by providing complementary information to ground the questions.

Furthermore, the best performance is achieved when systems have access to the video script (caption). Using such manual descriptions improves the performance of all systems. However, such summaries are unavailable in the real world, posing challenges during deployment. Recently, [43] proposed an approach to transfer the power of a teacher model that was trained using summaries to a student model that does not have access to summaries at test time.

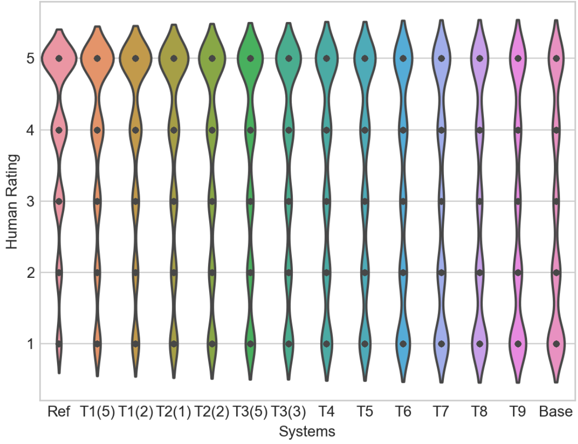


Fig. 3: Distribution of human rating scores across score values for each system. The figure shows each system's distribution of rating scores $(1, \ldots, 5)$ across all sentences and all raters. In this figure, the area of the violin plot for each score indicates the number of scores at each level on the Likert scale.

E. DSTC8 Results

In this section, we analyze the results from the AVSD challenge track at DSTC8 of each of the submitted systems, which are summarized in Table IX [44]. Most of the DSTC8 systems employed transformers, rather than recurrent networks using LSTM or GRU. The inclusion of transformers drastically improved performance on the AVSD task from DSTC7 to DSTC8, similar to transformer-powered improvements that have been observed in other applications such as machine translation and speech recognition and synthesis. Two of the most successful systems extracted semantic features of the word sequences by initializing network weights using a pre-trained model such as BERT or GPT-2, then fine-tuning on the AVSD dataset.

Figure 3 plots the human ratings for each team's best-scoring system. It is evident from the figure that the distribution of human rating scores across all systems appears to be bimodal—most answers are rated either highly (5) or poorly (1), with few examples in the middle. This is because the human ratings of each answer depend strongly on whether the answer is a correct response to the question: correct answers generally receive high human ratings, but incorrect answers receive low human ratings. The best systems generated mostly correct answers, while the worst systems generated more incorrect answers.

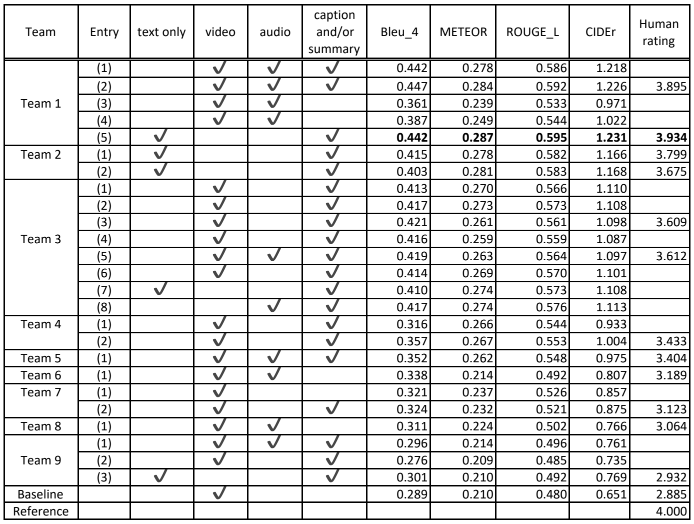
Table X presents the results (averaged across the test set) for each team's entries, using both word-overlap-based objective measures and subjective human ratings. Although the language-based transformer models such as BERT and GPT-2 demonstrate state-of-the-art performance on our tasks, these systems require features extracted from manually generated video captions (scripts), and such a text modality may be unavailable in real-world deployment scenarios. There are two other design difficulties that such text-based captions introduce that may skew the evaluation: (i) some captions already include parts of the answers that are used in the

[6] https://github.com/tylin/coco-caption

TABLE IX: Submitted systems to the AVSD Track. The individual system description papers contain more details about the systems.

Team	Encoder-decoder type	Multimodal fusion type	Features	Additional techniques/data
Baseline	LSTM	Context vector concatenation	I3D, VGGish	
Team 1	Transformer	Input sequence concatenation by universal multimodal transformer	I3D, VGGish,	Pre-trained GPT-2 model, fine-tuned on AVSD dataset using multi-task learning
Team 2	Transformer	Input sequence concatenation (text only)	I3D, VGGish	Pre-trained BERT model, fine-tuned on AVSD dataset
Team 3	Transformer	Fusion by multimodal transformer network	I3D, VGGish, ResNeXt	Pointer network and model ensemble
Team 4	Transformer	Semantically-controlled transformer with multi-head shuffled attention	ResNet-101	Spatio-temporal scene graph feature representation
Team 5	Transformer		I3D, VGGish	Multimodal semantic transformer
Team 6	Transformer	Hierarchical attention	I3D, VGGish	Pre-trained GPT
Team 7	LSTM	Multi-step joint-modality attention	I3D, VGGish	
Team 8	LSTM	Multimodal attention over frames across different modalities	VGG-19, VGGish	
Team 9	Dynamic memory network with GRU	Memory vector concatenation	I3D, VGGish	

TABLE X: Evaluation results for the AVSD track at DSTC8. Both word-overlap-based objective measures (based on 6 reference answers for each test question) and a subjective human rating measure (based on 5-level ratings) are shown.



Team	Entry	text only	video	audio	caption and/or summary	Bleu_4	METEOR	ROUGE_L	CIDEr	Human rating
Team 1	(1)		✓	✓	✓	0.442	0.278	0.586	1.218	
	(2)		✓	✓	✓	0.447	0.284	0.592	1.226	3.895
	(3)		✓	✓		0.361	0.239	0.533	0.971	
	(4)		✓	✓		0.387	0.249	0.544	1.022	
	(5)	✓			✓	**0.442**	**0.287**	**0.595**	**1.231**	**3.934**
Team 2	(1)	✓			✓	0.415	0.278	0.582	1.166	3.799
	(2)	✓			✓	0.403	0.281	0.583	1.168	3.675
Team 3	(1)		✓		✓	0.413	0.270	0.566	1.110	
	(2)		✓		✓	0.417	0.273	0.573	1.108	
	(3)		✓		✓	0.421	0.261	0.561	1.098	3.609
	(4)		✓		✓	0.416	0.259	0.559	1.087	
	(5)		✓	✓	✓	0.419	0.263	0.564	1.097	3.612
	(6)		✓		✓	0.414	0.269	0.570	1.101	
	(7)	✓			✓	0.410	0.274	0.573	1.108	
	(8)			✓	✓	0.417	0.274	0.576	1.113	
Team 4	(1)		✓		✓	0.316	0.266	0.544	0.933	
	(2)		✓		✓	0.357	0.267	0.553	1.004	3.433
Team 5	(1)		✓	✓	✓	0.352	0.262	0.548	0.975	3.404
Team 6	(1)		✓	✓		0.338	0.214	0.492	0.807	3.189
Team 7	(1)		✓			0.321	0.237	0.526	0.857	
	(2)		✓		✓	0.324	0.232	0.521	0.875	3.123
Team 8	(1)		✓	✓		0.311	0.224	0.502	0.766	3.064
Team 9	(1)		✓	✓	✓	0.296	0.214	0.496	0.761	
	(2)		✓		✓	0.276	0.209	0.485	0.735	
	(3)	✓			✓	0.301	0.210	0.492	0.769	2.932
Baseline			✓			0.289	0.210	0.480	0.651	2.885
Reference										4.000

evaluations, making audio-visual inference redundant, and (ii) language models trained using a simple (and limited) QA dataset may generate answers using frequently-occurring text patterns in the training data, without needing to use audio-visual cues (e.g., Q: How many people are in the scene? A: Two people). These observations are empirically supported by the results: without providing human-generated captions, the best performing model achieves only 0.387 in BLEU score, which is a relative reduction of 12% from its score when using human captions. This result suggests that there is still opportunity to design better audio-visual reasoning approaches to try to match the performance achieved using manually provided text captions.

F. Summary

We introduced a new challenge task and dataset for Audio Visual Scene-Aware Dialog (AVSD) in DSTC7, and we held a follow-up challenge in DSTC8. The participating teams built scene-aware dialog systems by combining end-to-end conversation models and end-to-end multimodal video description models into complete end-to-end differentiable systems. The DSTC8 winning system achieved an impressive 98.4% of human performance based on human ratings (a 9% improvement over DSTC7). The large performance improvement of this year's best systems was enabled by using transformers [45], including pre-trained GPT-2 and BERT models. It should be noted that in order to achieve its nearly human level of performance, the winning system used the human-generated video captions that were included in the dataset—it was not able to glean all of the necessary information directly from the video features, as would be required in a real-world, real-time interaction.

V. Schema-Guided Dialogue State Tracking Track

Today's virtual assistants such as the Google Assistant, Alexa, Siri, Cortana, etc. help users accomplish a wide variety of tasks including finding flights, searching for nearby events, surfacing information from the web etc. They provide this functionality by offering a unified natural language interface to a variety of services and APIs from the web. Building such large scale assistants offers many new challenges such as supporting a large variety of domains, data-efficient handling of APIs with similar functionality and reducing maintenance overhead for integration of new APIs among others. Despite tremendous progress in dialogue research, these critical challenges have not been sufficiently explored, owing to an absence of datasets matching the scale and complexity presented by virtual assistants. To this end, we created the Schema-Guided Dialogue (SGD) dataset, a large-scale corpus of over 18K multi-domain task-oriented conversations spanning 17 domains. This track explores the aforementioned challenges on this dataset, focusing on dialogue state tracking (DST).

A. Task definition

The dialogue state is a summary of the entire conversation till the current turn. In a task-oriented system, it is used to invoke APIs with appropriate parameters as specified by the user over the dialogue history. The state is also used by the assistant to

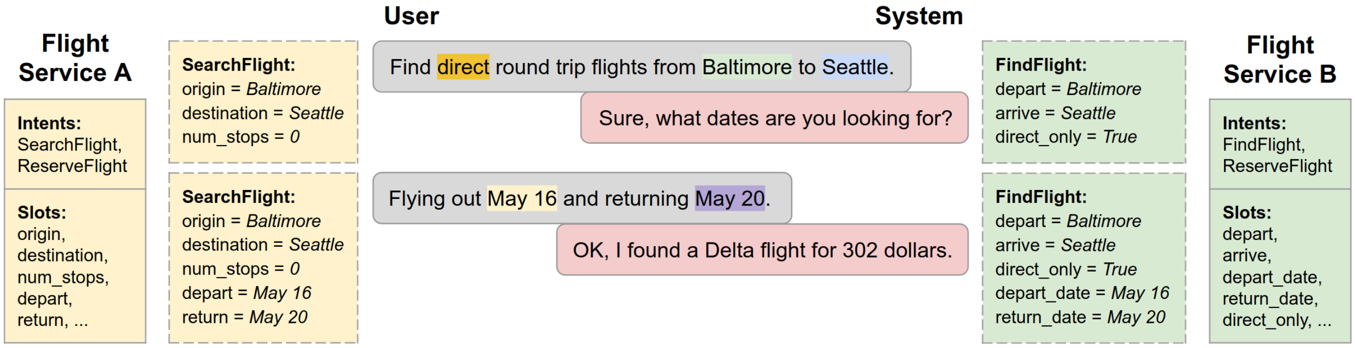

User
System
Flight Service A
Intents: SearchFlight, ReserveFlight
Slots: origin, destination, num_stops, depart, return, ...
SearchFlight: origin = Baltimore destination = Seattle num_stops = 0
SearchFlight: origin = Baltimore destination = Seattle num_stops = 0 depart = May 16 return = May 20
Find direct round trip flights from Baltimore to Seattle.
Sure, what dates are you looking for?
Flying out May 16 and returning May 20.
OK, I found a Delta flight for 302 dollars.
FindFlight: depart = Baltimore arrive = Seattle direct_only = True
FindFlight: depart = Baltimore arrive = Seattle direct_only = True depart_date = May 16 return_date = May 20
Flight Service B
Intents: FindFlight, ReserveFlight
Slots: depart, arrive, depart_date, return_date, direct_only, ...

Fig. 4: Illustration of Track 4: the dialogue state (dashed edges) for the same dialogue is conditioned on the domain/service schema under consideration (extreme left/right), provided as input.

generate the next actions to continue the dialogue. Dialogue State Tracking (DST), therefore, is a core component of virtual assistants. Deep learning-based approaches to DST have recently gained popularity. Some of these approaches estimate the dialogue state as a distribution over all possible slot values [46], [15] or individually score all slot-value combinations [47], [48]. Such approaches are, however, hard to scale to real-world virtual assistants, where the set of possible values for certain slots may be very large (date, time or restaurant name) and even dynamic (movie or event name). Other approaches utilizing a dynamic vocabulary of slot values [49], [50] still preclude zero-shot generalization to new services and APIs [51], since they use schema elements i.e. intents and slots as class labels.

The primary task of this challenge is to develop multi-domain models for DST with particular emphasis on joint modeling across different services or APIs (for data-efficiency) and zero-shot generalization (for handling new/unseen APIs). This takes the shape of a DST task where the dialogue state annotations are guided by the APIs under consideration. Figure 4 illustrates how the dialogue state representations can be conditioned on the corresponding schema for two different flight services (extreme left and right). In order to generate these schema-guided dialogue state representations, the systems are required to take the relevant schemas as additional inputs. The systems can also utilize the natural language descriptions of slots and intents supported by the APIs to yield distributed semantic representations, which can help in joint modeling of related concepts and generalization to new APIs. In addition, the participants are allowed to use any external datasets or resources to bootstrap their models.

B. Data and Baseline

The SGD dataset[7] consists of over 18K annotated multi-domain task-oriented conversations between a human and a virtual assistant. These conversations involve interactions with services/APIs spanning 17 domains (see Table XI). For most of these domains, SGD contains multiple APIs having overlapping functionalities but different interfaces - common in the real world; it is the first dataset set up this way. The schemas for all services/APIs pertinent to a dialogue, as well as natural language descriptions and other semantic features for a service and its intents and slots, are also included in the dataset. [52] contains more details about the dataset and the data collection methodology.

With annotations for slot spans, intent, dialogue state and system actions, our dataset is designed to serve as an effective testbed for intent prediction, slot filling, state tracking and language generation, among other tasks in large-scale virtual assistants. Furthermore, the evaluation set is tailored to contain many new services not present in the training set. This helps to quantify the robustness to changes in an API's interface or the addition of new APIs.

We also provide a baseline system [52], using user and system utterances and schema element descriptions as inputs to a model based on BERT [19]. The baseline model extends BERT-DST [53] by removing all domain-specific parameters, accomplishing zero-shot generalization to new APIs.

C. Evaluation

Joint goal accuracy, defined as the fraction of dialogue turns for which all slot values across all domains in the dialogue state are correctly predicted, is a popular metric for DST evaluation. We use it as our primary metric for comparison of different approaches, with two modifications in its definition. First, we use a fuzzy matching score for non-categorical slots (i.e. slots with large or unbounded sets of possible values) to reward partial matches, drawing from metrics used for slot tagging in spoken language understanding. Second, instead of including all services in the dialogue state, only the services which are active or pertinent in a turn are included. Thus, a service ceases to be a part of the dialogue state once its intent has been fulfilled. This is done because of the presence of a large number of services in our dataset. Including all services in the joint goal accuracy evaluation would result in near zero value if the traditional definition is used, reducing the insight into the performance on different services we may glean.

For a better understanding of the underlying models, we evaluated the submissions on other auxiliary metrics such as:

- **Active Intent Accuracy:** The fraction of user turns for which the active intent is predicted correctly.

[7] https://github.com/google-research-datasets/dstc8-schema-guided-dialogue

TABLE XI: The number of intents (services in parentheses) and dialogues per domain in the train and dev sets for Track 4. Multi-domain dialogues contribute to counts of each domain.

Domain	**#Intents**	**#Dialogs**	**Domain**	**#Intents**	**#Dialogs**	**Domain**	**#Intents**	**#Dialogs**
Alarm	2 (1)	37	Home	2 (1)	1027	Restaurant	4 (2)	2755
Bank	4 (2)	1021	Hotel	8 (4)	3930	RideShare	2 (2)	1973
Bus	4 (2)	2609	Media	4 (2)	1292	Service	8 (4)	2090
Calendar	3 (1)	1602	Movie	4 (2)	1758	Travel	1 (1)	2154
Event	5 (2)	3927	Music	4 (2)	1486	Weather	1 (1)	1308
Flight	8 (3)	3138	RentalCar	4 (2)	1966			

TABLE XII: Evaluation Results for Schema-Guided State Tracking track for the baseline and the top 3 submissions

Team	**Joint Goal Accuracy**			**Avg Goal Accuracy**			**Active Intent Accuracy**			**Requested Slots F1**		
	All	Seen	Unseen	All	Seen	Unseen	All	Seen	Unseen	All	Seen	Unseen
Baseline	0.254	0.413	0.200	0.560	0.678	0.519	0.906	0.951	0.892	0.965	0.996	0.955
Team 9	0.865	0.924	0.846	0.971	0.980	0.967	0.948	0.957	0.945	0.985	0.994	0.982
Team 14	0.773	0.991	0.730	0.922	0.961	0.908	0.969	0.958	0.971	0.995	0.996	0.992
Team 12	0.738	0.880	0.690	0.920	0.957	0.907	0.926	0.958	0.912	0.995	0.997	0.994

- **Requested Slot F1:** The macro-averaged F1 score for slots requested by the user over all valid turns.
- **Average Goal Accuracy:** The average accuracy of predicting the slot assignments for a turn correctly. Like the joint goal accuracy, this also uses a fuzzy matching score for non-categorical slots. In addition, we discard instances when both the ground truth and the predicted values for a slot are empty since, if naively evaluated, models can achieve a relatively high average goal accuracy just by predicting an empty assignment for each slot.

D. Results

We received submissions from 25 teams. Table XII lists the results for the top 3 teams (determined by joint goal accuracy) and the baseline system. The test set contains a total of 21 services, among which 6 services are also present in the training set (seen services), whereas the remaining 15 are not present in the training set (unseen services). Among these 15 unseen services are three entirely new domains - “Messaging", “Payment" and “Trains", the other unseen APIs being from domains present in training and dev sets. We observe that the submitted models are able to generalize well to new APIs and domains - partly attributable to the use of pre-trained models like BERT [19], XLNet [54] in most submissions.

Our most patent observation from the results is the higher joint goal accuracy metric than reported on other public datasets. This is because our dataset excludes the slots for APIs not under consideration in the current turn from the dialogue state for multi-domain dialogues, as opposed to other datasets which include slots for all domains and APIs present over the dialogue history. Thus, in our setup, an incorrect dialogue state prediction for a service only penalizes the joint goal accuracy metric for the turns in which that service is under consideration by the user or the system. Further, our fuzzy matching score rewards partial matches for non-categorical slots, leading to still higher joint and average goal accuracy values. The following trends were observed across all submissions:

- For unseen services, performance on categorical slots is comparable to that on non-categorical slots. On the other hand, for seen services, the performance on categorical slots is better. This could be because there is less signal to differentiate between the different possible values for a categorical slot when they are not seen during training.
- The winning team's performance on seen services is similar to that of the other top teams. However, the winning team has a considerable edge on unseen services, outperforming the second team by around 12% in terms of joint goal accuracy. This margin was observed across both categorical and non-categorical slots.
- Among unseen services, when looking at services belonging to unseen domains, the winning team was ahead of the other teams by at least 15%. The performance on categorical slots for unseen domains was about the same as that for seen services and domains. For other teams, there was at least a 20% drop in accuracy of categorical slots in unseen domains vs seen domains and services.
- The joint goal accuracy of most of the models was worse by 15 percentage points on an average on the test set as compared to the dev set. This could be because the test set contains a much higher proportion of turns with at least one unseen services as compared to the dev set (77% and 45% respectively).

VI. Conclusions

This paper summarizes the four tracks of the eighth dialog system technology challenges (DSTC8). Multi-domain task-completion track offered two sub-tasks: end-to-end multi-domain dialog task and fast adaptation task. NOESIS II track extended the response selection task of DSTC7 with new datasets with multi-party dialogs and two additional subtasks. Audio visual scene-aware dialog track explored further improvements from its first edition on DSTC7 with a new test dataset. Schema-guided dialog state tracking track introduced a new dialog state tracking task from a practical perspective.

From the evaluation results, we've got a common observation that Transformer-based large-scale pre-trained language models helped to achieve the state-of-the-art performances on all the challenge tasks. This is a significant difference from what we learned from previous DSTCs where most winning entries were based on RNN variants trained from scratch or minimal pre-training mostly for word embeddings only. This transition shows a recent trend in dialog research.

This challenge also leaves an open question about how we can make these benchmark results more realistic, reproducible and accumulable over time. Especially for the generation tasks, we've observed some limitations of the automated metrics with the gaps from the end-to-end human evaluation results. On the other hand, the human evaluations are too expensive to make it scalable and relatively less reproducible compared to the conventional corpus-based evaluation methods. We expect to address these issues further in the future challenges.

References

[1] J. Williams, A. Raux, D. Ramachandran, and A. Black, "The dialog state tracking challenge," in *Proceedings of the SIGDIAL 2013 Conference*, 2013, pp. 404–413.

[2] M. Henderson, B. Thomson, and J. Williams, "The second dialog state tracking challenge," in *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2014, p. 263.

[3] M. Henderson, B. Thomson, and J. D. Williams, "The third dialog state tracking challenge," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 324–329.

[4] S. Kim, L. F. D'Haro, R. E. Banchs, J. D. Williams, and M. Henderson, "The fourth dialog state tracking challenge," in *Dialogues with Social Robots*. Springer, 2017, pp. 435–449.

[5] S. Kim, L. F. D'Haro, R. E. Banchs, J. D. Williams, M. Henderson, and K. Yoshino, "The fifth dialog state tracking challenge," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 511–517.

[6] C. Hori, J. Perez, R. Higashinaka, T. Hori, Y.-L. Boureau, M. Inaba, Y. Tsunomori, T. Takahashi, K. Yoshino, and S. Kim, "Overview of the sixth dialog system technology challenge: Dstc6," *Computer Speech & Language*, vol. 55, pp. 1–25, 2019.

[7] K. Yoshino, C. Hori, J. Perez, L. F. D'Haro, L. Polymenakos, C. Gunasekara, W. S. Lasecki, J. K. Kummerfeld, M. Galley, C. Brockett *et al.*, "Dialog system technology challenge 7," *arXiv preprint arXiv:1901.03461*, 2019.

[8] L. F. D'Haro, K. Yoshino, C. Hori, T. K. Marks, L. Polymenakos, J. K. Kummerfeld, M. Galley, and X. Gao, "Overview of the seventh dialog system technology challenge: Dstc7," *Computer Speech & Language*, vol. 62, p. 101068, 2020.

[9] C. Gunasekara, J. K. Kummerfeld, L. Polymenakos, , and W. S. Lasecki, "Dstc7 task 1: Noetic end-to-end response selection," in *7th Edition of the Dialog System Technology Challenges at AAAI 2019*, January 2019. [Online]. Available: http://workshop.colips.org/dstc7/papers/dstc7_task1_final_report.pdf

[10] C. Gunasekara, J. K. Kummerfeld, L. Polymenakos, and W. Lasecki, "Dstc7 task 1: Noetic end-to-end response selection," in *Proceedings of the First Workshop on NLP for Conversational AI*, 2019, pp. 60–67.

[11] M. Galley, C. Brockett, X. Gao, J. Gao, and B. Dolan, "Grounded response generation task at dstc7," in *AAAI Dialog System Technology Challenges Workshop*, 2019.

[12] H. Alamri, C. Hori, T. K. Marks, D. Batr, and D. Parikh, "Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7," in *DSTC7 at AAAI2019 Workshop*, vol. 2, 2018.

[13] S. Lee, Q. Zhu, R. Takanobu, Z. Zhang, Y. Zhang, X. Li, J. Li, B. Peng, X. Li, M. Huang, and J. Gao, "ConvLab: Multi-domain end-to-end dialog system platform," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 64–69. [Online]. Available: https://www.aclweb.org/anthology/P19-3011

[14] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational ai," *Foundations and Trends® in Information Retrieval*, vol. 13, no. 2-3, pp. 127–298, 2019. [Online]. Available: http://dx.doi.org/10.1561/1500000074

[15] T. Wen, D. Vandyke, N. Mrkšíc, M. Gašíc, L. Rojas-Barahona, P. Su, S. Ultes, and S. Young, "A network-based end-to-end trainable task-oriented dialogue system," in *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017-Proceedings of Conference*, vol. 1, 2017, pp. 438–449.

[16] W. Lei, X. Jin, M.-Y. Kan, Z. Ren, X. He, and D. Yin, "Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1437–1447.

[17] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gasic, "Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 5016–5026.

[18] J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, and S. Young, "Agenda-based user simulation for bootstrapping a pomdp dialogue system," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. Association for Computational Linguistics, 2007, pp. 149–152.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[20] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019.

[21] O. Vinyals and Q. V. Le, "A neural conversational model," *arXiv:1506.05869*, 2015.

[22] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky, "Deep Reinforcement Learning for Dialogue Generation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016.

[23] T.-H. Wen, Y. Miao, P. Blunsom, and S. Young, "Latent Intention Dialogue Models," in *Proceedings of the International Conference on Machine Learning*, 2017.

[24] L. Maystre and M. Grossglauser, "Just sort it! a simple and effective approach to active preference learning," in *International Conference on Machine Learning (ICML)*, 2017.

[25] A. H. Copeland, "A 'reasonable' social welfare function," in *Seminar on Mathematics in Social Sciences*. University of Michigan, 1951.

[26] P. Hall, H. Miller *et al.*, "Using the bootstrap to quantify the authority of an empirical ranking," *The Annals of Statistics*, vol. 37, no. 6B, pp. 3929–3959, 2009.

[27] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[28] C. Callison-Burch, P. Koehn, C. Monz, and O. F. Zaidan, "Findings of the 2011 workshop on statistical machine translation," in *Proc. of the Workshop on Statistical Machine Translation*, 2011.

[29] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext.zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016.

[30] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, 1997.

[33] J. K. Kummerfeld, S. R. Gouravajhala, J. J. Peper, V. Athreya, C. Gunasekara, J. Ganhotra, S. S. Patel, L. Polymenakos, and W. S. Lasecki, "A large-scale corpus for conversation disentanglement," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July 2019, pp. 3846–3856. [Online]. Available: https://www.aclweb.org/anthology/P19-1374

[34] M. Meila, "Comparing clusterings–an information based distance," *Journal of Multivariate Analysis*, vol. 98, no. 5, pp. 873–895, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0047259X06002016

[35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[36] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, "Enhanced lstm for natural language inference," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1657–1668.

[37] G. A. Sigurdsson, G. Varol, X. Wang, I. Laptev, A. Farhadi, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," *ArXiv*, 2016. [Online]. Available: http://arxiv.org/abs/1604.01753

[38] H. Alamri, V. Cartillier, A. Das, J. Wang, A. Cherian, I. Essa, D. Batra, T. K. Marks, C. Hori, P. Anderson, S. Lee, and D. Parikh, "Audio visual scene-aware dialog," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[39] C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. G. Lopes, A. Das *et al.*, "End-to-end audio visual scene-aware dialog using multimodal attention-based video features," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2352–2356.

[40] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.

[41] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *ICASSP*, 2017.

[42] H. Alamri, C. Hori, T. K. Marks, D. Batra, and D. Parikh, "Track 3 overview: Audio visual scene-aware dialog (AVSD) track for natural language generation in dstc7," in *AAAI 2019 Workshop: DSTC7*, 2019, http://workshop.colips.org/dstc7/workshop.html.

[43] C. Hori, T. Hori, A. Cherian, and T. K. Marks, "Joint student-teacher learning for audio-visual scene-aware dialog," in *Interspeech 2019*. ISCA, 2019.

[44] C. Hori, A. Cherian, T. Hori, and T. K. Marks, "Audio visual scene-aware dialog (AVSD) track for natural language generation in DSTC8," in *The Eighth Dialog System Technology Challenge (DSTC8) at the 34th AAAI conference on Artificial Intelligence (AAAI)*, 2020.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[46] M. Henderson, B. Thomson, and S. Young, "Word-based dialog state tracking with recurrent neural networks," in *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2014, pp. 292–299.

[47] N. Mrkšić, D. Ó. Séaghdha, T.-H. Wen, B. Thomson, and S. Young, "Neural belief tracker: Data-driven dialogue state tracking," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2017, pp. 1777–1788.

[48] V. Zhong, C. Xiong, and R. Socher, "Global-locally self-attentive encoder for dialogue state tracking," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1458–1467. [Online]. Available: https://www.aclweb.org/anthology/P18-1135

[49] A. Rastogi, R. Gupta, and D. Hakkani-Tur, "Multi-task learning for joint language understanding and dialogue state tracking," in *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, 2018, pp. 376–384.

[50] R. Goel, S. Paul, and D. Hakkani-Tür, "Hyst: A hybrid approach for flexible and accurate dialogue state tracking," *arXiv preprint arXiv:1907.00883*, 2019.

[51] C.-S. Wu, A. Madotto, E. Hosseini-Asl, C. Xiong, R. Socher, and P. Fung, "Transferable multi-domain state generator for task-oriented dialogue systems," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 808–819. [Online]. Available: https://www.aclweb.org/anthology/P19-1078

[52] A. Rastogi, X. Zang, S. Sunkara, R. Gupta, and P. Khaitan, "Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset," *arXiv preprint arXiv:1909.05855*, 2019.

[53] G.-L. Chao and I. Lane, "Bert-dst: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer," *arXiv preprint arXiv:1907.03040*, 2019.

[54] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237*, 2019.