Overview of the Ninth Dialog System Technology Challenge: DSTC9

Chulakun Gunasekara, Seokhwan Kim, Senior Member, IEEE, Luis Fernando D’Haro, Member, IEEE, Abhinav Rastogi, Yun-Nung Chen, Member, IEEE, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkani-Tür, Fellow, IEEE, Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Minlie Huang, Senior Member, IEEE, Jianfeng Gao, Fellow, IEEE, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, Maxine Eskenazi, Ahmad Beirami, Eunjoon Cho, Paul A. Crook, Senior Member, IEEE, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar, and Rajen Subba

Abstract—This paper introduces the Ninth Dialog System Technology Challenge (DSTC-9). This edition of the DSTC focuses on applying end-to-end dialog technologies for four distinct tasks in dialog systems, namely, 1. Task-oriented dialog Modeling with Unstructured Knowledge Access, 2. Multi-domain task-oriented dialog, 3. Interactive evaluation of dialog and 4. Situated interactive multimodal dialog. 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.

Index Terms—Artificial intelligence, computational intelligence, chatbots, computational linguistics, human-machine systems, human computer interaction, natural language processing, natural language generation, personal voice assistants.

Manuscript received 21 November 2022; revised 11 August 2023; accepted 13 November 2023. Date of publication 11 July 2024; date of current version 16 September 2024. This work was supported in part by ASTOUND under Grant 101071191 HORIZON-EIC-2021-PATHFINDERCHALLENGES-01, in part by BEWORD under Grant PID2021-126061OB-C43, and in part by GENIUS under Grant APOYO-JOVENES-21-TAXTYC-32-K61X37. (All authors contributed equally to this work.) (Corresponding authors: Chulakun Gunasekara; Luis Fernando D’Haro.)

Chulakun Gunasekara is with the IBM Research-Yorktown Heights, NY 10598 USA (e-mail: chulakun.gunasekara@ibm.com).
Seokhwan Kim, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, and Yang Liu are with the Amazon Alexa AI, Enterprise Way, Sunnyvale, CA 94089 USA (e-mail: dr.seokhwan.kim@ieee.org; mihail911@gmail.com; ben.hedayat@gmail.com; karthkop@amazon.com; yangliud@amazon.com).
Luis Fernando D’Haro is with the Speech Technology and Machine Learning Group - Electronic Engineer Dept., ETSI de Telecomunicación - Universidad Politécnica de Madrid, Madrid 28040, Spain (e-mail: luisfernando.dharo@upm.es).
Abhinav Rastogi, Ahmad Beirami, and Shivani Poddar are with the Google DeepMind, Google, Mountain View, California 94041 USA (e-mail: abhrastagi@google.com; ahmad.beirami@gmail.com; shivani@fb.com).
Yun-Nung Chen and Chao-Wei Huang are with the National Taiwan University, Taipei 10617, Taiwan (e-mail: y.v.chen@ieee.org; f07922069@csie.ntu.edu.tw).
Dilek Hakkani-Tür is with the University of Illinois Urbana-Champaign, Urbana, IL 61801 USA (e-mail: dilek@illinois.edu).
Jinchao Li is with the Snap Inc., 110th Ave. NE, Bellevue, WA 98004, US (e-mail: lijinchao@ucla.edu).
Qi Zhu, Lingxiao Luo, Runze Liang, Zheng Zhang, and Minlie Huang are with the Tsinghua University, Beijing 100084, China (e-mail: luoz21@mails.tsinghua.edu.cn; z-zhang@tsinghua.edu.cn; alhuang@tsinghua.edu.cn).

I. INTRODUCTION

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

The first Dialogue System Technology Challenge [1] used human-to-bot dialogs in the bus timetable domain. Dialog State Tracking Challenges 2 [2] and 3 [3] used restaurant reservation
application, 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 dialogues in mono and cross-language settings. From the sixth challenge [6], the DSTC rebranded itself as the “Dialog System Technology Challenge” and organized multiple tracks in parallel to address a wider variety of dialog-related problems. The tracks in DSTC-6 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], [9], grounded response generation [10], and audio visual scene-aware dialog [11]. More recently in DSTC-8 [12] the focus has been on a diverse set of four tracks, including multidomain task completion, predicting responses, audio visual scene-aware dialog and schema-guided dialog state tracking.

For the ninth edition of the DSTC, we received nine track proposals from the leading research organizations and the top universities. The proposals went through a formal peer review process focusing on each task’s potential for (a) impact to the community, (b) novelty of the task, (c) feasibility of the proposal, and (d) potential participants. DSTC-8 participants 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 the four main tracks, including three newly introduced tasks and one follow-up task from DSTC-8.

The track, Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access (Track 1), aims to take the first step in expanding dialog research beyond datasets and challenges participants to develop dialog systems that can converse effectively in interactive environments with real users. SIMMC: Situated Interactive Multimodal Conversational AI (track 4) aims to lay the foundations for real-world assistant agents that can handle multimodal inputs and perform multimodal actions.

The following sections describe the details of each track.

II. TRACK 1 - BEYOND DOMAIN APIs - TASK-ORIENTED CONVERSATIONAL MODELING WITH UNSTRUCTURED KNOWLEDGE ACCESS

A. Track Overview

Most prior work on task-oriented dialog systems has been restricted to a limited coverage of domain APIs. However, users often have domain-related requests that are not covered by the APIs. This challenge track aims to expand the coverage of task-oriented dialog systems by incorporating external unstructured knowledge sources. There are three main tasks on this track as introduced in [13]: knowledge-seeking turn detection, knowledge selection, and knowledge-grounded response generation (Table I).

B. Data

This challenge track uses two different data sets (Table II). The first data is an augmented version of MultiWOZ 2.1 [14] that includes newly introduced knowledge-seeking turns in the MultiWOZ conversations. Data augmentation was done incrementally using crowdsourcing tasks described in [13]. A total of 22,834 utterance pairs were newly collected based on 2,900 knowledge candidates from the FAQ webpages about the domains and the entities in MultiWOZ databases. For the challenge track, we divided the whole data into three subsets: train, validation, and test. The first two sets were released in the development phase along with the ground-truth annotations and human responses for participants to develop their models.

In the evaluation phase, we released the test split of the augmented MultiWOZ 2.1 and the other conversations collected from scratch about tourist information for San Francisco. To evaluate the generalizability of models, the new conversations cover knowledge, locale, and domains that are unseen from the train and validation data sets. In addition, this test set includes not only written conversations, but also spoken dialogs to evaluate system performance across different modalities [15]. All back-end resources for this data collection were also released, which

| Task #1 Knowledge-seeking Turn Detection |
|------------------------------------------|
| Goal To decide whether to continue existing flow or trigger the knowledge access branch for a given utterance and dialog history |
| Input Current user utterance, dialog context, and domain API and knowledge sources |
| Output Binary class (requires knowledge access or not) |

| Task #2 Knowledge Selection |
|----------------------------|
| Goal To select proper knowledge sources from the domain knowledge-base given dialog context at each turn with knowledge access |
| Input Current user utterance, dialog context, and the entire set of knowledge candidates |
| Output Ranking of top-k knowledge candidates |

| Task #3 Knowledge-grounded Response Generation |
|-----------------------------------------------|
| Goal To generate a system response for a given triple of input utterance, dialog context, and the selected knowledge sources |
| Input Current user utterance, dialog context, and selected knowledge sources |
| Output Generated system response |

| TABLE I SUMMARY OF TRACK 1 TASKS |
|----------------------------------|

| Task | Description |
|------|-------------|
| #1   | Knowledge-seeking Turn Detection |
| #2   | Knowledge Selection |
| #3   | Knowledge-grounded Response Generation |

| TABLE II STATISTICS OF THE TRACK 1 DATA SETS |
|---------------------------------------------|
| Source | Split | # dialogs | total instances | # knowledge seeking turns |
|-------|-------|-----------|----------------|--------------------------|
| Train | 7,190 | 71,348 | 19,184 |
| MultiWOZ | Valid | 1,000 | 9,663 | 2,673 |
|          | Test  | 977    | 2,084 | 977  |
|        | SF    | 900    | 1,834 | 900  |
|        | Spoken| 107    | 263   | 104  |
includes 9,139 knowledge snippets and 855 database entries for San Francisco.

C. Evaluation Criteria

Each participating team submitted up to five system outputs, each of which contains the results for all three tasks on the unlabeled test instances. We first evaluated each submission using the task-specific objective metrics (Table III) by comparing to the ground-truth labels and responses. Considering the dependencies between the tasks in the pipelined architecture, the final scores for knowledge selection and knowledge-grounded response generation are computed by considering the first step knowledge-seeking turn detection recall and precision performance, as follows:

\[ f(x) = \begin{cases} 1 & \text{if } x \text{ is a knowledge-seeking turn}, \\ 0 & \text{otherwise} \end{cases} \]

\[ \tilde{f}(x) = \begin{cases} 1 & \text{if } x \text{ is predicted as a knowledge-seeking turn}, \\ 0 & \text{otherwise} \end{cases} \]

\[ S_p(X) = \frac{\sum_{x_i \in X} (s(x_i) \cdot f(x_i) \cdot \tilde{f}(x_i))}{\sum_{x_i \in X} f(x_i)}, \]

\[ S_r(X) = \frac{\sum_{x_i \in X} (s(x_i) \cdot f(x_i) \cdot \tilde{f}(x_i))}{\sum_{x_i \in X} f(x_i)}, \]

\[ S_f(X) = \frac{2 \cdot S_p(X) \cdot S_r(X)}{S_p(X) + S_r(X)}, \] \quad (1)

where \( s(x) \) is the knowledge selection or response generation score in a target metric for a single instance \( x \in X \).

Then, we aggregated a set of multiple scores across different tasks and metrics into a single overall score computed by the mean reciprocal rank, as follows:

\[ S_{\text{overall}}(e) = \frac{1}{|M|} \sum_{i=1}^{|M|} \frac{1}{\text{rank}_i(e)}, \] \quad (2)

where \( \text{rank}_i(e) \) is the ranking of the submitted entry \( e \) in the \( i \)-th metric against all the other submissions and \( M \) is the number of metrics we considered.

Based on the overall objective score, we selected the finalists to be manually evaluated by the following 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.

In both tasks, we assigned each instance to three crowd workers and took their average as the final human evaluation score for the instance. These scores were then aggregated over the entire test set following (1), i.e., weighted by the knowledge-seeking turn detection performance. Finally, we used the average of the Appropriateness and Accuracy scores to determine the official ranking of the systems in the challenge track.

D. Results

We received 105 entries in total submitted from 24 participating teams. To preserve anonymity, the teams were identified by numbers from 1 to 24, while our baseline [13] was marked as team 0. Table IV shows the objective evaluation results of the best entry from each team in the featured metrics. The full scores with all the submitted entries and the other metrics are available on the track repository. Most entries outperformed the baseline in all three tasks. In particular, the best entry from Team 3 achieved over 99% F-measure for knowledge-seeking turn detection, and also the highest scores in the BLEU and ROUGE variants for the response generation task. On the other hand, Team 19 was the best in the knowledge selection metrics, and Team 15 was better than all the other teams in METEOR for generation. We calculated the overall score (2) of each entry.

1https://github.com/alexa/alexa-with-dstc9-track1-dataset

| Task | Metrics |
|------|---------|
| Task #1 | Precision/Recall/F-measure |
| Task #2 | MKR@5, Recall@1, Recall@5 |
| Task #3 | BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-1, ROUGE-2, ROUGE-L |

| Table III | Objective Evaluation Metrics for the Track 1 Tasks |
|-----------|-----------------------------------------------|
| Task #1   | Precision/Recall/F-measure                     |
| Task #2   | MKR@5, Recall@1, Recall@5                     |
| Task #3   | BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-1, ROUGE-2, ROUGE-L |

| Table IV | Objective Evaluation Results of the Best Entry From Each Team in the Featured Metrics for the Track 1 Tasks. Team 0 is the Baseline |
|----------|-------------------------------------------------|
| Team     | Task #1 | Task #2 | Task #3 |
|          | F1      | R@1    | BLEU-1  | METEOR | ROUGE-L |
| 1        | 0.9721  | 0.8255 | 0.3368  | 0.3324 | 0.3364  |
| 2        | 0.9644  | 0.8584 | 0.3338  | 0.3322 | 0.3330  |
| 3*       | 0.9911  | 0.9013 | 0.3879  | 0.3914 | 0.3885  |
| 4        | 0.8998  | 0.6950 | 0.3025  | 0.3001 | 0.2990  |
| 5        | 0.9428  | 0.7055 | 0.3218  | 0.3266 | 0.3174  |
| 6        | 0.9838  | 0.8531 | 0.3371  | 0.3341 | 0.3326  |
| 7        | 0.9702  | 0.8988 | 0.3752  | 0.3854 | 0.3702  |
| 8        | 0.9530  | 0.7403 | 0.3135  | 0.3097 | 0.3119  |
| 9        | 0.9242  | 0.7931 | 0.3154  | 0.3159 | 0.3101  |
| 10*      | 0.9726  | 0.9158 | 0.3684  | 0.3719 | 0.3692  |
| 11*      | 0.9675  | 0.8702 | 0.3743  | 0.3854 | 0.3797  |
| 12       | 0.9540  | 0.8011 | 0.3411  | 0.3526 | 0.3401  |
| 13*      | 0.9819  | 0.8434 | 0.3787  | 0.3902 | 0.3619  |
| 14       | 0.9252  | 0.6466 | 0.3019  | 0.2974 | 0.3003  |
| 15*      | 0.9803  | 0.8975 | 0.3779  | 0.3931 | 0.3765  |
| 16       | 0.9549  | 0.7327 | 0.3351  | 0.3334 | 0.3364  |
| 17*      | 0.9839  | 0.8734 | 0.3699  | 0.3724 | 0.3687  |
| 18*      | 0.9635  | 0.8994 | 0.3806  | 0.3864 | 0.3726  |
| 19*      | 0.9886  | 0.9235 | 0.3803  | 0.3869 | 0.3738  |
| 20*      | 0.9711  | 0.8628 | 0.3619  | 0.3637 | 0.3535  |
| 21*      | 0.9396  | 0.8269 | 0.3551  | 0.3594 | 0.3576  |
| 22       | 0.7803  | 0.2831 | 0.1792  | 0.1583 | 0.1852  |
| 23*      | 0.9618  | 0.8959 | 0.3534  | 0.3565 | 0.3519  |
| 24       | 0.3471  | 0.0017 | 0.0835  | 0.0796 | 0.0866  |

Bold denotes the best result in each column and * indicates the finalists.
The dialog system is evaluated by comparing the performance of component-wise models using corpus-based evaluation and that of the entire system using the end-to-end evaluation. These findings are consistent with the landscape of dialog development technology stacks in the industry. However, interestingly, the winning team built their model based on GPT-2 [19], and achieved a significant improvement over other teams with respect to the success rate, the understanding score and the response score in the human evaluation phase. Meanwhile, by using similar model training paradigms, SOLOIST [20] and SimpleTOD [21] shortly achieved top performance in the MultiWOZ leaderboard by leveraging GPT-2 [22].

This year, we continue with the end-to-end task-oriented dialog task, with the aim of promoting the technology of building end-to-end dialog systems one step further. Like last year, participants are encouraged to explore all possible approaches and there is no restriction on the architecture of the dialog system.

1) Data: Participants are expected to build dialog systems based on MultiWOZ 2.1 [14], a multi-domain dialog dataset spanning 7 distinct domains, containing over 10,000 dialogs under the travel planning setting. Compared to MultiWOZ 2.0 [23], MultiWOZ 2.1 re-annotated states to fix the noisy annotation and incorporated user dialog act annotation. Although the dialog system is evaluated under MultiWOZ 2.1, participants can leverage any public datasets, pre-trained models, or other resources to build the dialog system.

2) Evaluation Criteria: ConvLab-2 [24] is employed as the platform for dialog development and evaluation. As the successor of ConvLab [25], ConvLab-2 provides a user simulator and evaluator for MultiWOZ 2.1 so that participants can effectively run offline experiments and evaluations. Specifically, we offer two evaluation approaches:

a) Automatic Evaluation: The dialog system is evaluated via conversing with an end-to-end user simulator. The simulator is constructed by assembling a BERT-based natural language understanding model [26], an agenda-based user simulator [27] and a rule-based natural language generation module. A dialog is successful only if all requested slots are filled with grounded values in the database, and the booking is successful. We report metrics including success rate, book rate, number of turns for each dialog, and precision/recall/F1 score for slot prediction.

b) Human Evaluation: In human evaluation, Amazon Mechanical Turkers (MTurkers) communicate with the dialog systems via natural language, judge whether the dialog is successful, and provide scores based on correctness for language understanding and appropriateness of responses on a 5-point Likert scale. Since MTurkers do not directly access the back-end database, we also report the success rate with grounding after verifying whether the requested slot values returned by the dialog systems match the database record. We take the average value of the success rate with grounding and without grounding as the final ranking.

3) Results: As per our submission policy, each team is allowed to submit up to 5 models. We received 34 models in total from 10 teams. Table VI lists the automatic evaluation result for the best models of each team. We filter out low-performance models based on the automatic evaluation result while keeping the best model for each team, and sent the remaining models for evaluation.

Table V shows the final ranking of the Track 1 participating teams based on the human evaluation scores of the finalist entries. The top three teams (Team 19, 3 and 10) commonly used ensemble of large-scale pre-trained language models in their best entries. Team 19 won the challenge track with the highest scores for both Accuracy and Appropriateness, most likely due to their better performance in the knowledge selection task (as in the objective evaluation results). To compare the importance of each task towards end-to-end performance, we calculated the Spearman’s rank correlation coefficient of the ranked lists of all the entries in every pair of objective and human evaluation metrics. As a result, Recall@1 for the knowledge selection task shows a strong correlation with the averaged human evaluation ranking at 0.8601, which is significantly higher than 0.7692 and 0.6503 with F-measure for the knowledge-seeking turn detection and BLEU-1 for response generation, respectively. This implies that knowledge selection is a key task to improve end-to-end performance.

| Rank | Team | Entry | Accuracy | Appropriateness | Average |
|------|------|-------|----------|----------------|---------|
| 1    | 19   | 2     | 4.3917   | 4.3922         | 4.3920  |
| 2    | 3    | 1     | 4.3480   | 4.3634         | 4.3557  |
| 3    | 10   | 0     | 4.3544   | 4.3201         | 4.3373  |
| 4    | 15   | 3     | 4.3793   | 4.2755         | 4.3274  |
| 5    | 17   | 0     | 4.3360   | 4.3076         | 4.3218  |
| 6    | 7    | 4     | 4.3308   | 4.2989         | 4.3149  |
| 7    | 18   | 3     | 4.3309   | 4.2859         | 4.3084  |
| 8    | 13   | 3     | 4.3763   | 4.2360         | 4.3061  |
| 9    | 23   | 0     | 4.3082   | 4.2665         | 4.2874  |
| 10   | 11   | 3     | 4.2722   | 4.2619         | 4.2670  |
| 11   | 20   | 4     | 4.2283   | 4.2486         | 4.2384  |
| 12   | 21   | 3     | 4.1060   | 4.1560         | 4.1310  |
| Baseline | | | 3.7135 | 3.9386 | 3.8271 |

and selected 12 finalists, corresponding to the best entry from each of the top 12 teams.

Table V shows the final ranking of the Track 1 participating teams based on the human evaluation scores of the finalist entries. The top three teams (Team 19, 3 and 10) commonly used ensemble of large-scale pre-trained language models in their best entries. Team 19 won the challenge track with the highest scores for both Accuracy and Appropriateness, most likely due to their better performance in the knowledge selection task (as in the objective evaluation results). To compare the importance of each task towards end-to-end performance, we calculated the Spearman’s rank correlation coefficient of the ranked lists of all the entries in every pair of objective and human evaluation metrics. As a result, Recall@1 for the knowledge selection task shows a strong correlation with the averaged human evaluation ranking at 0.8601, which is significantly higher than 0.7692 and 0.6503 with F-measure for the knowledge-seeking turn detection and BLEU-1 for response generation, respectively. This implies that knowledge selection is a key task to improve end-to-end performance.

III. TRACK 2 - MULTI-DOMAIN TASK-ORIENTED DIALOG CHALLENGE II

We provide two tasks in the multidomain task-oriented dialog setting. One is the end-to-end task-oriented dialog task aiming at solving the complexity of building end-to-end dialog systems. The other is cross-lingual dialog state tracking (DST) to address the language adaptation problem for the DST task.

A. End-to-End Task-Oriented Dialog Task

This task is a continuation of last year at DSTC8 [16]. Participants will develop an end-to-end task-oriented dialog system that takes natural language as input and generates natural language response as output in the travel planning setting. Both the evaluation result of last year’s challenge [17] and empirical analysis of models in ConvLab [18] show that the best rule-based pipeline systems outperform systems assembled using state-of-the-art component-wise machine learning models. From the results, we have also observed a discrepancy between

and selected 12 finalists, corresponding to the best entry from each of the top 12 teams.

Table V shows the final ranking of the Track 1 participating teams based on the human evaluation scores of the finalist entries. The top three teams (Team 19, 3 and 10) commonly used ensemble of large-scale pre-trained language models in their best entries. Team 19 won the challenge track with the highest scores for both Accuracy and Appropriateness, most likely due to their better performance in the knowledge selection task (as in the objective evaluation results). To compare the importance of each task towards end-to-end performance, we calculated the Spearman’s rank correlation coefficient of the ranked lists of all the entries in every pair of objective and human evaluation metrics. As a result, Recall@1 for the knowledge selection task shows a strong correlation with the averaged human evaluation ranking at 0.8601, which is significantly higher than 0.7692 and 0.6503 with F-measure for the knowledge-seeking turn detection and BLEU-1 for response generation, respectively. This implies that knowledge selection is a key task to improve end-to-end performance.

III. TRACK 2 - MULTI-DOMAIN TASK-ORIENTED DIALOG CHALLENGE II

We provide two tasks in the multidomain task-oriented dialog setting. One is the end-to-end task-oriented dialog task aiming at solving the complexity of building end-to-end dialog systems. The other is cross-lingual dialog state tracking (DST) to address the language adaptation problem for the DST task.

A. End-to-End Task-Oriented Dialog Task

This task is a continuation of last year at DSTC8 [16]. Participants will develop an end-to-end task-oriented dialog system that takes natural language as input and generates natural language response as output in the travel planning setting. Both the evaluation result of last year’s challenge [17] and empirical analysis of models in ConvLab [18] show that the best rule-based pipeline systems outperform systems assembled using state-of-the-art component-wise machine learning models. From the results, we have also observed a discrepancy between
Compared with previous datasets [5], [32], VII and [14], 2. To collect new dialogs, we and Cross-WOZ. In DSTC8, out of 11 teams with valid submissions, 1 team uses cross-lingual transfer from English to Chinese using MultiWOZ 2.1 dataset and 2) cross-lingual transfer from Chinese to English using CrossWOZ dataset.

Following a similar scheme as in DSTC-5 [5], we provide machine translations of the original dataset. We collected 500 new dialogs in the target language as the test set. The performance of each dialog state tracker is evaluated on the test set and compared with reference annotation.

1) Data: Compared with previous datasets [5], [31], [32] for cross-lingual transfer learning in task-oriented dialog, MultiWOZ 2.1 and CrossWOZ are much larger. MultiWOZ 2.1 contains over 10,000 dialogs, and CrossWOZ contains over 6,000 dialogs. They are also more challenging due to the multi-domain setting. For each subtask, we prepared data in a similar way: a) collected 500 new dialogs in the source language, b) translated the ontology to the target language, and c) translated the original dialogs and the new dialogs. We released 250 new dialogs without any annotation as a public test set and reserved the other 250 dialogs as a private test set.

a) Test Data Collection: To collect new dialogs, we adapted the data collection website of CrossWOZ where paired workers can converse synchronously and make annotations. New user goals were generated by the ConvLab-2 goal generator. Following the Wizard-of-Oz setting, one worker acts as the user who needs to accomplish the allocated goal, and the other acts as the system that uses the database to provide information. During the conversation, both sides need to annotate the dialog acts of their utterances, and the system should also annotate the dialog states that are queries over the database.

b) Ontology Translation: We extracted the ontology from the dialog act and the dialog state annotations of the original and test datasets. We then used Google Translate to translate them into the target language. For some slots that may not be faithfully translated, such as “name” and “address”, we employed human translators to correct the translations. This process is vital to ensure the translation consistency of the same values in different contexts.

c) Dialog Translation: To make sure that the translations of values in a dialog are faithful to the ontology dictionary, we first replaced the values that appeared in the dialog with their translations in the dictionary. We then used Google Translate to translate the resulting code-switching sentences from the original dataset and test set. In this way, the translated dialogs achieved much better performance in human evaluation than the systems at DSTC8.²

### TABLE VI
**Automatic Evaluation Result (Best Submissions)**

| Team | SR  | CR  | BR  | Inform | Turn | S/A |
|------|-----|-----|-----|--------|------|-----|
| 1    | 93  | 95.2| 94.6| 84.1/96.2/88.1 |       | 12.5/12.7 | 
| 2    | 91.4| 96.9| 96.2| 80.2/97.3/86.0 |       | 15.3/15.7 | 
| 3    | 90.8| 94.4| 96.7| 81.0/95.4/85.9 |       | 13.4/13.6 | 
| 4    | 89.8| 94.6| 96.3| 72.4/96.8/80.1 |       | 15.1/15.8 | 
| 5    | 83.3| 88.5| 89.1| 81.9/90.3/83.5 |       | 13.5/13.8 | 
| 6    | 67.7| 88.5| 90.8| 70.4/85.6/75.2 |       | 12.8/14.2 | 
| 7    | 57.8| 87.1| 85. | 68.7/81.6/72.6 |       | 13.7/16.4 | 
| 8    | 52.6| 66.9| 66.7| 57.5/80.7/64.8 |       | 13.2/22.3 | 
| 9    | 44.4| 50  | 26.5| 57.9/64.5/58.9 |       | 12.2/14.5 | 
| 9    | 21.4| 40.7| 0   | 55.4/60.0/64.1 |       | 11.0/25.9 | 
| Baseline | 85  | 92.4| 91.4| 79.3/94.9/84.5 |       | 13.8/14.9 | 

²The success rate in DSTC8 human evaluation is success rate w/o grounding.

### TABLE VII
**Human Evaluation Result (Best Submissions)**

| Team | SRa | SRw | SRog | Under | Appr | Turn | Rank |
|------|-----|-----|------|-------|------|------|------|
| 1    | 74.8| 70.2| 79.4 | 4.54  | 4.47 | 18.5 | 1    |
| 2    | 74.8| 68.8| 80.8 | 4.51  | 4.45 | 19.4 | 1    |
| 3    | 72.3| 62  | 82.6 | 4.53  | 4.41 | 17.1 | 3    |
| 4    | 70.6| 60.8| 80.4 | 4.41  | 4.41 | 20.1 | 4    |
| 5    | 67.8| 60  | 75.6 | 4.56  | 4.42 | 21   | 5    |
| 6    | 60.3| 51.4| 69.2 | 4.49  | 4.22 | 17.7 | 6    |
| 7    | 58.4| 50  | 66.4 | 4.15  | 4.06 | 19.7 | 7    |
| 8    | 55.2| 43.2| 67.2 | 4.15  | 3.98 | 19.2 | 8    |
| 9    | 35  | 26  | 44   | 3.27  | 3.15 | 18.5 | 9    |
| 10   | 35  | 33  | 32.3 | 2.93  | 2.93 | 18.8 | 10   |
| Baseline | 69.6| 56.8| 82.4 | 4.34  | 4.18 | 18.5 | N/A  |

SR: success rate. CR: complete rate. BR: book rate. Inform: F/R/Fl score of slot prediction. Turn S/A: Turn for successful and all dialogues, respectively.

### Summary

Team 1 achieves the top 1 performance in both automatic and human evaluation by constructing an end-to-end dialog system with the pre-trained dialog generation model PLATO-2. This model generates the dialog state, system action, and system response simultaneously, given the dialog context. The dialog state is used as the constraint for database query, and the system action is then refreshed according to the queried results to regenerate the final system response. Team 2 achieves the same ranking as Team 1 in the human evaluation using a similar hybrid end-to-end neural model. It borrows ideas from [20] and [29], uses GPT-2 as the back-end for pretraining and fine-tuning, and adds various pre/post-processing modules to improve model generalization ability. An additional fault tolerance mechanism is also added to correct errors.

4) Summary: Compared with the challenge results at DSTC8, there is a trend of shifting from building dialogs by assembling component-wise modules to end-to-end learning. In DSTC8, out of 11 teams with valid submissions, 1 team uses GPT-2 based models, 1 team uses word DST + word policy, with the rest 9 teams use component-wise models. This year, out of 10 teams, 8 teams used the end-to-end learning mechanism by leveraging transformer-based models. The top three systems in both automatic evaluation and human evaluation are all built using transformer-based end-to-end learning, and they have achieved much better performance in human evaluation than the systems at DSTC8.
and the corresponding annotations do not conflict. 250 dialogs were sampled from the original dataset as the development set. Human translators were employed to proofread the translations of the development and test set.

2) Evaluation Criteria: We evaluate the performance of the dialog state tracker using the following metrics: a) Joint Goal Accuracy. This metric evaluates whether the predicted dialog state is exactly equal to the ground truth. b) Slot Accuracy. This metric evaluates whether each slot’s predicted label is exactly equal to the ground truth, averaged over all slots. c) Slot Precision/Recall/F1. These metrics evaluate the overlap between the predicted labels and the ground truth for non-empty slots, micro-averaged over dialog turns. Each submission contains the predictions for the public test set and the model that is used to make predictions for the private test set. The results are averaged over the public and private test sets. The final ranking is solely based on the joint goal accuracy.

3) Results: The results of MultiWOZ (en→zh) and CrossWOZ (zh→en) subtasks are shown in Tables VIII and IX respectively. During the evaluation, we found that the CrossWOZ test data miss many “name” labels when the user accepts the attraction/hotel/restaurant recommended by the system. Therefore, we utilized the database search results and heuristic rules to correct empty “name” labels and provided an updated leaderboard for CrossWOZ in Table X. Both of the CrossWOZ leaderboards are valid, but the updated one is preferred.

We adapted SUMBT [33] as the baseline model and used the translated training set of the original dataset to train for both subtasks. We have received 10 models for MultiWOZ (en→zh) and 8 models for CrossWOZ (zh→en) from the same 3 teams. We briefly introduce their best models here. Team 1 incorporated a four-class state operation prediction task into the CHAN model [34]. Team 2 modified SOM-DST [35] and used ontology and some handcraft rules to postprocess the generated values. Team 3 formulated the dialog state tracking as a sequence generation problem and used mBART to generate pairs of slot names and slot values. All of their best models were trained using the translated data in the target language. Although the best teams (measured by joint goal accuracy, JGA) are different in three leaderboards, Team 2 consistently outperforms the other two teams in slot F1, which is a smoother metric than JGA. Recently, there have been some discussions about the limitation of the JGA [36], [37]. However, at the time of the challenge, JGA is the mainstream metric.

4) Summary: To our surprise, all the best models are trained on monolingual machine translated data instead of both the original data and translations. Team 2 and 3 even got negative results when training XLM/mBART on the original data and the translations simultaneously. The performance of “Translate-Train” partially depends on the machine translator, which may be why team 1 and 2 augment the data by using another translator to translate the original dataset. Team 1 and 2 modified DST models that are state-of-the-art on English MultiWOZ 2.1 dataset and got strong performance on Chinese MultiWOZ 2.1, verifying these models’ language portability.

### IV. TRACK 3: INTERACTIVE EVALUATION OF DIALOG

#### A. Track Overview

The purpose of dialog research is to create systems that can be used effectively in interactive settings by real users [38]. Despite this, the majority of research is done on static datasets. For example, the task of response generation is usually done by producing a response for a static dialog context [39]. This track is intended to move dialog research beyond datasets and evaluate models in interactive environments with real users.

This track consists of two subtasks: (1) static evaluation and (2) interactive evaluation. The first subtask challenges participants to build response generation models that are evaluated in a static manner, using the Topical-Chat corpus [40]. The second subtask aims to extend dialog models beyond datasets and assess them in an interactive setting with real users, using DialPort [41]. In the first subtask, models must generate a response to a fixed dialog context. In contrast, in the second subtask, they must have a back-and-forth interaction with a real user. Through the two subtasks, this track challenges participants to take strong response generation models and develop strategies to make them effective in interactive settings.

#### B. Data

Participants in this track were free to train on any publicly available data or use any pre-trained models. The static evaluation in the first subtask was carried out on the Topical-Chat corpus [40]. Topical-Chat is a large collection of human-human knowledge-grounded open-domain conversations that consists of 11,319 dialogs and 248,014 utterances. For each conversational turn, several relevant facts are provided. Models must

---

**TABLE VIII**

| MultiWOZ Leaderboard | (Best Submissions) |
|-----------------------|--------------------|
| Team | JGA | SA | Slot Precision/Recall/F1 | JGA/pub/prti | Rank |
| 1 | 62.37 | 98.09 | 92.15/94.02/93.07 | 62.70/62.03 | 1 |
| 2 | 62.08 | 98.10 | 90.61/96.20/93.32 | 62.35/60.91 | 2 |
| 3 | 30.13 | 94.40 | 87.0/77/67 | 30.53/29.72 | 3 |
| BS | 55.56 | 97.68 | 92.02/91.10/91.56 | 55.81/55.31 | N/A |

JGA: joint goal accuracy. SA: slot accuracy. Slot Precision/Recall/F1, pub/prti. public/private test set.

**TABLE IX**

| CrossWOZ Leaderboard | (Best Submissions) |
|-----------------------|--------------------|
| Team | JGA | SA | Slot Precision/Recall/F1 | JGA/pub/prti | Rank |
| 1 | 16.86 | 89.11 | 68.26/62.85/65.45 | 16.82/16.89 | 1 |
| 2 | 15.28 | 90.37 | 65.94/78.87/71.82 | 15.19/15.37 | 2 |
| 2 | 13.99 | 91.92 | 72.63/78.20/75.64 | 14.11/13.58 | 3 |
| BS | 7.21 | 85.13 | 55.27/46.15/50.30 | 7.41/7.00 | N/A |

**TABLE X**

| CrossWOZ Leaderboard | (Updated Evaluation, Best Submissions) |
|-----------------------|--------------------|
| Team | JGA | SA | Slot Precision/Recall/F1 | JGA/pub/prti | Rank |
| 1 | 32.30 | 94.35 | 81.39/82.25/81.82 | 32.70/31.89 | 1 |
| 2 | 23.96 | 92.94 | 74.6/83.41/78.56 | 23.45/24.47 | 2 |
| 3 | 15.31 | 89.70 | 74.7/64.06/69.01 | 14.25/16.37 | 3 |
| BS | 13.02 | 87.97 | 67.18/52.18/58.74 | 13.30/12.74 | N/A |
leverage these facts and generate a response. This dataset was chosen because it is the largest knowledge-grounded open-domain dataset presently available, to the best of our knowledge. Additionally, the choice of usable facts provides a mechanism for systems to tailor responses to a specific user’s interests.

Since we continuously performed human evaluation over the duration of the challenge and used reference-free evaluation metrics [42], it was not strictly necessary for the models to be trained in the Topical-Chat corpus. A strong pre-trained dialog model may perform well on the first subtask, despite not training on the corpus.

The second subtask was not tied to a dataset. The interactive evaluation was carried out on DialPort3 [41] with real users recruited through Facebook Advertising.

C. Evaluation Criteria

The first subtask was evaluated using ongoing (1) human evaluation and (2) three automatic metrics: METEOR [43], BERTscore [44], and USR [42]. Human evaluation was carried out on Amazon Mechanical Turk with the annotation questionnaire used to collect the FED dataset [45]. During the duration of the challenge, we carried out an evaluation on the Topical-Chat frequent validation set. For human evaluation, 30 context-response pairs were sampled and each one was labeled by 3 annotators. For the final evaluation, we performed an automatic evaluation on the frequent test set and performed a human evaluation on 100 randomly sampled context-response pairs. For the final evaluation, the 100 dialog contexts used for the evaluation were consistent across the different systems.

The evaluation for the second subtask consists of (1) collecting dialogs through conversations with real users on DialPort and (2) post hoc assessment of the collected dialogs. Participants submitted dialog models (via an API) to DialPort. Real users were recruited through Facebook Advertising to interact with the submitted dialog systems. After gathering a sufficient number of conversations, we performed post hoc assessment of the dialogs with the FED metric [45] and human evaluation on Amazon Mechanical Turk with the annotation questionnaire used to collect the FED dataset [45].

Throughout the challenge, our aim was to collect at least 100 conversations for each submitted system, discounting any dialogs with offensive terms (e.g., curse words, racist phrases). For each system, 100 conversations were evaluated with the FED metric and on Amazon Mechanical Turk, with 3 annotators labeling each dialog.

For the final submission, we gather dialogs for all systems over the same time period. Ultimately, given a Facebook Advertising budget of $2500 and 11 systems (including two baselines), we obtained 4651 conversations (after removing offensive dialogs) with a total of 41,640 turns. We consider only the conversations that are at least four turns in length (total of 2960) for the final post hoc assessment. For each system, we carry out a human evaluation with 200 conversations of suitable length. Throughout the challenge, all individuals who interact with the system on DialPort do so for free, of their own volition, thereby avoiding common problems observed with paid users [46].

D. Results

The challenge received 33 submissions to the first subtask and 9 submissions to the second subtask.

Table XI shows the results of the static evaluation on the Topical-Chat corpus [40], for the 10 best performing systems according to the human evaluation. Among the top-ranked systems, System 1 introduces a dialog planning module to model the conversation flow more explicitly [47], and System 2 employs a two-stage training methodology where the second stage is specifically trained to generate diverse and coherent responses [48]. All of the top 10 systems used either pre-trained models or additional data, highlighting the importance of pre-training for open-domain response generation. This observation aligns with previous research, which has seen strong performance in open-domain response generation through the use of large-scale pre-training [49], [50].

In addition to performing ongoing human evaluation throughout the challenge, we assess systems in the first subtask using three evaluation metrics. METEOR [43] and BERTscore [44], are referenced evaluation metrics that compare a generated output to a ground-truth response. On the contrary, USR [42] is a reference free evaluation metric that uses pre-trained models and self-supervised training objectives to estimate the quality of a response. Although none of the evaluation metrics is a perfect predictor of the final ranking, we find that USR better correlates with the system-level human performance (Spearman: 0.35, \( p < 0.05 \)) than either METEOR (Spearman: 0.23, \( p > 0.05 \)) or BERTscore (Spearman: 0.22, \( p > 0.05 \)). The relatively low system-level correlation highlights the importance of performing ongoing human evaluation throughout the challenge.

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

\[^{3}\text{http://dialog.speech.cs.cmu.edu:3000/}^{\text{3}}\]

### Table XI: Results for Subtask 1

| System | METEOR | BERTScore | USR | Human | Rank |
|--------|--------|-----------|-----|-------|------|
| 1      | 9.06   | 84.91     | 4.26| 4.281 | 1    |
| 2      | 13.11  | 86.17     | 4.59| 4.280 | 1    |
| 3      | 6.83   | 84.36     | 3.86| 4.280 | 1    |
| 4      | 8.96   | 85.15     | 4.26| 4.260 | 4    |
| 5      | 12.37  | 86.21     | 4.83| 4.293 | 5    |
| 6      | 12.31  | 86.32     | 4.73| 4.231 | 6    |
| 7      | 13.96  | 86.84     | 4.48| 4.229 | 6    |
| 8      | 12.51  | 85.91     | 4.45| 4.229 | 6    |
| 9      | 12.14  | 85.91     | 4.46| 4.216 | 9    |
| 10     | 10.87  | 85.65     | 4.53| 4.210 | 10   |

For brevity, we only show the top 10 submissions (out of 33) according to the human evaluation. This table only reports the overall USR metric and the overall impression of the response from the human evaluation. The full evaluation results may be found here.
TABLE XII
RESULTS FOR SUBTASK 2

| System | Avg. Turns | P@1 | Human Rank |
|--------|------------|-----|------------|
| 1      | 12.44      | 4.97| 4.15       |
| 2      | **13.47**  | 4.79| 4.14       |
| 3      | 8.89       | 4.61| 4.08       |
| 4      | 9.36       | 4.68| 4.03       |
| 5      | 9.82       | 4.53| 3.93       |
| 6      | 8.75       | 4.72| 3.87       |
| 7      | 8.51       | 4.41| 3.85       |
| 8      | 7.67       | 4.30| 3.85       |
| 9      | 5.53       | 4.64| 3.83       |
| 10     | 7.35       | 4.80| 3.69       |
| 11     | 5.80       | 3.69| 3.60       |

This table reports for each system: the overall P@1 metric, the overall impression of the dialog from the human evaluation, and the average number of dialog turns. The full results can be found here. System 6 and 11 are our DialoGPT and Transformer baselines, respectively.

(METEOR, BERTscore) or not resembling the utterances in the Topical-Chat corpus (USR).

The results for the second subtask are shown in Table XII. System 6 is our DialoGPT baseline [49], fine-tuned on the Topical-Chat corpus without knowledge grounding. System 11 is our Transformer baseline which was trained on the Topical-Chat corpus and uses tf-idf sentence similarity to retrieve relevant knowledge at inference time. The best-performing model, System 1, leverages large-scale pre-training in addition to strategies for producing more diverse responses. This system achieved first place in both subtasks: System 1 in Table XII corresponds to System 2 in Table XI.

FED [45], which is an unsupervised evaluation metric for interactive dialog is shown to be a moderate predictor of the final ranking with a system-level Spearman correlation of 0.49 (p = 0.13), although it correctly predicts the two top systems. In the context of these evaluations, we also observe a strong correlation (Spearman: 0.94, p < 0.01) between the average number of turns per dialog and its final ranking. This indicates that real users are more inclined to engage longer with higher-quality systems, suggesting that user engagement can serve as a proxy for system quality in interactive settings [51], provided sufficient sample size.

In the first subtask, static evaluation, many of the submissions perform similarly as shown by the human evaluation scores in Table XI. This similarity in performance is due to the systems’ similar architectures and methodologies, resembling realistic evaluation of incremental systems. The scores in Table XII, however, are much more varied. This diversity in scores signifies that interactive evaluation provides a more exhaustive testing environment, better identifying and highlighting the weaknesses of the systems. This observation has been shown by prior work [45], when analyzing dialogs from Meena [50].

The Interactive Evaluation of Dialog track demonstrates both the feasibility and the importance of evaluating dialog systems in interactive settings with real users. We show that with an advertising budget of $2500, we collect more than 4000 dialogs on DialPort (2960 dialogs with at least 4 turns). The results of interactive evaluation are more varied (Table XII) suggesting that back-and-forth interactions with real users are challenging for dialog systems and that interactive evaluation is a better reflection of a system’s capabilities.

V. TRACK 4 - SIMMC: SITUATED INTERACTIVE MULTI-MODAL 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. [52] gives more details on the provided datasets and models.

B. Data

SIMMC contains approximately 13 k human-to-human dialogs (totaling about 169 k 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 four 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 appearance logs of each item that appears. SIMMC tasks emphasize semantic processing of the input modalities, while work in this area has traditionally focused heavily on raw image processing.

2) Compared with the conventional task-oriented conversational datasets, the agent actions in the SIMMC datasets span across a diverse multimodal action space (e.g., “rotate”, “search”, and “add to cart”).

3) Agent actions can be enacted on both the object level (e.g. changing the view of a specific object within a scene) and the scene level (e.g. introducing a new scene/image).

4) 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.

Datasets were collected through the SIMMC Platform [53], an extension to ParlAI [54] for multimodal conversational data collection and system evaluation that allows human annotators to each play the role of either the assistant or the user.

C. Evaluation Criteria

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

Subtask 1: Structural API Call Prediction focuses on predicting the assistant action as an API call given the dialog and the multimodal contexts as inputs. Since accuracy does not account for the existence of multiple valid actions, we use perplexity
TABLE XIII
SUMMARY OF EACH TEAM’S RESULTS ON TEST-STD SPLIT, AVERAGE OF FURNITURE AND FASHION

| Teams       | Architecture                                           | 1. API Prediction | 2a. Response Gen | 3. DST |
|-------------|--------------------------------------------------------|-------------------|-----------------|--------|
|             |                                                        | Acc↑ | A.Acc↑ | Perp↓ | BLEU↑ | Slot F1↑ | Intent F1↑ |
| Baseline    | LSTM (subtask 1, 2a.), GPT-2 (subtask 3.)             | 79.3 | 63.7   | 1.9   | 0.061 | 62.4     | 62.1       |
| Team 1      | GPT-2 + fully connected layer                          | 80.2 | 74.6   | 2.0   | 0.105 | 77.8     | 76.7       |
| Team 2      | LSTM + GPT-2 + MAG/MMI Ens.                            | 82.5 | 69.8   | 1.8   | 0.082 | 79.1     | 78.1       |
| Team 3      | GPT-2 Ens. + Ens. beam search                          | 79.4 | 73.2   |       |       | 0.128     |            |
| Team 4      | BART-Large                                            | 81.3 | 73.9   | 3.5   | 0.108 | 78.0     | 77.7       |

The best results from each team are shown. (1) API prediction via accuracy, perplexity and attribute accuracy, and, (2a) Response generation via BLEU. (3) Dialog State Tracking (DST), via slot and intent prediction F1↑. ↑: higher is better, ↓: lower is better. Winner: Runner-up.

TABLE XIV
SUMMARY OF EACH TEAM’S RESULTS ON TEST-STD SPLIT, AVERAGE OF FURNITURE AND FASHION

| Teams       | Architecture | 2b. Rep. Retrieval | MRR↑ | r@1↑ |
|-------------|--------------|--------------------|------|------|
| Baseline    | LSTM → log-like. rank | 0.145 | 7.2   |
| Team 1      | GPT-2 + BLEUMETEOR rank | 0.326 | 21.1  |
| Team 2      | LSTM + GPT-2 + MAG/MMI Ens. | 0.074 | 2.5   |
| Team 3      | GPT-2 Ens. + cosine sim. | 0.381 | 26.3  |
| Team 4      | BART-Large + Bi-Encoder | 0.073 | 52.6  |
| Team 5*     | BERT → log-like. rank | 0.39  | 26.7  |

(2b) Response retrieval via mean reciprocal rank (MRR) and recall@k (k=1). (*Team 5 submitted results only for Fashion). Winner: Runner-up.

D. Results

The challenge saw a total of 11 model entries from 5 teams across the world, setting a new state-of-the-art in all three subtasks (Tables XIII & XIV).

For each subtask, we listed metrics in a priority order. Subtask winners were determined based on the entry with the most favorable performance on the highest priority metric, where its result was separated from other entries by more than one standard error, with ties broken by the next metric.

The majority of teams utilized generative models, e.g., GPT-2 [22] or BART [55]. Many were used in ensembles (Ens.) either with like models where training or model sizes were varied (Team 3), or with multimodal fusion models (Team 2).

The structural API call prediction subtask (subtask 1) was won by a BART-Large model from Team 4 that jointly predicted the dialog state (subtask 3), API call (subtask 1), and response (subtask 2a) as single target given the dialog history, multimodal context, and user utterance. This simpler architecture prevailed over a LSTM, GPT-2, multimodal fusion ensemble (MAG/MMI Ens. [56]) by demonstrating higher attribute accuracy, and was runner up in other subtasks.

The response retrieval subtask (subtask 2b) winner was a BART-based Bi-encoder [57], [58], [59], also from Team 4, whose weights were initialized from the jointly trained BART model that won subtask 1. The significant margin by which this model outperformed the others (0.29 points) is likely attributable to the use of discriminative training. The winner of the response generation and DST subtasks (subtask 2a and subtask 3) was an ensemble of GPT-2 models from Team 3 that were of differing sizes (large and small) and used differing portions of the training and development sets. An ensemble-beam-search over each model’s prediction was used to generate the final prediction.

None of the systems explicitly access the product catalogue information, relying instead on the models implicitly capturing it during training. This leads to hallucination in response generation. Future challenges should aim to discourage memorization of information that is mutable, such as pricing.

An ablation study where a BART-Large model was trained on versions of the datasets with and without the multimodal information showed that its absence significantly impacts subtask 3, slot-F1, and marginally impacts attribute accuracy, subtask 1 [60]. Response generation (subtask 2a) is largely unaffected. This implies that there is additional information in the multimodal input, however it is either not fully utilized or is somewhat marginal given the dialogue utterances.

VI. CONCLUSION

This paper summarizes the four tracks of the ninth Dialog System Technology Challenge (DSTC9). Beyond Domain APIs track expands the coverage of current task-oriented dialog systems by incorporating external unstructured knowledge sources. Multi-domain Task-oriented Dialog Challenge II, focuses on end-to-end multi-domain task completion dialog and cross-lingual multi-domain dialog state tracking. Interactive Evaluation of Dialog Track expands dialog research beyond
datasets, encourages to develop dialog systems that can converse effectively in interactive environments. The Situated Interactive Multi-Modal Conversational AI track focuses on real-world assistant agents that can handle multi-modal inputs and perform multimodal actions. All the datasets and resources introduced for every track are 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 Conf., 2013, pp. 404–413.

[2] M. Henderson, B. Thomson, and J. D. Williams, “The second dialog state tracking challenge,” in Proc. 15th Annu. Meeting Special Int. Group Discourse Dialogue (SIGDIAL), 2014, pp. 263–272.

[3] M. Henderson, B. Thomson, and J. D. Williams, “The third dialog state tracking challenge,” in Proc. IEEE Spoken Lang. Technol. Workshop (SLT), 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. Berlin, Germany: 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. Spoken Lang. Technol. Workshop (SLT), 2016, pp. 511–517.

[6] C. Horii et al., “Overview of the sixth dialog system technology challenge: DSTC6,” Comput. Speech Lang., vol. 55, pp. 1–25, 2018.

[7] K. Yoshino et al., “Overview of the seventh dialog system technology challenge: DSTC7,” Comput. Speech & Lang., vol. 62, 2020, Art. no. 101068.

[8] C. Gunasekara, J. K. Kummerfeld, L. Polymenakos, and W. S. Lasecki, “Dstc7 task 1: Noetic end-to-end response selection,” in Proc. 1st Workshop NLP Conversational AI, Florence, Italy, 2019, pp. 60–67.

[9] C. Gunasekara, J. K. Kummerfeld, L. Polymenakos, and W. Lasecki, “Dstc7 task 1: Noetic end-to-end response selection,” in Proc. First Workshop NLP Conversational AI, 2019, pp. 60–67.

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

[11] H. Alamri, C. Horii, T. K. Marks, D. Batr, and D. Parikh, “Audio visual scene-aware dialog (AVSD) track for natural language generation in dstc7,” in Proc. DSTC7 AAAZ2019 Workshop, 2018.

[12] S. Kim et al., “Overview of the eighth dialog system technology challenge,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 2529–2540, 2021.

[13] S. Kim, M. Eric, K. Gopalakrishnan, B. Hedayatnia, Y. Liu, and D. hakkani-Tür, “Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access,” in Proc. 21st Annu. Meeting Special Int. Group Discourse Dialogue, vol. 2020, pp. 278–289. [Online]. Available: https://www.aclweb.org/anthology/2020.sigdial-1.35

[14] M. Eric et al., “MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines,” in Proc. 12th Lang. Resour. Eval. Conf., Marseille, France, 2020, pp. 422–428.

[15] K. Gopalakrishnan, B. Hedayatnia, L. Wang, Y. Liu, and D. hakkani-Tür, “Are neural open-domain dialog systems robust to speech recognition errors in the dialog history? an empirical study,” in Proc. INTERSPEECH, 2020, pp. 911–915.

[16] S. Kim et al., “The Eighth Dialog System Technology Challenge,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, arXiv:1911.06481.

[17] J. Li et al., “Results of the multi-domain task-completion dialog challenge,” in Proc. 34th AAAI Conf. Artif. Intell. 8th Dialog System Technology Challenge Workshop, 2020.

[18] R. Takanobu, Q. Zhu, J. Li, B. Peng, J. Gao, and M. Huang, “Is your goal-oriented dialog model performing really well? empirical analysis of system-wise evaluation,” in Proc. 21st Annu. Meeting Special Int. Group Discourse Dialogue, 2020, pp. 297–310.

[19] D. Ham, J.-G. Lee, Y. Jang, and K.-E. Kim, “End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2,” in Proc. 58th Annu. Meeting Assoc. Computut. Linguistics, 2020, pp. 583–592.

[20] B. Peng, C. Li, J. Li, S. Shayyandeh, L. Liden, and J. Gao, “Soloist: Building task bots at scale with transfer learning and machine teaching,” Trans. Assoc. Comput. Linguistics, vol. 9, pp. 807–824, 2021.

[21] E. Hosseini-Asl, B. McCann, C.-S. Wu, S. Yavuz, and R. Socher, “A simple language model for task-oriented dialogue,” Advances Neural Inf. Process. Syst., vol. 33, pp. 20179–20191, 2020.

[22] R. Takanobu, Q. Zhu, J. Li, B. Peng, J. Gao, and M. Huang, “Is your goal-oriented dialog model performing really well? empirical analysis of system-wise evaluation,” in Proc. 21st Annu. Meeting Special Int. Group Discourse Dialogue, 2020, pp. 297–310.

[23] D. Ham, J.-G. Lee, Y. Jang, and K.-E. Kim, “End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2,” in Proc. 58th Annu. Meeting Assoc. Computut. Linguistics, 2020, pp. 583–592.

[24] B. Peng, C. Li, J. Li, S. Shayyandeh, L. Liden, and J. Gao, “Soloist: Building task bots at scale with transfer learning and machine teaching,” Trans. Assoc. Comput. Linguistics, vol. 9, pp. 807–824, 2021.
[44] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with BERT,” 2019, arXiv:1904.09675.

[45] S. Mehri and M. Eskenazi, “Unsupervised evaluation of interactive dialog with dialoGPT,” in Proc. 21th Annu. Meeting Special Interest Group Discourse Dialogue, 2020, pages 225–235.

[46] H. Ai, A. Raux, D. Bohus, M. Eskenazi, and D. Litman, “Comparing spoken dialog corpora collected with recruited subjects versus real users,” in Proc. 8th SIGdial Workshop Discourse Dialogue, 2007, pp. 124–131.

[47] Z. Li, Z. Li, J. Zhang, Y. Feng, and J. Zhou, “Wechat AI & ICT’s submission for DSTC9 interactive dialogue evaluation track,” 2021, arXiv:2101.07947.

[48] S. Bao et al., “A unified pre-training framework for conversational AI,” 2021, arXiv:2105.02482.

[49] Y. Zhang et al., “Dialogpt: Large-scale generative pre-training for conversational response generation,” in Proc. 58th Annu. Meeting Assoc. for Comput. Linguistics: Syst. Demonstrations, 2020, pp. 270–278.

[50] D. Adiwidana et al., “Towards a human-like open-domain chatbot,” 2020, arXiv:2001.09977.

[51] A. Ram et al., “Conversational AI: The science behind the alexa prize,” 2018, arXiv:1801.03604.

[52] S. Moon et al., “Situated and interactive multimodal conversations,” in Proc. 28th Int. Conf. Comput. Linguistics (COLING), Barcelona, Spain, 2020, pp. 1103–1121.

[53] P. A. Crook et al., “Simmc: Situated interactive multi-modal conversational data collection and evaluation platform,” 2019, arXiv:1911.02690.

[54] A. H. Miller et al., “Parlai: A dialog research software platform,” in Proc. 2017 Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations, Copenhagen, Denmark, 2017, pp. 79–84.

[55] M. Lewis et al., “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.703

[56] B. Kim, I. Lee, Y. Jeong, K. Youngjoong, M.-W. Koo, and J. Seo, “Improving multimodal API prediction via adding dialog state and various multimodal gates,” AAAI 2021 DSTC9 Workshop, 2021.

[57] S. Humeau, K. Shuster, M. Lachaux, and J. Weston, “Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring,” in 8th Int. Conf. Learn. Representations, Addis Ababa, Ethiopia, 2020, 2020. [Online]. Available: https://openreview.net/forum?id=Skxg0n4FyH

[58] P.-E. Mazaré, S. Humeau, M. Raison, and A. Bordes, “Training millions of personalized dialogue agents,” in 2018 Conf. Empirical Methods Natural Lang. Process. Brussels, Belgium, 2018, pp. 2775–2779. [Online]. Available: https://www.aclweb.org/anthology/D18-1298

[59] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, “Wizard of wikipedia: Knowledge-powered conversational agents,” in Proc. Int. Conf. Learn. Representations (ICLR), 2019.

[60] S. Kottur et al., “An analysis of state-of-the-art models for situated interactive MultiModal conversations (SIMMC),” in Proc. 22nd Annu. Meeting Special Int. Group Discourse Dialogue (SIGDial), Jul. 2021, pp. 144–153. [Online]. Available: https://aclanthology.org/2021.sigdial-1.15