MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Overview of the Tenth Dialog System Technology Challenge: DSTC10

Yoshino, Koichiro; Chen, Yun-Nung; Crook, Paul; Kottur, Satwik; Li, Jinchao; Hedayatnia, Behnam; Moon, Seungwhan; Fe, Zhengcong; Li, Zekang; Zhang, Jinchao; Fen, Yang; Zhou, Jie; Kim, Seokhwan; Liu, Yang; Jin, Di; Papangelis, Alexandros; Gopalakrishnan, Karthik; Hakkani-Tur, Dilek; Damavandi, Babak; Geramifard, Alborz; Hori, Chiori; Shah, Ankit; Zhang, Chen; Li, Haizhou; Sedoc, João; D'Haro, Luis F.; Banchs, Rafael; Rudnicky, Alexander

TR2023-109 September 02, 2023

Abstract

This paper introduces the Tenth Dialog System Technology Challenge (DSTC-10). This edition of the DSTC focuses on applying end-to-end dialog technologies for five distinct tasks in dialog systems, namely 1. Incorporation of Meme images into open domain dialogs, 2. Knowledge-grounded Task- oriented Dialogue Modeling on Spoken Conversations, 3. Situated Interactive Multimodal dialogs, 4. Reasoning for Audio Visual Scene-Aware Dialog, and 5. Automatic Evaluation and Moderation of Open-domain Dialogue Systems. This paper describes the task definition, provided datasets, baselines, and evaluation setup for each track. We also summarize the results of the submitted systems to highlight the general trends of the state-of-the-art technologies for the tasks.

IEE/ACM Transactions on Audio, Speech, and Language Processing 2023

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Overview of the Tenth Dialog System Technology Challenge: DSTC10

Koichiro Yoshino, Yun-Nung Chen, Paul Crook, Satwik Kottur, Jinchao Li, Behnam Hedayatnia, Seungwhan Moon, Zhengcong Fei, Zekang Li, Jinchao Zhang, Yang Feng, Jie Zhou, Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Dilek Hakkani-Tur, Babak Damavandi, Alborz Geramifard, Chiori Hori, Ankit Shah, Chen Zhang, Haizhou Li, João Sedoc, Luis F. D'Haro, Rafael Banchs, Alexander Rudnicky

***Abstract*—This paper introduces the Tenth Dialog System Technology Challenge (DSTC-10). This edition of the DSTC focuses on applying end-to-end dialog technologies for five distinct tasks in dialog systems, namely 1. Incorporation of Meme images into open domain dialogs, 2. Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations, 3. Situated Interactive Multimodal dialogs, 4. Reasoning for Audio Visual Scene-Aware Dialog, and 5. Automatic Evaluation and Moderation of Open-domain Dialogue Systems. This paper describes the task definition, provided datasets, baselines, and evaluation setup for each track. We also summarize the results of the submitted systems to highlight the general trends of the state-of-the-art technologies for the tasks.**

I. Introduction

The Dialog System Technology Challenge (DSTC) is one of the leading series of research competitions in the space of dialog systems. Since the inception in 2013, DSTC has been accelerating the development of dialog technologies by bringing the leading researchers and engineers together to solve important problems in dialog systems. The challenge has been evolving every year to cater to the demand and the interest of the dialog community to foster the development of technology.

The first version of the challenge (initially called Dialog State Tracking Challenge) [1] used human-to-bot dialogs in the bus timetable domain. Dialog State Tracking Challenges 2 [2] and 3 [3] used restaurant reservation applications, which introduced more complicated and dynamic dialog states. Dialog State Tracking Challenge 4 [4] and Dialog State Tracking Challenge 5 [5] moved to tracking human-to-human dialogs in mono and cross-language settings. From the sixth challenge [6], the DSTC rebranded itself as "Dialog System Technology Challenge" and organized multiple tracks in parallel to address a wider variety of dialog-related problems. The tracks in DSTC-6 were focused on end-to-end conversation modeling and dialog breakdown detection. DSTC-7 [7] focused on developing end-to-end dialog technologies for noetic response selection [8], grounded response generation [9], and audio visual scene aware dialog [10]. Then, in DSTC-8 [11], the focus was on a diverse set of four tracks that included multidomain task completion, predicting responses, audio-visual scene-aware dialog, and schema-guided dialog state tracking. More recently, DSTC-9 [12] focused on unstructured knowledge access in dialogue systems, multi-domain task-oriented dialogs, dialog evaluation, and situated multi-modal dialog modeling.

For the tenth edition, we received track proposals from leading research organizations and top universities. The proposals went through a formal peer review process focusing on each task's potential for (a) impact on the community, (b) novelty of the task, (c) feasibility of the proposal, and (d) potential participants. Participants in previous DSTC editions were also asked to provide their feedback on the presented track proposals through a survey, and responses were also considered in the evaluation. Finally, we ended up with five main tracks, including two newly introduced tasks and three follow-up and extensive tasks from previous challenges.

Track 1, *MOD: Internet Meme Incorporated Open-Domain Dialog*, aims to incorporate contextualized internet memes into multi-turn open dialogues. Track 2, *Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations*, focuses on benchmarking the robustness of the conversational models against the gaps between written and spoken conversations, where it extends the last-year challenge about unstructured knowledge access in task-oriented dialogues. Track 3 of this year, *SIMMC 2.0: Situated Interactive Multimodal Conversational AI*, is a continuation of last year, aimed at laying the foundations for the real-world assistant agents that can handle multi-modal inputs and perform multi-modal actions. Track 4, *Reasoning for Audio Visual Scene-Aware Dialog*, aims to promote the combination of conversation systems and multimodal reasoning algorithms into a single framework, where the system needs to learn to produce the answers without the captions of videos. Finally, *Automatic Evaluation and Moderation of Open-domain Dialogue Systems* (track 5) mainly focuses on developing effective automatic evaluation metrics that perform robustly across a range of dialogue evaluation tasks. The following sections describe the details of each track.

II. Track 1 - MOD: Internet Meme Incorporated Open-domain Dialog

A. Track Overview

Internet memes have become one of the most important approaches for expression and emotions in social media and messaging communication [13], [14], [15]. Meme, which is a type of content that features a visual format of images, GIF, or short videos, can inject humor into conversations and

Every author has equal contribution.

TABLE I: Summary of DSTC10 Track 1 tasks

Task #1	**Text Response Modeling**
Goal	To generate a coherent and natural text response given the multi-modal history context
Input	Multi-modal dialogue history $(u_1, u_2, ..., u_{t-1})$, where $u_i = (S_i, m_i)$, and S_i represent text-only response and m_i represents suitable meme id
Output	Natural, fluent, and informative machine text response S_t in line with dialogue history
Task #2	**Meme Retrieval**
Goal	To select a suitable internet meme from candidates given the multi-modal history context and generated text response
Input	Multi-modal dialogue history $(u_1, u_2, ..., u_{t-1})$ and generated text response S_t, where $u_i = (S_i, m_i)$, and S_i represent text-only response and m_i represents suitable meme id
Output	Suitable and vivid internet meme m_t in line with dialogue history
Task #3	**Meme Emotion Classification**
Goal	To predict the emotion type when respond with an internet meme
Input	Multi-modal dialogue history $(u_1, u_2, ..., u_t)$, where $u_i = (S_i, m_i)$, and S_i represent text-only response and m_i represents suitable meme id
Output	Emotion type c_t for current meme usage

create an emotional context [16]. Compared to emojis which is limited in variety and size, memes are more expressive and engaging. Although there is an increasing interest for chatbots that can converse using multiple modalities with humans [17], [18], incorporating contextualized internet memes into multi-turn open dialogues under different situations is still under explored. This challenge aims to deal with a new task – Meme incorporated Open Dialogue (MOD), where models are required to generate a vivid response in text-only, meme-only, or mixed information, provided with a multimodal dialogue context. There are three main tasks as introduced in [19]: text response modeling, meme retrieval, and meme emotion classification, as listed in Table I. The data and baseline system are publicly available [1].

B. Task and Data

1) Meme incorporated Open-domain Dialogue task: Participants are expected to build multi-modal dialogue systems based on the MOD dataset. Provided with the dialogue history consisting of utterances filled with Internet memes, the dialogue system aims to build an interesting response in the form of text-only, meme-only, or a mixed category of both.

We further split the current scope of MOD into the following three tasks as shown in Table I: (1) **Text Response Modeling**: given the multimodal history context, the task aims to generate a coherent and natural text response. (2) **Meme Retrieval**: given a multimodal historical context and a generated text response, the goal here is to select a suitable meme as feedback. (3) **Meme Emotion Classification**: given the multimodal history, the goal is to predict the emotion type when responding with an internet meme.

2) Data Collection:

a) Step 1: Pre-processing: For Internet meme sets, the meme candidates are firstly collected from the Internet and then chosen carefully by annotators to maintain good quality. In addition, if textual information appears in the selected Internet meme content, we will also annotate it manually. To avoid the model only utilizing the textual information and ignoring visual features, we control the proportion of memes without appeared texts in the final set to 40%. Meanwhile, to avoid multiple appropriate memes being selected under one dialogue condition, we filter out the memes with highly similar or duplicate semantic content. Finally, we obtain a total of 307 Internet memes for the subsequent data annotating process. To facilitate the arrangement and annotating process, the Internet meme set is further split into four groups: *atmosphere adjustment, basic expression, basic emotion,* and *common semantics*, respectively.

b) Step 2: Internet meme incorporated response construction: The annotators, who are well-educated and familiar with dialogue research, are tasked to take two operations using the prepared Internet meme candidates: use one most suitable Internet meme to replace part of the text conversation or insert an Internet meme into the utterance to enhance the emotion of the current dialogues. In particular, we also ask annotators to label the emotional states when utilizing the current Internet memes. The annotators are specially instructed based on the following criteria: (i) behave naturally, and the meme usage is in line with real daily chats, (ii) the number of different Internet memes in the dataset is kept balanced to avoid meaningless gatherings and biased data.

c) Step 3: Quality control: Before formal annotation, annotators are asked to annotate training samples until their results pass our examination. During the annotation, to eliminate the subjective inconsistency and make the annotation reliable, several specialized workers consistently monitor the collected dialogue data and perform a periodic quality check on samples. After the checking, we sample 10% data and manually check the samples ourselves.

d) Dataset Statistics: The total detailed statistics of the MOD dataset are summarized in Table II. MOD dataset has an average of 13.93 turns, and each turn contains 11.6 tokens. The text is tokenized by Chinese BERT tokenizer [20] and the vocabulary size is 13,086. We also plot the usage frequency of Internet memes and corresponding emotion in Figures 1 and 2, respectively. Although the dialogue system is evaluated under MOD, participants can leverage any public datasets and pre-trained models to build models. In the evaluation phase, we released the test set that is divided into easy test version for all internet meme seen in the training set and hard test version for some unseen internet memes.

TABLE II: Statistics of the Track 1 data sets

Split	**# dialogs**	**# sentences**	**# memes**
Train	66,219	927,331	274
Valid	1,000	13,666	274
Easy test	1,000	10,398	274
Hard test	2,183	29,046	307

C. Evaluation Criteria

Each participating team submitted up to five system outputs each of which contains the results for all three tasks on the two unlabeled test sets. We first evaluated each submission using

[1]https://github.com/lizekang/DSTC10-MOD

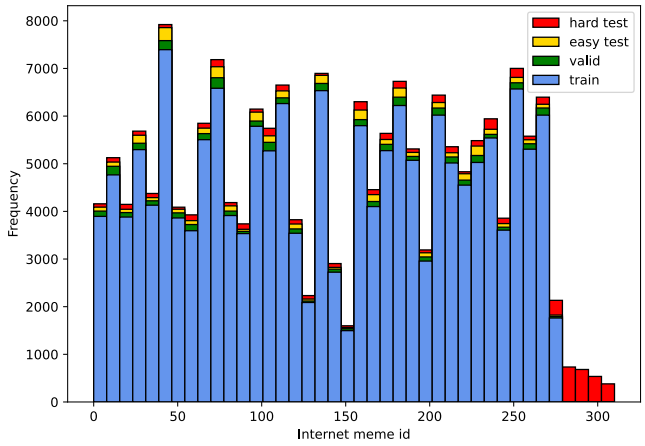


Fig. 1: Internet meme frequency in the Track 1 dataset. The meme usage balances without significant bias. Meme ids greater than 274 only occur in hard test set.

TABLE III: Evaluation metrics of the best entry from each team for the Track 1 tasks

Task	Evaluation Metrics
Task #1	BLEU, Dist Human Evaluation
Task #2	Recall$_n$@k, MAP
Task #3	Accuracy@k

the automatic task-specific objective metrics as show in Table III by comparing to the ground-truth labels and responses.

Considering the limitation of text response evaluation metrics, we selected the top-3 finalists based on the metric score to be manually evaluated for task #1, following the four aspects as:

- *Correctness*: whether there are grammatical errors in the machine generated text response.
- *Relevance*: whether the generated text response related to the historical content of the conversation.
- *Fluency*: whether the generated response is natural and smooth, in line with persons' conversation habits.
- *Informativeness*: whether the generated text response contains sufficient information. General replies are considered to be missing valid information.

Besides, we also required annotators to give an *overall score* based on the above four aspects. All of the scores are ranged from 1 to 5 with integers. The annotated data is randomly chosen from the submitted entries of each team, 2000 history-answer pairs for easy and hard test, respectively.

D. Results and Analysis

Generally, we received 22 entries in total submitted from 5 participating teams, setting a new state-of-the-art in all three subtasks. To preserve anonymity, the teams were identified by numbers from 1 to 5, while our baseline [19] was listed as team 0. Table IV presents the evaluation results of the best entry from each team in the automatic metrics for different tasks.

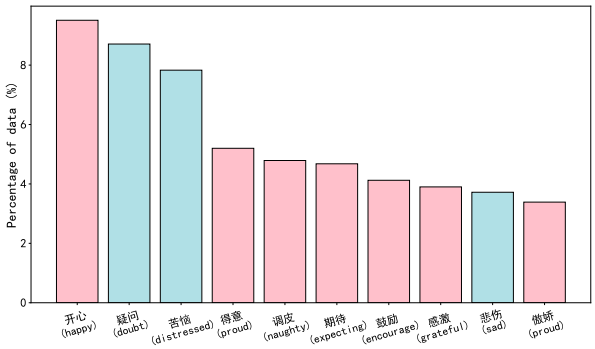


Fig. 2: Histogram of top-10 annotated emotions when memes are used in Track 1. Positive emotions (pink) occur significantly more often than negative emotions (blue).

a) For Task #1: Text Response Modeling: Table V presents the human evaluation results of the task #1 participating teams. We can find that team 1 wins the text response modeling task in easy test set while team 3 achieves the first in the hard test set. Team 1 focuses more on correctness and relevance while Team 3 gains the highest scores in fluency and informativeness. The big gap in automatic evaluation and relatively small gap in human evaluation between teams show that the automatic metrics are not reliable for the open-domain dialogue.

b) For Task #2: Meme Retrieval: Team 3 achieves over 90% Recall$_{10}$@5 in the easy test set, and also the highest scores in the hard test set. They treat the meme retrieval task as a matching problem and employ the cross-encoder architecture for relevance estimation using negative sampling. The big gap in performance between easy and hard tests also reveals that the generalization ability is limited for meme retrieval.

c) For Task #3: Meme Emotion Classification: Team 3 achieves the highest score of 89.5% in the easy test and 49.9% in the hard test. In particular, they devise an auxiliary method called Emotion-Enhanced Masked LM to improve the ability of meme emotion recognition. Meantime, Team 2 integrated historical memes and constructed a good-quality candidate set to reduce the difficulty of model learning and advance multimodal content understanding. There is also a big gap between easy and hard tests.

E. Conclusion

In this section, we describe the task definition, provided datasets, and evaluation set-up for DSTC10-MOD tracks. The top systems are all built with transformer-based end-to-end learning and follow the pre-training and fine-tuning paradigm. The incorporation of extra data for contrastive learning can effectively improve the robustness and generalization of the model. Well-designed self-supervised tasks can boost the multi-modal information fusion and understanding of the system. Although there is a lot of advancement compared with the baseline, we believe that the MOD task is worth further exploring and can benefit the modeling of multi-modal open-domain dialogue intelligence in the future, especially in how to exploit the visual features of memes better.

TABLE IV: Evaluation results of Track 1. Bold denotes the best results in each column and * indicates the finalists.

	Task #1				Task #2				Task #3		
Team	**B-2**	**B-4**	**D-1**	**D-2**	**R_{10}@1**	**R_{10}@3**	**R_{10}@5**	**MAP**	**A@1**	**A@3**	**A@5**
Easy test											
0	3.35	1.31	1.72	18.2	31.2	54.8	72.1	49.2	53.7	70.1	73.6
1*	**5.08**	**4.25**	1.90	**26.7**	34.2	59.6	76.0	52.3	-	-	-
2*	3.78	1.89	**2.20**	20.2	34.4	60.4	76.5	52.3	58.3	74.3	78.9
3*	3.57	1.32	1.93	21.5	**56.8**	**84.7**	**94.4**	**72.0**	**62.3**	**83.4**	**89.5**
4	3.70	1.65	2.00	19.5	33.5	58.0	75.3	50.3	54.5	70.8	74.5
5	3.54	1.40	1.85	20.3	34.0	56.8	74.5	51.0	57.3	72.0	77.6
Hard test											
0	3.15	1.22	1.05	15.8	25.5	49.2	64.0	40.5	25.6	36.0	45.5
1*	**5.04**	**3.65**	1.10	**20.0**	26.8	50.6	65.5	42.3	-	-	-
2*	4.03	1.65	**1.36**	16.6	27.9	50.8	66.7	45.1	**29.7**	**40.6**	**49.9**
3*	3.65	1.30	1.17	17.7	**42.0**	**69.7**	**80.9**	**58.8**	27.3	39.2	47.5
4	3.52	1.35	1.00	16.8	27.5	51.0	67.6	50.3	26.5	36.8	46.5
5	3.30	1.26	1.23	15.6	25.7	49.5	63.8	40.9	27.0	37.6	47.2

TABLE V: Human evaluation results for the Track 1 task #1

Team	**Cor**	**Rel**	**Flu**	**Info**	**Overall Score**
Easy test					
1	**3.82**	**3.88**	3.65	3.15	**3.74**
2	3.68	3.86	3.62	3.23	3.60
3	3.75	3.77	**3.67**	**3.35**	3.69
Hard test					
1	**3.80**	3.75	3.66	3.10	3.68
2	3.72	3.76	3.58	3.18	3.65
3	3.76	**3.80**	**3.70**	**3.32**	**3.72**

III. Track 2 - Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations

A. Track Overview

Recently, more public data sets and benchmarks have become available for dialogue research on task-oriented conversations in various domains [21], [22], [23], [24]. However, most data sets include only written conversations collected by crowdsourcing via web interfaces, which differ from spoken conversations for the following reasons. First, there are differences between the style of spoken and written conversations, even for the same context, intention, and semantics. Second, spoken conversations tend to have extra noise from grammatical errors, disfluencies or barge-ins, which are rarely encountered when processing written text. Finally, speech recognition output is not perfect and contains errors, which brings in additional challenges for developing spoken dialogue systems in practice.

There have been extensive studies towards robust language understanding against spoken input in dialog systems, especially for single-turn intent classification and slot filling tasks [25], [26], [27], [28], [29], [30], [31], [32], [33]. Nonetheless, the research communities have rarely addressed these issues on more contextual dialogue tasks including dialogue state tracking, dialogue policy learning, or end-to-end dialogue response generation, which are as important as the single-turn understanding tasks in fully working dialogue systems. This is mainly due to the lack of rich, annotated spoken data for such multi-turn dialogue tasks.

TABLE VI: Summary of Track 2 tasks

Task #1	**Multi-domain Dialogue State Tracking**
Goal	To estimate the system's belief states after each user turn
Input	Current user utterance, dialog context, and domain DB
Output	Belief state with a set of slot-value pairs
Metrics	JGA, Slot accuracy, Value P/R/F, None P/R/F
Task #2-1	**Knowledge-seeking Turn Detection**
Goal	To decide whether to continue existing flow or trigger the knowledge access branch for a given dialogue context
Input	Current user utterance, dialog context, and domain API and knowledge sources
Output	Binary class (requires knowledge access or not)
Metrics	Precision/Recall/F-measure
Task #2-2	**Knowledge Selection**
Goal	To select proper knowledge sources from the domain knowledge-base for each knowledge-seeking turn
Input	Current user utterance, dialog context, and the entire set of knowledge candidates
Output	Ranking of top-k knowledge candidates
Metrics	MRR@5, Recall@1, Recall@5
Task #2-3	**Knowledge-grounded Response Generation**
Goal	To generate a system response for given dialogue context and the selected knowledge snippets
Input	Current user utterance, dialog context, and selected knowledge sources
Output	Generated system response
Metrics	BLEU, METEOR, ROUGE, Human Ratings

To benchmark the robustness of conversational models on spoken conversations, this challenge track introduces a new data set with spoken task-oriented dialogues for two subtasks: 1) multi-domain dialogue state tracking [23] and 2) knowledge-grounded dialogue modeling [34], as summarized in Table VI. Our new data includes the ASR output instead of manual transcripts for the user turns, which aims to evaluate how robust each model is against ASR errors. The remainder of this section presents the data details and reports the evaluation results of the submitted entries from the challenge track participants.

B. Data

To study speech-based task-oriented dialogue modeling, we collected spoken human-human dialogues about touristic information for San Francisco. Each session was collected by pairing two participants: one as a user and the other as an agent. We provided a set of specific goals to the user-side participant before each session. The agent-side participant

TABLE VII: Statistics of the Track 2 data sets.

Split	Total # dialogues	Task 1: # instances	Task 2: # instances	Task 2: # positive
Val	107	936	263	104
Test	783	6588	1988	683

had access to the domain database including both structured information and unstructured text snippets. We recorded 890 sessions, which are around 45 hours in total, and manually transcribed all the utterances. Table VII shows the statistics of DSTC10 data. For each of the user turns we provide the ASR output instead of manual transcripts. Our ASR model is based on the wav2vec 2.0 model [35] that was pre-trained on 960 hours of Librispeech [36] and then fine-tuned with 10% of our validation data. This model achieved a WER of 26.25% at 1-best and 24.31% oracle WER at 10-best hypotheses on the user utterances on our test set.

C. Evaluation Criteria

Each participating team submitted up to five system outputs for either or both tasks. For task 1, we performed only automatic evaluations by comparing the submitted DST predictions with the ground-truth labels. We calculated the joint goal accuracy (JGA) as the main evaluation metric as well as the slot-level scores listed in Table VI.

For task 2, we use the same evaluation criteria and metrics as in the DSTC9 Track 1 [37]. First, for each submission we calculated the task-specific objective metrics (Table VI) by comparing to the ground-truth labels and responses. Then, we aggregated a set of multiple scores across different tasks and metrics into a single overall score computed by the mean reciprocal rank. Based on the overall objective score, we selected the finalists to be manually evaluated by two crowd-sourcing tasks:

- Appropriateness: This task asks crowd workers to score how well a system output is naturally connected to a given conversation on a scale of 1-5.
- Accuracy: This task asks crowd workers to score the accuracy of a system output based on the provided reference knowledge on a scale of 1-5.

Finally, we used the average of the Appropriateness and Accuracy scores to determine the official ranking of the submissions to task 2.

D. Results

We received a total of 99 submissions, including 40 entries from 11 teams for task 1 and 59 entries from 16 teams for task 2. Six of the teams participated in both tasks. To preserve anonymity, the teams were identified by A01 - A11 for task 1 and B01 - B16 for task 2.

1) Task 1 Results: Table VIII shows the task 1 evaluation results of the best entries from each team selected based on JGA. We differentiated between the single-model and ensemble-based entries and categorized the core methods into value classification, span extraction, value generation, or hybrid approaches combining more than one of them. A key

TABLE VIII: Task 1 results of the best entries from each team. Cls., Ext., Gen., and Ens. correspond to Value Classification, Span Extraction, Value Generation and Model Ensemble, respectively. Bold denotes the best JGA score, the winner of task 1 was A11; underline indicates the best single model performance.

Team - Entry	Proposed Methods				Joint Goal Accuracy
	Cls.	Ext.	Gen.	Ens.	
A11 - 1			✓	✓	**0.4616**
A11 - 4			✓		<u>0.4071</u>
A01 - 0			✓		0.3605
A01 - 2			✓		0.3553
A07 - 1			✓	✓	0.2773
A10 - 4	✓			✓	0.2679
A09 - 4		✓		✓	0.1821
A06 - 3	✓			✓	0.1691
A05 - 3	✓	✓	✓	✓	0.1615
A03 - 1		✓		✓	0.0524
A04 - 0		✓			0.0050
A08 - 0		✓			0.0018
A02 - 0		✓			0.0014
Baseline: TripPy [38]	✓	✓			0.0039

observation is that the generative models outperformed the other classification or extraction-based methods, consistent with findings on written conversations. We suppose this demonstrates the benefit of the generation-based DST in terms of its robustness against unseen values, different styles, as well as noisy transcriptions in our test data. On the other hand, most span extraction models failed to predict accurate dialogue states, because many of the extracted spans from spoken dialogue contexts with lexical variations and ASR errors are not correct dialogue state values. Another finding from the highly-ranked teams is that they commonly made huge efforts in data augmentation to account for the difference between the training and test data sets. Especially, Team A11 achieved the best performance by trying various data augmentation methods including value substitutions, synthetic data generation and speech/ASR simulation. In addition, the model ensemble also helped to boost the performance from the single-model results. We observed the performance gains by the model ensemble from all three teams: A11, A10, and A09 who submitted the entries in both settings. In particular, the ensemble-based entry from the winning team A11 was significantly better than their single model and also all the entries from other teams.

2) Task 2 Results: Table IX shows the objective evaluation results of the best entry from each team selected based on the overall score. Most entries show the improved performance from the DSTC9 [37] and Knover [39] baseline models in all three tasks. Team B10 achieved significantly better knowledge selection results than all the other teams, which may be attributed to the huge amount of augmented data they generated as well as the enhanced negative sampling methods. For response generation, the top 8 teams achieved at least two to three times higher scores than the baselines in the key automatic generation metrics. This is mainly because of their efforts on style transfer from written to spoken languages in response generation. For example, Team B08 introduced a noisy channel model to guide the generated responses towards more spoken styles and it helped to get the best scores in all the automated generation metrics compared to the reference

TABLE IX: Task 2 objective evaluation results of the best entries from each team. B-1, M and R-L correspond to BLEU-1, METEOR and ROUGE-L, respectively. Bold denotes the best score for each metric; underline indicates the best results among the single models with no ensemble; and * indicates the finalists selected for the human evaluations.

Team-Entry	Task #2-1 F	Task #2-2 R@1	Task #2-3: Generation B-1	M	R-L
DSTC9 [37]	0.7954	0.4583	0.1153	0.1215	0.1143
Knover [39]	0.7692	0.4950	0.1248	0.1364	0.1229
B01 - 0	0.8875	0.1840	0.3017	0.3514	0.3389
B02 - 3*	0.9037	0.6929	0.3732	0.4387	0.4115
B03 - 2	0.1260	0.0000	0.0063	0.0076	0.0090
B04 - 1*	0.9179	0.7481	0.3380	0.4070	0.3868
B05 - 0	0.8835	0.4499	0.1553	0.1689	0.1616
B06 - 1	0.7907	0.4887	0.1184	0.1285	0.1177
B07 - 3*	0.8817	0.6527	0.3335	0.4172	0.3915
B08 - 0	0.9045	0.7028	0.3961	0.4526	0.4372
B08 - 3*	0.9109	0.7097	**0.4013**	**0.4597**	**0.4405**
B09 - 1	0.8459	0.3067	0.1269	0.1365	0.1241
B10 - 0	0.9147	0.7428	0.1427	0.1919	0.2009
B10 - 1*	0.9228	**0.7933**	0.1617	0.2096	0.2187
B11 - 0	0.7990	0.4131	0.1106	0.1257	0.1154
B12 - 1*	0.9028	0.6915	0.3327	0.4039	0.3742
B13 - 1	0.8889	0.5716	0.1452	0.1580	0.1420
B14 - 0*	**0.9241**	0.6204	0.2705	0.3167	0.3184
B15 - 1	0.8559	0.3135	0.0496	0.0835	0.1036
B16 - 3*	0.8774	0.7105	0.3083	0.3547	0.3523

TABLE X: Human evaluation results. Bold indicates the best score for each metric, the winner of task 2 was B10, and † denotes the statistical significance ($p < 0.01$) from the paired t-test.

Rank	Team	Entry	Accuracy	Appropriateness	Average
Ground-truth			3.5769	3.4814	3.5292
1	B10	1	**3.4947**†	**3.3523**	**3.4235**†
2	B04	1	3.3356	3.3021	3.3189
3	B08	3	3.3433	3.2559	3.2996
4	B14	0	3.2935	3.2815	3.2875
5	B02	3	3.2932	3.2271	3.2602
6	B12	1	3.2546	3.2336	3.2441
7	B16	3	3.1874	3.1251	3.1563
8	B07	3	3.1315	3.1007	3.1161
Baseline: DSTC9			2.7425	2.7894	2.7659
Baseline: Knover			2.7793	2.7435	2.7614

responses from spoken human-human conversations.

We selected 8 finalists to be manually evaluated, corresponding to the best entry from each of the top 8 teams in the overall objective score. Table X shows the official ranking of the finalists based on the human evaluation results. Team B10 won the task 2 with the highest scores for both Accuracy and Appropriateness. A notable observation is that Team B10 was just in the middle rank in the automatic NLG metrics, due to the lack of style transfer mechanisms in their systems. Nonetheless, their system responses were more preferred by the crowd-workers in the human evaluation compared to the other entries even with much higher objective scores.

Consistently with our DSTC9 track results [37], the best team on the knowledge selection task again ended up with the final winner after the human evaluation. Most participating teams took the pipelined system architecture as the baselines, including three models for detection, selection, and generation, each of which was fine-tuned from the large-scale pre-trained language models. On the other hand, three of the top-4 teams introduced a separate entity tracking component for knowledge selection to narrow down the search space before the document ranking. In addition, all the top-4 teams for task 2 utilized the augmented data to train their models. Finally, model ensembles further improved performance.

E. Conclusions

We presented the official evaluation results of our DSTC10 track on the Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations. This challenge track addressed the multi-domain dialogue state tracking and the knowledge-grounded conversational modeling tasks on spoken task-oriented conversations. We released the validation and test data sets including 890 dialogues collected from spoken human-human conversations. A total of 21 teams participated with an overall number of 99 entries submitted. From the evaluation results, we learned the following two key factors to achieve high performance in both tasks: data augmentation for better generalization to unseen data and ensemble of different model outputs.

IV. Track 3 - SIMMC 2.0: Situated Interactive Multimodal Conversational AI

A. Track overview

The SIMMC challenge aims to lay the foundations for the real-world assistant agents that can handle multimodal inputs, and perform multimodal actions. We thus focus on task-oriented dialogs that encompass a situated multimodal user context in the form of a co-observed image or virtual reality (VR) environment. The context is dynamically updated on each turn based on the user input and the assistant action. Moon *et al.* [40] (SIMMC 1.0) and Kottur *et al.* [41] (SIMMC 2.0) provide more details on the datasets and the models we provide.

B. Data

SIMMC 2.0 dataest contains about $11k$ human-to-human dialogs (totaling about $117k$ utterances). We chose shopping experiences—specifically furniture and fashion—as the domain for the SIMMC datasets because of the dynamic environment created by these domains, where rich multimodal interactions happen around visually grounded items.

SIMMC offers many key advantages over previous multimodal dialog datasets:

1) SIMMC assumes a co-observed multimodal context between a user and an assistant and records the ground-truth item reference. SIMMC tasks emphasize semantic processing of the input modalities, while work in this area has traditionally focused heavily on raw image processing.
2) SIMMC emphasizes semantic processing. The proposed SIMMC annotation schema allows for a more systematic and structural approach for visual grounding of conversations, which is essential for solving challenging problems in real-world scenarios.
3) SIMMC 2.0 provides photo-realistic scenes that change over time (via viewpoint updates), moving away from the sanitized contexts present in many multimodal datasets.

TABLE XI: Summary of the results on Test-Std split. Best results from each system are shown. **(1) Multimodal Disambiguation (Disamb.)**, via classification accuracy, **(2) Multimodal Coreference Resolution (MM-Coref)**, via coref prediction F1, **(3) Dialog State Tracking (DST)**, via slot and intent F1, **(4) Response Generation** via BLEU, recall@k (k=1,5,10), Mean rank, and mean reciprocal rank (MRR). ↑: higher is better. Baselines: GPT-2-based [40] and MTN-based [42].

Team	1. Disamb.	2. MM-Coref	3. MM-DST		4. Response Retrieval & Generation					
	Acc↑	Coref F1↑	Slot F1↑	Intent F1↑	MRR↑	r@1↑	r@5↑	r@10↑	Mean↓	BLEU↑
GPT-2	73.5	44.1	83.8	94.1	·	·	·	·	·	0.202
MTN	·	·	76.7	92.8	·	·	·	·	·	0.211
Team 1	·	52.1	88.3	96.3	53.5	42.8	65.4	74.9	11.0	0.285
	·	51.9	88.4	96.3	51.7	41.2	62.8	72.5	11.9	0.279
Team 2	·	78.3	·	·	·	·	·	·	·	·
Team 3	89.5	42.2	87.8	96.2	61.2	49.6	74.7	84.5	6.6	0.256
Team 4	93.9	75.8	90.3	95.9	81.5	71.2	95.0	98.2	1.9	0.295
Team 5	93.8	56.4	89.3	96.4	32.0	19.9	41.8	61.2	12.9	0.322
Team 6	94.7	59.5	91.5	96.0	·	·	·	·	·	·
	94.5	·	·	·	·	·	·	·	·	0.309
Team 7	93.1	57.3	·	·	·	·	·	·	·	·
Team 8	93.1	63.4	4.0	41.4	·	·	·	·	·	0.297
	·	63.0	·	·	·	·	·	·	·	·
	·	66.7	·	·	·	·	·	·	·	·
	·	68.2	·	·	·	·	·	·	·	·
Team 9	·	73.3	·	·	·	·	·	·	·	·
Team 10	·	50.6	·	·	·	·	·	·	·	·
	93.6	68.2	87.7	95.8	·	·	·	·	·	0.327

C. Evaluation Criteria

We present four subtasks primarily aimed at replicating human-assistant actions in order to enable rich and interactive shopping scenarios.

1) Subtask 1: Multimodal Disambiguation: identifying whether a given user turn contains ambiguity in referencing to objects in the scene. As defined in [41], given the dialog history and the current user utterance, multimodal disambiguation requires the agent to predict a binary label conditioned on the multimodal context, to indicate the presence of a referential ambiguity in the user utterance. We use accuracy to measure and compare model performances for this task.

2) Subtask 2: Multimodal Coreference Resolution: requires the dialog system to resolve referential mentions in user utterances to their canonical object IDs as defined for each scene. These mentions can be resolved through (1) the dialog context (e.g., A: *'This shirt comes in XL and is $29.'* → U: *'Please add it to cart.'*, or (2) the multimodal context (e.g., U: *'How much is that red shirt?'*), or (3) both (e.g., U: *'How much is the one next to the one you mentioned?'*). The main evaluation metric includes F1, precision and recall performance.

3) Subtask 3: Dialog State Tracking (DST): aims to systematically track the dialog acts and the associated slot pairs across multiple turns, as represented in the flexible ontology developed to represent the SIMMC multimodal context. We use the intent and slot prediction metrics (F1), inline with prior work in DST.

4) Subtask 4: Response Prediction: examines the relevance of the assistant response in the current turn. We evaluate in two ways; *(a)* as a conditional language modeling problem, where the closeness between the generated and ground-truth response is measured through using BLEU-4 score, and, *(b)* as a retrieval problem, where we measure the model performance when retrieving ground-truth responses from a pool of 100 candidates (randomly chosen and unique to each turn).

D. Results

The challenge saw a total of 16 model entries from 10 teams across the world, setting a new state-of-the-art in all four subtasks (Table XI).

For each subtask, we listed metrics in a priority order and the entry with the most favorable performance on the highest priority metric was considered to be a candidate winner.

The winner of the multimodal disambiguation subtask (subtask 1) was the BART+ResNet model from Team 6. This model was the winner for the MM-DST subtask (subtask 3) as well. The winner of the multimodal coreference resolution task (subtask 2) and the response retrieval task (subtask 4a) was a BART-based multimodal model from Team 4. The joint winners of the response generation (4b) were Team 5 and 10.

V. Track 4 - Reasoning for Audio Visual Scene-Aware Dialog

A. Track Overview

Recent artificial intelligence (AI) research activities have accelerated the development of technologies required for advanced human-like capabilities in machines, such as robots. For instance, current computer vision technologies can accurately perceive visual scenes, and spoken dialog systems can transcribe speech and understand speakers' intention. However, one important piece of technology is missing: natural and context-aware human-machine interaction, where machines understand their surrounding scene from the human perspective,

TABLE XII: AVSD dataset for DSTC10

	Train	Validation	Test
#dialogs	7,659	1,787	1,804
#turns	153,180	35,740	28,406
#words	1,450,754	339,006	272,606

and they are able to share their understanding with humans using natural language.

To invent machines that can communicate with humans about objects and events in surrounding scenes, the project to work on Audio-visual Scene-aware Dialog (AVSD) was kicked-off [6], [43], [44], [11]. An automated system that can converse with humans on video scenes via natural dialogs is a challenging research problem. The goal of AVSD in DSTC is to have question-answering based conversations on videos from daily life. To this end, the AVSD challenge task was designed based on the popular Charades dataset [45], with the goals: (1) generate answers to questions about objects and events in the video clips and (2) hold a meaningful dialog with humans about objects and events using conversational frameworks.

To promote further advancements into real-world applications of the AVSD setup, a third challenge was proposed in DSTC10, progressively improving the challenge from the previous video-based scene-aware dialog tracks. The new task is to generate sentences for a system response to a query that occurs during a dialog about a video using reasoning features without using the human-created video description. Participants used the video, audio, and dialog text data to train end-to-end models without the manual descriptions. This challenge used the AVSD datasets that were collected and used in the previous challenges. The additional datasets for temporal reasoning for QA datasets were collected and used in DSTC10.

B. Audio-Visual Scene-Aware Dialog data set

The AVSD in DSTC10, the same AVSD data collected by [43] have been used. Table XII shows the size of the data used for DSTC10. For DSTC10, additional data for temporal reasoning were collected, in which humans watched the videos and read the dialogues, then identified segments of the video containing evidence to support each given answer. Figure 3 shows the annotation tool for reasoning. With this tool, humans identified temporal segments based on visual evidence and/or audio evidence and filled in the appropriate fields with begin and end timestamps to provide temporal reasoning.

C. Baseline Model

A baseline system has been built for the DSTC10 AVSD track, which utilizes an AV-transformer architecture [46]. The system employs a transformer-based encoder-decoder, including a bimodal attention mechanism [47], [48] that lets it learn interdependencies between audio and visual features.

The audio-visual encoder extracts VGGish [49] and I3D [50] features from the audio and video tracks, respectively, and encodes these using self-attention, bimodal attention, and feed-forward layers. The decoder receives the encoder outputs and the dialog history until the current question, and starts

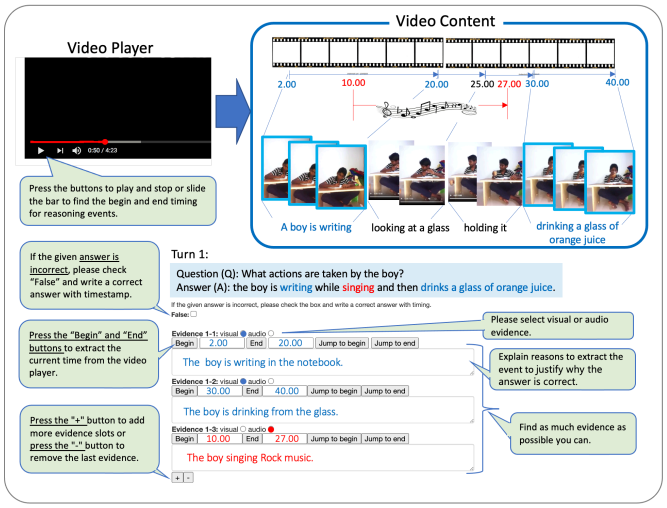


Fig. 3: Temporal reasoning data collection tool for AVSD.

generating the answer sentence. At each iteration step, it receives the preceding word sequence and predicts the next word by applying M decoder blocks and a prediction network. The self-attention layer converts the word vectors to high-level representations considering temporal dependency. The bimodal source attention layers update the word representations based on the relevance to the encoded multi-modal representations. A feed-forward layer is then applied to the outputs of the bimodal attention layers. Finally, a linear transform and softmax operation are applied to the output of the M-th decoder block to obtain the probability distribution of the next word.

D. Temporal Reasoning

Temporal reasoning is the task of finding evidence supporting the generated answers, where the evidence corresponds to human-annotated time regions of the video that have been identified as supporting each ground-truth answer. Human annotators were allowed to choose multiple time regions for each question-answer pair, but most of the reasons consist of a single region.

E. Submitted Systems and Evaluation

The AVSD Task received 12 system submissions from 5 teams. This section summarizes the techniques used in the submitted systems to the AVSD challenge, including the baseline system. Table XIII lists the baseline and submitted systems with brief specifications including the encoder-decoder model type, multimodal fusion type, audio-visual video features used, and additional techniques or data sets.

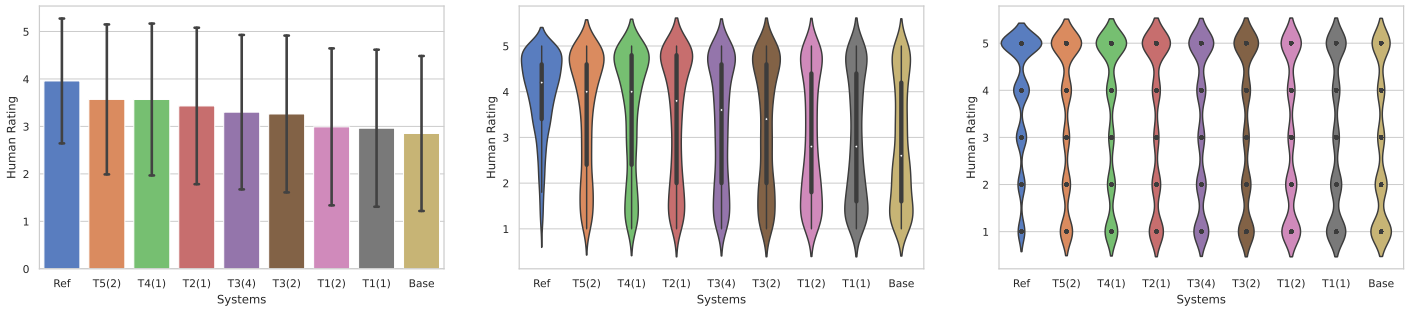
In this challenge, the quality of a system's automatically generated sentences is evaluated using objective measures to determine the level of similarity between the system-generated responses and ground-truth responses provided by humans. For this purpose, we needed to collect more human-generated responses to each test question (the original dialog, of course, contains only a single human response to each question). To collect more possible human answers in response to the test question for each test video, we asked 5 humans to watch the

TABLE XIII: Submitted systems to the DSTC10-AVSD track

Team	Encoder-decoder type	Multimodal fusion type	Features	Additional techniques/data
Baseline	Transformer	Audio-visual bi-modal att.	I3D, VGGish, QA history	
Team 1	Transformer	(1) Triple cross attention (2) (1) + Attentional fusion	I3D, VGGish, QA history	Student-teacher learning (STL)
Team 2: [51]	LSTM & Transformer	Attentional multimodal fusion	I3D, VGGish, last question	(1) Linear comb. of LSTM & Transformer + STL (2) (1) + Cross student-teacher loss [52]
Team 3: [53]	Transformer	Input multimodal labels and text to GPT-2	Action/event labels, QA history	GPT-2 with Action/Event labels by Video Swim/Audio Spectrum Transformers. Reasoning with 2D-TAN net.
Team 4: [54]	Transformer	Input video features and text to GPT-2	TimeSformer video feature, QA history	GPT-2 + TimeSformer video features with (1) fixed 32 frames or (2) variable-length input
Team 5: [55]	(1) Attention-based encoder decoder (2) UniVL model	(1) Concat. of encoder states from different modalities (2) Cross-attention in early fusion of multimodal features	(1) S3D, VGGish, BERT-encoded QA history (2) (1) + object feature	(2) UniVL model + D2Det object detector

TABLE XIV: DSTC10-AVSD evaluation results with word-overlap-based objective measures based on 6 references, a subjective measure based on 5-level ratings by humans (HR), and reasoning performance based on Intersection-over-Union (IoU).

System	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE_L	CIDEr	IoU-1	IoU-2	HR
Baseline	0.572	0.422	0.320	0.247	0.191	0.439	0.566	0.361	0.380	2.851
Team 1 (1)	0.601	0.451	0.347	0.270	0.196	0.456	0.607	0.360	0.378	2.962
Team 1 (2)	0.598	0.449	0.345	0.270	0.198	0.458	0.613	0.362	0.380	2.990
Team 2 (1)	**0.695**	**0.564**	**0.462**	0.381	**0.248**	**0.540**	0.888	-	-	3.431
Team 2 (2)	0.692	0.563	**0.462**	0.381	0.246	0.537	0.880	-	-	-
Team 3 (1)	0.641	0.489	0.379	0.298	0.225	0.502	0.804	0.506	0.534	-
Team 3 (2)	0.624	0.475	0.366	0.286	0.231	0.503	0.786	**0.516**	**0.544**	3.262
Team 3 (3)	0.651	0.490	0.376	0.295	0.227	0.502	0.789	0.505	0.533	-
Team 3 (4)	0.646	0.489	0.380	0.299	0.225	0.499	0.787	0.505	0.533	3.300
Team 4 (1)	0.680	0.558	0.461	**0.385**	0.247	0.539	**0.957**	-	-	3.567
Team 4 (2)	0.679	0.554	0.456	0.379	0.246	0.536	0.945	-	-	-
Team 5 (1)	0.670	0.541	0.441	0.365	0.241	0.526	0.906	0.485	0.510	-
Team 5 (2)	0.673	0.545	0.448	0.372	0.243	0.530	0.912	0.479	0.505	**3.569**
Reference	-	-	-	-	-	-	-	-	-	3.958



(a) Mean and standard deviation. (b) Distribution over sentence scores. (c) Distribution over score values.

Fig. 4: Statistics of human rating scores

video, read a dialogue (up to the test question) about the video between a questioner and an answerer, and then provide an answer in response to the test question.

To evaluate the systems, we compared them with 6 ground-truth human answers, which consisted of the one original answer and these 5 newly collected answers. We used the MSCOCO evaluation tool for objective evaluation of system outputs. The supported metrics include metrics based on word overlap, such as BLEU, METEOR, ROUGE_L, and CIDEr. In addition, we collected human ratings for each system response using a 5-point Likert scale, in which humans rated system responses given a dialog context. We asked the human raters to consider correctness of the answers as well as naturalness, informativeness, and appropriateness of the response according to the given context. The reasoning performance was measured by Intersection over Union (IoU), which indicates the ratio of overlap between the predicted and ground-truth time regions (higher is better). IoU-1 is obtained as an average IoU computed between each ground truth and the predicted region that gives the highest IoU to the ground truth. IoU-2 is computed by frame-level matching among all predicted and ground-truth regions for each answer.

TABLE XV: Comparison of answer qualities while the videos are shuffled or not.

System	Videos	BLEU-4	METEOR	CIDEr
Baseline	matched	0.247	0.190	0.566
	shuffled	0.241	0.188	0.559

Table XIV reports the numerical results of all qualifying submitted systems (entries) from all teams. The subjective human ratings described above are given in the rightmost column of the table, and the others are the objective scores that were computed using word-overlap metrics (Bleu, METEOR, ROUGE_L, and CIDEr) and reasoning metrics (IoU-1 and IoU-2). Figure 4 plots the human ratings for each system in several ways. In all three figures, the systems are shown in the same order on the x-axis.

We tested our baseline model in two settings: giving matched and shuffled videos. As indicated in Table XV, we can see a certain degradation in the scores of the two systems, but the performance gaps are relatively small. Thus, the result suggests that text information is dominant in the AVSD task, and at the same time, the expressive power of the baseline video features, i.e., I3D and Vggish, is insufficient. In other words, developing more advanced video features was one of the important issues in this challenge; better video features are important. In the DSTC10-AVSD challenge, some teams applied more advanced video features and reported substantial improvement. For example, Team 4 introduced TimeSformer to extract video features.

F. Conclusion

The third AVSD challenge promoted further advancements for real-world applications, where 1) the human-created description is unavailable at inference time, and 2) systems must demonstrate temporal reasoning by finding evidence from the video to support each answer. The submitted systems provided high-quality answers and reasoning even without human-generated descriptions at inference time. The DSTC10 winning system achieved 90.2% of the human performance based on human ratings. The result is considerable, but the gap with human performance is actually larger than the DSTC8 result (98.4%). This shows that continued research is still needed to achieve human performance. The data setup, baseline system, and evaluation tools are released, which facilitate continuous improvement by the community after the DSTC10.

VI. Track 5 - Automatic Evaluation and Moderation of Open-domain Dialogue Systems

A. Track Overview

Our track consists of two tasks: (1) Automatic Open-domain Dialog Evaluation. (2) Safe Chatbots Development. The goal of the first task is for participants to design robust automatic dialogue evaluation metrics that correlate well with human judgements across multiple dialogue domains as well as across different dialogue evaluation dimensions, such as naturalness, appropriateness, etc. The goal of the second task is for the participants to build generative models that first detect a toxic user's comment, and then generate appropriate and polite responses that keep the dialogue fluid and nontoxic.

B. Data

1) Task 1 - Automatic Dialogue Evaluation: As evaluation benchmark we released 14 publicly available datasets for the participants to tune their proposed metrics during the development phase. During the final evaluation phase, we collected five hidden test evaluation datasets for assessing participants' submissions. The datasets and final leaderboards are publicly available on the ChatEval platform[2].

Each turn-level dataset is a collection of context-response pairs. The context refers to a list of consecutive utterances that are extracted from a human-human conversation. The response is produced by a dialog model conditioned on the context. Each context-response pair was assessed by several human judges along different evaluation criteria.

The 19 evaluation datasets cover a large number of distinct dialogue domains, such as daily chitchat [56], knowledge exchange [57], and persona-based conversations [58], and a large number of different evaluation criteria, such as naturalness, interestingness, response appropriateness, etc. More details can be found in [59]

2) Task 2 - Safe Chatbots Development: Several datasets were preprocessed and formatted from their original sources as part of the Chat/Dialogue Modeling and Evaluation task (CHANEL) held during the 2020 Seventh Frederick Jelinek Memorial Summer Workshop[3]. All selected datasets are organized into turn of pairs (prompt-answer) and processed using Microsoft Azure Cognitive Services to automatically detect toxic turns. Then, we selected those pairs where the prompt was detected as toxic but the answer was not. To reduce false positives in the prompts or false negatives in the answers, we filtered the Azure results by passing all detected turns through a dictionary consisting of the 320 most common swear words in English. Concretely, the datasets we used include: (1) MovieDic [60] (2) Cornell Movie Dataset [61] (3) ChatCorpus [60][4] (4) DSTC8-Reddit [62][5] Refer to [59] for more details of the four datasets, which are annonymized. Besides the toxicity detection process, we extract additional features: humour scores and detected emotion. Humour scores are extracted using Colbert pretrained model [63]. Emotion detection were trained by four different datasets [64], [56], [65], [66], distinguishing up to 7 different emotions: happiness, sadness, fear, angry, surprise, disgust, and neutral [67] .

To further assess task difficulty, we manually annotated a subset of the test data. In total, 1290 prompt-answer pairs were annotated by 7 annotators from three different geographical zones (3 in the USA, 3 in Europe, and 1 in Asia). An annotation guideline, with no examples, was prepared to avoid biasing responses. Refer to [59] for the annotation guideline details.

[2]https://chateval.org/dstc10
[3]https://www.clsp.jhu.edu/workshops/20-workshop/
[4]https://github.com/Marsan-Ma/chat_corpus/
[5]https://github.com/microsoft/dstc8-reddit-corpus

C. Baselines

For each task, we provide a baseline system. For "Automatic Dialogue Evaluation" task, We adopt the deep AM-FM framework [68], an ensemble metric, as the baseline for the automatic dialogue evaluation task[6]. We modify the framework to a reference-free version whereby for AM, we compute the cosine similarity between the sentence-level embedding of the response and that of the last sentence in the corresponding dialogue context. For FM, we use the formulation of the context-response coherence metric in HolisticEval [69].

For the "Safe Chatbots Development" task, participants are provided with a baseline system based on DialoGPT: a GPT-2 model pretrained on 147M multi-turn dialogues from Reddit threads [70] and finetuned on our provided training data[7].

D. Evaluation Criteria

In task 1, we adopted Spearman correlation to assess the participants' submissions. We rank the submissions only based on their performance on the five test evaluation datasets. We compute the Spearman correlation between the submitted metric scores and the corresponding mean human annotation scores per evaluation dimension for each evaluation dataset.

In task 2, we conduct both automatic and human evaluation. For automatic evaluation, we adopt four different objective metrics: a) BLEU [71], b) ROUGE-L [72], c) BERTScore [73], and d) BLEURT [74]. For human evaluation, we perform a pairwise ranking of the system-generated responses given a toxic prompt. A subset of 160 toxic prompts are randomly selected from the golden test set for pairwise analysis.

E. Results

1) Task 1 - Automatic Dialogue Evaluation: In task 1, we received 21 and 35 submissions from nine different teams for development and testing, respectively. Table XVI presents the main correlation results of each team on the five test datasets. For each row in the table, we show the Spearman rank correlation w.r.t. each team's best submission. Each entry in row 6 is computed by averaging the 11 dimension-wise correlation scores over all five test datasets. Each dimension-wise correlation score is computed between the metric scores assigned to all data instances within a test dataset and the corresponding human annotated scores along one evaluation criterium of that particular dataset.

Remarkably, Team 1, 5, and 8 all rely on ensembling multiple sub-metrics for evaluation. The weights of combining different sub-metrics are dynamically learnt from the data. This finding is inline with the observation made in Yeh et al. [75], which highlights the advantage of combining multiple sub-metrics.

2) Task 2 - Safe Chatbots Development: Unfortunately, there was no submission for this task. Hence, we decided to test the performance of three existing state-of-the-art chatbots on our annotated golden test set (described in Section VI-B2). The three chatbots include: a) the pretrained baseline released to the participants (a finetuned version of DialogGPT [70]). b) BlenderBot Vs 2.0 (including its safety layer) [76], [77], and c) GPT-3 [78] (the DaVinci version)[8].

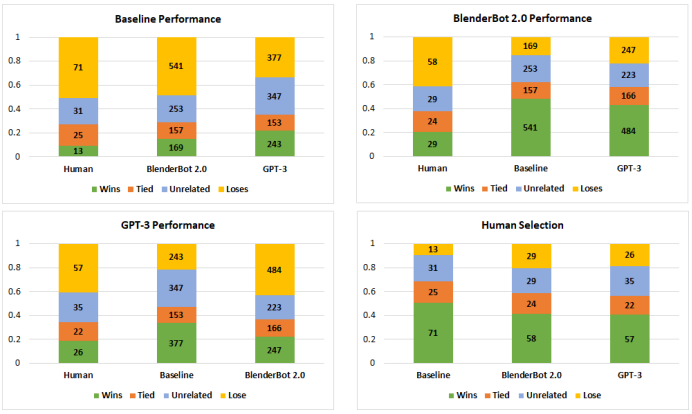


Fig. 5: Comparative performance between the different chatbots and human answers on the annotated test set

Table XVII shows the automatic evaluation results for each chatbot. The results for the word-overlap metrics (BLEU and ROUGE) are very low due to the high differences in the system generated responses and the corresponding human references. On the other hand, semantic metrics (i.e., BERTScore and BLEURT) show marginal differences between chatbots, with BlenderBot Vs 2.0 performing slightly better. In Table XVIII and Figure 5 shows performance of chatbots and humans.

F. Conclusion & Future Work

We conclude the track with several important points that can benefit future development of automatic dialogue evaluation metrics and safe dialogue systems. (a) in task 1, we notice that all the teams' performance on the development data is much better than that on the hidden test data (around 31.4% in average) except the baseline, which performs better on the test data (around 34.5% better) probably due to the usage of a more simple mechanism for combining different evaluation dimensions and some topic overlap between the training data and test sets (e.g., topical and persona datasets). Hence, future work should explore models with better generalization to out-of-distribution evaluation (i.e., robustness). This research direction towards robust and generalizable metrics is also highlighted in Mehri et al. [79]. (b) we standardize a large number of dialogue evaluation datasets and release a ready-to-use and high-quality benchmark to meta-evaluate different capabilities of automatic dialogue evaluation metrics, such as domain generalization, multi-dimensionality, and robustness. The benchmark serves to help dialogue researchers and practitioners holistically assess their newly-proposed automatic dialogue evaluation metrics. (c) The second task is just scratching the surface on how to deal with toxic users. As there is currently not enough resources on this topic, we provide the data and baseline systems that can help advance the development of safe chatbots. (d) Future work may focus on more advanced techniques in detecting different types of toxicity and how to address them. In addition, efforts

[6]https://github.com/e0397123/dstc10_metric_track

[7]https://github.com/lfdharo/DSTC10_Track5_Toxicity

[8]Using OpenAI API at https://beta.openai.com/?app=chat

TABLE XVI: Mean Spearman correlations (%) for the baseline, with each team's best submission on the 5 test datasets. The best score for each row is highlighted in bold. The second best is underlined. The third best is italicized. Row 6 is averaged over 11 dimension-wise correlation scores of all 5 datasets instead of over the entries of rows 1-5 in the same column.

Row	Datasets	Baseline	Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	Team 8	Team 9
1	JSALT	5.09	**27.74**	3.10	10.54	4.96	11.66	*12.73*	4.07	8.75	<u>26.42</u>
2	ESL	32.29	<u>43.18</u>	19.86	28.75	9.34	*40.01*	32.92	3.28	36.10	**45.58**
3	NCM	16.49	**29.91**	1.98	22.08	8.24	<u>29.60</u>	*26.60*	2.01	25.57	19.11
4	Topical-DSTC10	17.48	*21.32*	10.85	14.56	8.33	**23.68**	20.00	1.43	<u>22.77</u>	17.41
5	Persona-DSTC10	19.61	30.67	7.77	25.80	16.59	**37.50**	*35.78*	2.54	<u>37.22</u>	33.82
6	Average	18.38	*27.81*	8.95	20.20	10.29	**29.63**	26.86	2.30	<u>28.19</u>	26.89

TABLE XVII: Objective metrics for tested chatbots in subtask 2

System	BLEU	ROUGE	BERTScore	BLEURT
Baseline	0.008	0.072	0.832	-1.180
BlenderBot 2.0	0.009	0.097	0.836	-1.183
GPT-3	0.008	0.065	0.831	-1.201

TABLE XVIII: Human performance for the subtask 2 test set. Percentages use total annotated items for each chatbot.

	Wins	Tied	Unrelated	Loses
Baseline	425 17.9%	335 14.1%	631 26.5%	989 41.6%
BlenderBot 2.0	1054 44.3%	347 14.6%	505 21.2%	474 19.9%
GPT-3	650 27.3%	341 14.3%	605 25.4%	784 32.9%
Human	186 44.3%	71 16.9%	95 22.6%	68 16.2%
Total	2315	1094	1836	2315

to avoid the use of toxic words is just a first step in reducing toxicity. There are other complex toxic scenarios to address.

VII. Conclusions

This paper summarizes five tracks in the tenth dialog system technology challenge (DSTC10). MOD: Internet Meme Incorporated Open-domain Dialog incorporates interbet memes into open-domain dialogues. Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations focuses on robustness in spoken conversations. The Situated Interactive Multi-Modal Conversational AI track focuses on real-world assistant agents that can handle multi-modal inputs and perform multi-modal actions. Reasoning for Audio Visual Scene-Aware Dialog promotes a multimodal reasoning task in conversational scenarios. Finally, Automatic Evaluation and Moderation of Open-domain Dialogue Systems target the proposal of new metrics, self-supervised methods, and non-toxic generation of responses for open-domain dialog systems. All datasets and resources introduced for each track are kept publicly available even after the challenge period to support future dialog system research.

References

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

[2] M. Henderson, B. Thomson, and J. D. Williams, "The second dialog state tracking challenge," in *Proc. SIGDIAL*, 2014, pp. 263–272.

[3] ——, "The third dialog state tracking challenge," in *Proc. SLT*. 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 *Proc. SLT*. IEEE, 2016, pp. 511–517.

[6] C. Hori, J. Perez, R. Higasinaka, 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," *CSL*, 2018.

[7] 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," *CSL*, vol. 62, p. 101068, 2020.

[8] C. Gunasekara, J. K. Kummerfeld, L. Polymenakos, and W. Lasecki, "DSTC7 task 1: Noetic end-to-end response selection," in *Proc. ConvAI*, 2019, pp. 60–67.

[9] M. Galley, C. Brockett, X. Gao, J. Gao, and B. Dolan, "Grounded response generation task at DSTC7," in *AAAI DSTC WS*, 2019.

[10] H. Alamri, V. Cartillier, R. G. Lopes, A. Das, J. Wang, I. Essa, D. Batra, D. Parikh, A. Cherian, T. K. Marks, *et al.*, "Audio visual scene-aware dialog challenge at DSTC7," *arXiv preprint arXiv:1806.00525*, 2018.

[11] S. Kim, M. Galley, C. Gunasekara, S. Lee, A. Atkinson, B. Peng, H. Schulz, J. Gao, J. Li, M. Adada, M. Huang, L. Lastras, J. K. Kummerfeld, W. S. Lasecki, C. Hori, A. Cherian, T. K. Marks, A. Rastogi, X. Zang, S. Sunkara, and R. Gupta, "Overview of the eighth dialog system technology challenge: DSTC8," *IEEE/ACM TASLP*, vol. 29, pp. 2529–2540, 2021.

[12] C. Gunasekara, S. Kim, L. F. D'Haro, A. Rastogi, Y.-N. Chen, M. Eric, B. Hedayatnia, K. Gopalakrishnan, Y. Liu, C.-W. Huang, *et al.*, "Overview of the ninth dialog system technology challenge: DSTC9," *arXiv preprint arXiv:2011.06486*, 2020.

[13] Y. Wang, Y. Li, X. Gui, Y. Kou, and F. Liu, "Culturally-embedded visual literacy: A study of impression management via emoticon, emoji, sticker, and meme on social media in China," *Proc. ACM-HCI*, vol. 3, no. CSCW, pp. 68:1–68:24, 2019.

[14] D. M. Beskow, S. Kumar, and K. M. Carley, "The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning," *Inf. Process. Manag.*, vol. 57, no. 2, p. 102170, 2020.

[15] C. Chen, "Research on sticker cognition for elderly people using instant messaging," in *Proc. HCI-CCD*, ser. Lecture Notes in Computer Science, P. P. Rau, Ed., vol. 12192. Springer, 2020, pp. 16–27.

[16] C. Posey, P. B. Lowry, T. L. Roberts, and T. S. Ellis, "Proposing the online community self-disclosure model: the case of working professionals in france and the U.K. who use online communities," *Eur. J. Inf. Syst.*, vol. 19, no. 2, pp. 181–195, 2010.

[17] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra, "Visual dialog," in *Proc. IEEE-CVPR*, 2017, pp. 1080–1089.

[18] 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 *Proc. IEEE-CVPR*, 2019, pp. 7558–7567.

[19] Z. Fei, Z. Li, J. Zhang, Y. Feng, and J. Zhou, "Towards expressive communication with internet memes: A new multimodal conversation dataset and benchmark," *arXiv preprint arXiv:2109.01839*, 2021.

[20] Y. Wang, P. Ke, Y. Zheng, K. Huang, Y. Jiang, X. Zhu, and M. Huang, "A large-scale Chinese short-text conversation dataset," in *Proc. NLP and CC - CCF*, ser. Lecture Notes in Computer Science, X. Zhu, M. Zhang, Y. Hong, and R. He, Eds., vol. 12430. Springer, 2020, pp. 91–103.

[21] L. El Asri, H. Schulz, S. K. Sarma, J. Zumer, J. Harris, E. Fine, R. Mehrotra, and K. Suleman, "Frames: a corpus for adding memory to goal-oriented dialogue systems," in *Proc. SIGdial*, 2017, pp. 207–219.

[22] T.-H. Wen, D. Vandyke, N. Mrkšić, M. Gasic, L. M. R. Barahona, P.-H. Su, S. Ultes, and S. Young, "A network-based end-to-end trainable task-oriented dialogue system," in *Proc. EACL*, 2017, pp. 438–449.

[23] 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 *Proc. EMNLP*, 2018, pp. 5016–5026.

[24] A. Rastogi, X. Zang, S. Sunkara, R. Gupta, and P. Khaitan, "Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset," in *Proc. AAAI*, vol. 34, no. 05, 2020, pp. 8689–8696.

[25] G. Tur, J. Wright, A. Gorin, G. Riccardi, and D. Hakkani-Tür, "Improving spoken language understanding using word confusion networks," in *Proc. ICSLP*, 2002.

[26] D. Hakkani-Tür, F. Béchet, G. Riccardi, and G. Tur, "Beyond asr 1-best: Using word confusion networks in spoken language understanding," *CSL*, vol. 20, no. 4, pp. 495–514, 2006.

[27] M. Henderson, M. Gašić, B. Thomson, P. Tsiakoulis, K. Yu, and S. Young, "Discriminative spoken language understanding using word confusion networks," in *Proc. IEEE-SLT*, 2012, pp. 176–181.

[28] G. Tur, A. Deoras, and D. Hakkani-Tür, "Semantic parsing using word confusion networks with conditional random fields." in *Proc. Interspeech*. Citeseer, 2013, pp. 2579–2583.

[29] R. Masumura, Y. Ijima, T. Asami, H. Masataki, and R. Higashinaka, "Neural confnet classification: Fully neural network based spoken utterance classification using word confusion networks," in *Proc. IEEE-ICASSP*, 2018, pp. 6039–6043.

[30] F. Ladhak, A. Gandhe, M. Dreyer, L. Mathias, A. Rastrow, and B. Hoffmeister, "Latticernn: Recurrent neural networks over lattices," in *Proc. Interspeech*, 2016, pp. 695–699.

[31] L. Velikovich, "Semantic model for fast tagging of word lattices," in *Proc. IEEE-SLT*, 2016, pp. 398–405.

[32] C.-W. Huang and Y.-N. Chen, "Adapting pretrained transformer to lattices for spoken language understanding," in *Proc. IEEE-ASRU*, 2019, pp. 845–852.

[33] ——, "Learning spoken language representations with neural lattice language modeling," in *Proc. ACL*, Online, July 2020, pp. 3764–3769.

[34] S. Kim, M. Eric, K. Gopalakrishnan, B. Hedayatnia, Y. Liu, and D. Hakkani-Tur, "Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access," in *Proc. SIGdial*, 2020, pp. 278–289.

[35] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. IEEE-ICASSP*. IEEE, 2015, pp. 5206–5210.

[37] S. Kim, M. Eric, B. Hedayatnia, K. Gopalakrishnan, Y. Liu, C.-W. Huang, and D. Hakkani-Tur, "Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access track in DSTC9," 2021.

[38] M. Heck, C. van Niekerk, N. Lubis, C. Geishauser, H.-C. Lin, M. Moresi, and M. Gasic, "Trippy: A triple copy strategy for value independent neural dialog state tracking," in *Proc. SIGdial*, 2020, pp. 35–44.

[39] H. He, H. Lu, S. Bao, F. Wang, H. Wu, Z. Niu, and H. Wang, "Learning to select external knowledge with multi-scale negative sampling," 2021.

[40] S. Moon, S. Kottur, P. A. Crook, A. De, S. Poddar, T. Levin, D. Whitney, D. Difranco, A. Beirami, E. Cho, R. Subba, and A. Geramifard, "Situated and interactive multimodal conversations," *Proc. COLING*, 2020.

[41] S. Kottur, S. Moon, A. Geramifard, and B. Damavandi, "SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations," in *Proc. EMNLP*, Nov. 2021, pp. 4903–4912.

[42] H. Le, D. Sahoo, N. Chen, and S. Hoi, "Multimodal transformer networks for end-to-end video-grounded dialogue systems," in *Proc. ACL*, 2019, pp. 5612–5623.

[43] 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 *Proc. IEEE-CVPR*, June 2019.

[44] 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 *Proc. IEEE-ICASSP*, 2019, pp. 2352–2356.

[45] 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.

[46] V. Iashin and E. Rahtu, "A better use of audio-visual cues: Dense video captioning with bi-modal transformer," in *Proc. BMVC*, 2020.

[47] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[48] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. NeurIPS*, 2015, pp. 577–585.

[49] 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 *Proc. IEEE-ICASSP*, Mar. 2017.

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

[51] A. P. Shah, T. Hori, J. Le Roux, and C. Hori, "DSTC10-AVSD submission system with reasoning using audio-visual transformers with joint student-teacher learning," in *Proc. AAAI-DSTC10*, 2022.

[52] C. Hori, A. Cherian, T. K. Marks, and T. Hori, "Joint student-teacher learning for audio-visual scene-aware dialog," in *Proc. Interspeech*, Sept. 2019, pp. 1886–1890.

[53] Y. Heo, "Interpretable multimodal dialogue system with natural language-based multimodal integration," in *Proc. AAAI-DSTC10*, 2022.

[54] Y. Yamazaki, S. Orihashi, R. Masumura, M. Uchida, and A. Takashima, "Audio visual scene-aware dialog generation withtransformer-based video representations," in *Proc. AAAI-DSTC10*, 2022.

[55] X. Huang, H. L. Tan, M. C. Leong, Y. Sun, L. Li, R. Jiang, and J. J. Kim, "Investigation on transformer-based multi-modal fusionfor audio-visual scene-aware dialog," in *Proceedings of DSTC10 Workshop at AAAI-2022*, 2022.

[56] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proc. IJCNLP*, Nov. 2017, pp. 986–995.

[57] K. Gopalakrishnan, B. Hedayatnia, Q. Chen, A. Gottardi, S. Kwatra, A. Venkatesh, R. Gabriel, D. Hakkani-Tür, and A. A. AI, "Topical-chat: Towards knowledge-grounded open-domain conversations." in *Proc. Interspeech*, 2019, pp. 1891–1895.

[58] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" in *Proc. ACL*, July 2018, pp. 2204–2213.

[59] C. Zhang, J. Sedoc, L. F. D'Haro, R. Banchs, and A. Rudnicky, "Automatic evaluation and moderation of open-domain dialogue systems," *arXiv preprint arXiv:2111.02110*, 2021.

[60] R. E. Banchs, "Movie-dic: a movie dialogue corpus for research and development," in *Proc. ACL*, 2012, pp. 203–207.

[61] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs," *arXiv preprint arXiv:1106.3077*, 2011.

[62] J. Li, B. Peng, S. Lee, J. Gao, R. Takanobu, Q. Zhu, M. Huang, H. Schulz, A. Atkinson, and M. Adada, "Results of the multi-domain task-completion dialog challenge," in *Proc. AAAI-DSTC8*, 2020.

[63] I. Annamoradnejad and G. Zoghi, "Colbert: Using bert sentence embedding for humor detection," *arXiv preprint arXiv:2004.12765*, 2020.

[64] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "Carer: Contextualized affect representations for emotion recognition," in *Proc. EMNLP*, 2018, pp. 3687–3697.

[65] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proc. ACL*, July 2019, pp. 5370–5381.

[66] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, L.-W. Ku, *et al.*, "Emotionlines: An emotion corpus of multi-party conversations," *arXiv preprint arXiv:1802.08379*, 2018.

[67] M. Rodríguez-Cantelar, D. de la Cal, M. Estecha, A. G. Gutiérrez, D. Martín, N. R. N. Milara, R. M. Jiménez, and L. F. D'Haro, "Genuine^2: An open domain chatbot based on generative models," *Proceedings Alexa Socialbot Grand Challenge SGC4*, 2021.

[68] C. Zhang, L. F. D'Haro, R. E. Banchs, T. Friedrichs, and H. Li, *Deep AM-FM: Toolkit for Automatic Dialogue Evaluation*. Singapore: Springer Singapore, 2021, pp. 53–69.

[69] B. Pang, E. Nijkamp, W. Han, L. Zhou, Y. Liu, and K. Tu, "Towards holistic and automatic evaluation of open-domain dialogue generation," in *Proc. ACL*, July 2020, pp. 3619–3629.

[70] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *Proc. ACL*, July 2020, pp. 270–278.

[71] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. ACL*, 2002, pp. 311–318.

[72] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," 2004, pp. 74–81.

[73] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *Proc. ICLR*, 2020.
[74] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proc. ACL*, July 2020, pp. 7881–7892.
[75] Y.-T. Yeh, M. Eskenazi, and S. Mehri, "A comprehensive assessment of dialog evaluation metrics," in *Proc. EANCS*, Online, 2021, pp. 15–33.
[76] J. Xu, A. Szlam, and J. Weston, "Beyond goldfish memory: Long-term open-domain conversation," *arXiv preprint arXiv:2107.07567*, 2021.
[77] M. Komeili, K. Shuster, and J. Weston, "Internet-augmented dialogue generation," *arXiv preprint arXiv:2107.07566*, 2021.
[78] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Proc. NeurIPS*, vol. 33, 2020, pp. 1877–1901.
[79] S. Mehri, J. Choi, L. F. D'Haro, J. Deriu, M. Eskenazi, M. Gasic, K. Georgila, D. Hakkani-Tur, Z. Li, V. Rieser, S. Shaikh, D. Traum, Y.-T. Yeh, Z. Yu, Y. Zhang, and C. Zhang, "Report from the NSF future directions workshop on automatic evaluation of dialog: Research directions and challenges," 2022.