# Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations Track at DSTC10

**Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis,**
**Behnam Hedayatnia, Karthik Gopalakrishnan, Dilek Hakkani-Tür**
Amazon Alexa AI, Sunnyvale, CA, USA
{seokhwk, yangliud, djinamzn, papangea, behnam, karthgop, hakkanit}@amazon.com

## Abstract

A lot of recent work in dialogue modeling has been on written conversations, partly because of available data sets. However, written dialogues are not sufficient to fully capture the nature of spoken conversations as well as the potential effect of speech recognition errors on practical spoken dialogue systems. This challenge track aims to provide a new benchmark on spoken task-oriented conversations. We introduce the validation and test data sets for multi-domain dialogue state tracking and knowledge-grounded dialogue modeling tasks. The challenge track received a total of 99 entries from 21 participating teams in total. From the evaluation results, we observe that the data augmentation and model ensemble methods are two major factors that help enhance models' generalization capabilities to unseen spoken data and achieve good performance on both tasks.

## Introduction

Recently, more public data sets and benchmarks have become available for dialogue research on task-oriented conversations in various domains (El Asri et al. 2017; Wen et al. 2017; Budzianowski et al. 2018; Rastogi et al. 2020). 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. Figure 1 compares a written and a spoken conversation, and shows many differences in terms of wording and expressions between the two examples even for the same content. The spoken example includes disfluencies and speech recognition errors (highlighted by underlined text). Moreover, there is no punctuation, capitalization, or sentence segmentation in raw speech recognizer outputs.[1]

| **Written Conversation** | |
|---|---|
| User | I need a hotel in Fisherman's Wharf |
| Agent | Is there a particular price range you are looking for? |
| User | I'm looking in the expensive price range |
| Agent | The Suite at Fisherman's Wharf may work for you |
| User | Do you know how much the parking is? |
| Agent | It would cost 25 dollars per day. |

| **Spoken Conversation** | |
|---|---|
| User | hi ummm i'm looking for a place at uhhh to stay at fisherman's wharf at a hotel in the expensive pressure engine |
| Agent | sure let me see ok so there is one called the suites at fisherman's wharf is that something that would be interesting to you |
| User | can you tell me how much parking coast |
| Agent | sure okay this hotel charges twenty five dollars per day |

Figure 1: Examples of written and spoken conversations

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 (Tur et al. 2002; Hakkani-Tür et al. 2006; Henderson et al. 2012; Tur, Deoras, and Hakkani-Tür 2013; Masumura et al. 2018; Ladhak et al. 2016; Velikovich 2016; Huang and Chen 2019, 2020; Kim et al. 2021b). 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.

To benchmark the robustness of conversational models on spoken conversations, this challenge track introduces a new data set with spoken task-oriented dialogues for the following two subtasks: 1) multi-domain dialogue state tracking (Budzianowski et al. 2018) and 2) knowledge-grounded dialogue modeling (Kim et al. 2020). The remainder of this paper presents the data and task details and reports the evaluation results of the submitted entries from the challenge track participants.

[1] There are automatic models to recover such information, but they are still error prone.

## Related Work

There have been a lot of studies towards improving the single-turn spoken language understanding (SLU) robustness to automatic speech recognition (ASR) errors. The most common line of work has focused on utilizing multiple ASR hypotheses in the form of word confusion networks (Tur et al. 2002; Hakkani-Tür et al. 2006; Henderson et al. 2012; Tur, Deoras, and Hakkani-Tür 2013; Masumura et al. 2018) or word lattices (Ladhak et al. 2016; Velikovich 2016; Huang and Chen 2019, 2020). Recently, more proactive approaches have been explored for ASR error correction (Weng et al. 2020; Namazifar et al. 2021) and data augmentation (Wang et al. 2020; Jindal et al. 2020; Sawhney et al. 2021) as well.

The earlier dialog state tracking challenges (DSTCs) aimed to address the multi-turn dialogue problems in spoken conversations. DSTC2 (Henderson, Thomson, and Williams 2014a) and DSTC3 (Henderson, Thomson, and Williams 2014b) data sets included $n$-best ASR hypotheses as well as word confusions, which was intended for speech-oriented studies. However, the dialogue research community hasn't paid much attention to this aspect, due to the lack of critical ASR issues on these single-domain human-machine conversations that were restricted only to a small domain ontology. DSTC4 (Kim et al. 2017) and DSTC5 (Kim et al. 2016) extended from single-domain human-machine dialogs to multi-domain human-human conversations, but only manual transcriptions were included in the challenge data sets without ASR outputs.

On the other hand, many recent task-oriented dialogue datasets sets (El Asri et al. 2017; Wen et al. 2017; Budzianowski et al. 2018; Rastogi et al. 2020) consist of written conversations collected by crowdsourcing with no consideration of speech-specific aspects in spoken dialogue systems. There have been studies focusing on the speech robustness issues on task-oriented (Peng et al. 2020) and open-domain conversations (Gopalakrishnan et al. 2020), but they were restricted to simulated ASR errors based on written conversations.

## 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 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 1 summarizes the DSTC10 data details in comparison with two other data sets. MultiWOZ 2.0 (Budzianowski et al. 2018) and its variants (Eric et al. 2019; Zang et al. 2020) include crowd-sourced written conversations about seven different domains including hotel, restaurant, attraction, train, taxi, hospital, and police station in Cambridge, UK. Following the MultiWOZ data collection set-up, we recently released new written dialogues as a part of the official test set for the DSTC9 Track 1 (Kim et al. 2021a). This data was collected for a new locale, San Francisco, for three target domains: hotel, restaurant, and attraction, but with almost three times more entities than the MultiWOZ ontology entries. In addition, the data includes the turns grounded on the knowledge snippets from the FAQ list for the entities. This infrastructure and domain ontology for this written data were re-used for our spoken data collection. The difference is that the DSTC10 data came from the recorded conversations instead of the written text from crowd-sourcing.

To benchmark the robustness of models in practical spoken dialogues systems, 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 (Baevski et al. 2020) that was pre-trained on 960 hours of Librispeech (Panayotov et al. 2015). We fine-tuned it with 10% of our validation data. Then, we run 10-best decoding with an external language model built with KenLM (Heafield 2011) on all the written texts from MultiWOZ and DSTC9 data sets. For the user utterances on the test set, this ASR pipeline achieved a WER of 26.25% at 1-best and 24.31% oracle WER at 10-best hypotheses.

## Tasks

This section describes two benchmark tasks in this challenge track. As shown in Figure 2, we decouple between turns that could be handled by conventional task-oriented conversational models with no extra knowledge and turns that require external knowledge resources, following the architecture in  (Kim et al. 2020). In the first API/DB-based pipeline, we focus only on dialogue state tracking as the first target task. For the other knowledge access branch, our task 2 includes three subtasks introduced in  (Kim et al. 2020): 1) Knowledge-seeking Turn Detection, 2) Knowledge Selection, and 3) Knowledge-grounded Response Generation.

### Task 1: Multi-domain Dialogue State Tracking

Dialogue state tracking (DST) aims to estimate the system's belief states after each interaction with the user, which is a key problem in task-oriented conversational modeling. The belief states are defined to represent the latest user goals in a dialogue context from the beginning to the target user turn of a given conversation.

In this challenge track, we address the multi-domain DST task on human-human conversations, which have been actively explored by the dialogue research community especially using MultiWOZ data and its variants (Budzianowski et al. 2018; Eric et al. 2019; Zang et al. 2020). Following previous work, we also represent the user goals as a set of slot-value pairs defined for each domain and take the slot-level value prediction performance and the joint goal accuracy (Henderson, Thomson, and Williams 2014a) as the evaluation metrics. Figure 3 presents an example conversation with the ground-truth DST annotations for the first two user turns.

Our task differs from most previous DST benchmarks in the following two aspects. First, we focus on the DST performance on spoken conversations rather than written ones.

Table 1: Comparisons of the data sets.

| Data | Split | Locale | Modality | Dialogues |  | Domain Ontology |  |  |
|---|---|---|---|---|---|---|---|---|
|  |  |  |  | # sessions | # turns | # domains | # entities | # snippets |
| MultiWOZ 2.0 | all | Cambridge | written | 8,438 | 113,556 | 7 | 289 | - |
| DSTC9: Track 1 | test | San Francisco | written | 903 | 8,501 | 3 | 855 | 15,086 |
| DSTC10: Track 2 | val | San Francisco | spoken | 107 | 2,292 | 3 | 855 | 15,086 |
| DSTC10: Track 2 | test | San Francisco | spoken | 783 | 19,686 | 3 | 855 | 15,086 |

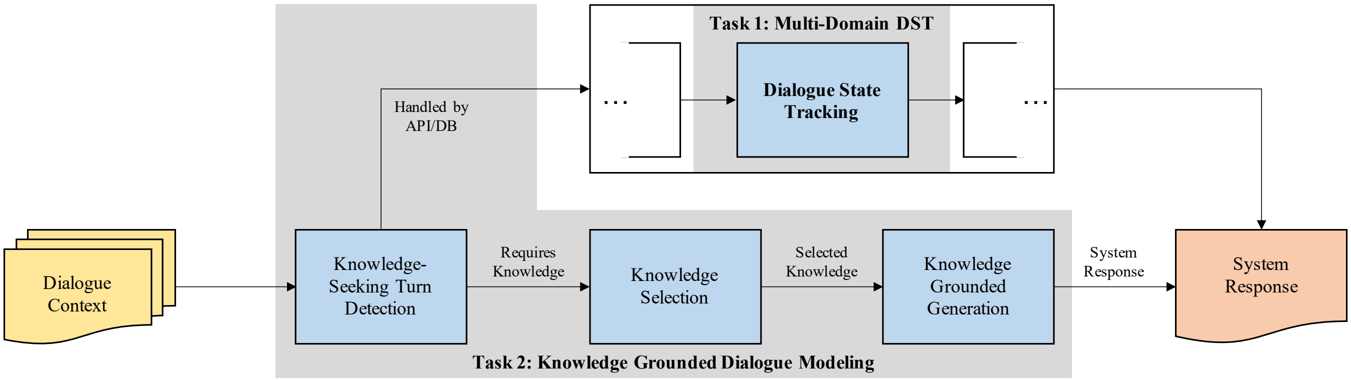

Figure 2: An overview of the benchmark tasks: multi-domain dialogue state tracking and knowledge-grounded dialogue modeling.

The latter has been widely used by previous DST studies because of its cost-efficiency in large scale data collection. We believe that those written data sets are not enough to fully reflect the actual human behaviors for spoken conversations. So we propose to take DST models trained on existing written data, evaluate on our new collection of spoken conversations, and eventually try to improve the model performance in face of a mismatch between training and test data sets.

In addition, our new data uses the ASR output instead of manual transcripts for the user turns, as described earlier. The goal is to evaluate how robust each DST model is against ASR errors. Although ASR errors are expected to be a very critical issue in developing spoken dialogue systems in practice, it hasn't been studied for the DST task.

### Task 2: Knowledge-grounded Dialogue Modeling

Recently, we introduced a new benchmark on task-oriented conversational modeling with unstructured knowledge access (Kim et al. 2020), which aims to incorporate external unstructured knowledge into the end-to-end dialogue response generation problem (Jin, Kim, and Hakkani-Tur 2021). As shown in Figure 2, this task focuses on the knowledge access branch, including the following three sub-tasks:

- **Knowledge-seeking Turn Detection** decides whether to trigger the knowledge access branch for a given utterance and dialogue history.
- **Knowledge Selection** selects appropriate knowledge snippets from the domain knowledge base for the knowledge seeking turn.
- **Knowledge-grounded Response Generation** generates a system response given a triple of input utterance, dialog context, and the selected knowledge snippet.

Figure 3 shows an example knowledge-seeking turn along with its ground-truth knowledge snippet and a reference response (in the last turn). We organized a challenge track for this task under DSTC9 (Kim et al. 2021a; Gunasekara et al. 2020), which had more than 100 submissions from 24 teams in total.

In DSTC10, we extend our DSTC9 track by replacing written conversations with spoken ones, as in the first task. The issue between written and spoken conversations has been partially discussed in our DSTC9 track that included a spoken subset in the test data (Kim et al. 2021a). But only manual transcripts were provided in the DSTC9 track, while the new data for this challenge includes the ASR outputs for the user turns.

## Baseline Models

To investigate the existing model behaviors on our spoken data, we adopt state-of-the-art models on written data for both tasks as baselines.

### Task 1 Baseline

We use TripPy [2] (Heck et al. 2020) as a baseline for the DST task. It is based on a fine-tuned BERT (Devlin et al. 2019) on the DST objectives with copy mechanisms from three different sources: user utterance, system utterance, and previous dialogue states. This model achieved 55.30% in joint goal accuracy on MultiWOZ 2.1, which was also used as the DST baseline for DialoGLUE (Mehri, Eric, and Hakkani-Tur 2020).

[2] https://gitlab.cs.uni-duesseldorf.de/general/dsml/trippy-public

| Speaker | Utterance | Task | Ground-truth Annotations |
|---|---|---|---|
| User | hi ummm i'm looking for a place at uhhh to stay at fisherman's wharf at a hotel in the moderate pressure engine | Task 1 | hotel-area: fisherman's wharf<br>hotel-type: hotel<br>hotel-pricerange: moderate |
| Agent | sure let me see what can i find for you ok unfortunately we're not showing any result is there any specification that i can change to possibly find you something | | |
| User | is there wholesale in the expensive pressure engine student | Task 1 | hotel-area: fisherman's wharf<br>hotel-type: hotel<br>hotel-pricerange: expensive |
| Agent | sure let me see ok so there is one called the suites at fisherman's wharf is that something that would be interesting to you | | |
| User | can you tell me how much parking coast | Task 2 | Q: What is the parking cost at the Suites at Fisherman's Wharf? - A: Parking costs $25 per day. |
| Agent | sure okay this hotel charges twenty five dollars per day | | |

Figure 3: An example conversation with the ground-truth labels for both tasks.

### Task 2 Baselines

For task 2, we take two baselines: one is the official baseline of DSTC9[3] (Kim et al. 2021a) and the other is Knover[4] (He et al. 2021) from the DSTC9 track winner. The official DSTC9 baseline uses the fine-tuned GPT-2 (Radford et al. 2019) for each sub-task. For Knover, we use the model for their submission #0. It differs from the GPT-2 baseline in the following: 1) replacing GPT-2 with PLATO-2 (Bao et al. 2021); 2) multi-scale negative sampling for knowledge selection; and 3) response generation with beam search instead of nucleus sampling. Note that the DSTC9 winning entry was a further enhanced version from this model by schema guided turn detection and model ensemble, however, that model is not publicly available.

## 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 following slot-level scores: slot-level accuracy and value/none prediction in precision, recall, and F1. If there exist multiple values for a single slot in either or both predictions and references, we treat the value F1 as a partial matching score and use it as the numerator of the slot-level metrics.

For task 2, we use the same evaluation criteria and metrics as in the DSTC9 Track 1 (Kim et al. 2021a). First, for each submission we calculated the task-specific objective metrics (Table 2) 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, as follows:

$$S_{overall}(e) = \frac{1}{|M|} \sum_{i=1}^{|M|} \frac{1}{rank_i(e)} \quad (1)$$

where $rank_i(e)$ 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.

Table 2: Task 2 objective evaluation metrics

| Sub-task | Metrics |
|---|---|
| Detection | Precision/Recall/F-measure |
| Selection | MRR@5, Recall@1, Recall@5 |
| Generation | BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-1, ROUGE-2, ROUGE-L |

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.

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

## 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.

### Task 1 Results

Table 3 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 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 bene-

[3] https://github.com/alexa/alexa-with-dstc9-track1-dataset

[4] https://github.com/PaddlePaddle/Knover

Table 3: Task 1 evaluation results of the best entries from each team. Bold denotes the best score for each metric; underline indicates the best results among the single models with no ensemble

| | Proposed Methods | | | | Joint Goal | Slot | Value Prediction | | | None Prediction | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Team-Entry | Classification | Extraction | Generation | Ensemble | Accuracy | Accuracy | P | R | F | P | R | F |
| A11 - 1 | | | ✓ | ✓ | **0.4616** | **0.9498** | 0.9224 | **0.9009** | **0.9115** | **0.9707** | 0.9858 | **0.9782** |
| A11 - 4 | | | ✓ | | 0.4071 | 0.9352 | 0.9006 | 0.8705 | 0.8853 | 0.9612 | 0.9826 | 0.9718 |
| A01 - 0 | | | ✓ | | 0.3605 | 0.9367 | 0.9208 | 0.8671 | 0.8931 | 0.9491 | 0.9861 | 0.9672 |
| A01 - 2 | | | ✓ | | 0.3553 | 0.9362 | **0.9236** | 0.8649 | 0.8933 | 0.9468 | **0.9870** | 0.9665 |
| A07 - 1 | | | ✓ | ✓ | 0.2773 | 0.8948 | 0.8154 | 0.7756 | 0.7950 | 0.9464 | 0.9773 | 0.9616 |
| A10 - 4 | ✓ | | | ✓ | 0.2679 | 0.9079 | 0.8486 | 0.8254 | 0.8368 | 0.9486 | 0.9659 | 0.9571 |
| A09 - 4 | | ✓ | | ✓ | 0.1821 | 0.8759 | 0.8767 | 0.7332 | 0.7986 | 0.8778 | 0.9738 | 0.9233 |
| A06 - 3 | ✓ | | | ✓ | 0.1691 | 0.8595 | 0.7375 | 0.7689 | 0.7529 | 0.9498 | 0.9227 | 0.9361 |
| A05 - 3 | ✓ | ✓ | ✓ | ✓ | 0.1615 | 0.8624 | 0.8275 | 0.7121 | 0.7655 | 0.8837 | 0.9661 | 0.9231 |
| A03 - 1 | | ✓ | | ✓ | 0.0524 | 0.7803 | 0.6725 | 0.5251 | 0.5897 | 0.8314 | 0.9532 | 0.8881 |
| A04 - 0 | | ✓ | | | 0.0050 | 0.6996 | 0.6577 | 0.2843 | 0.3970 | 0.7097 | 0.9791 | 0.8229 |
| A08 - 0 | | ✓ | | | 0.0018 | 0.7045 | 0.6362 | 0.2962 | 0.4042 | 0.7217 | 0.9795 | 0.8311 |
| A02 - 0 | | ✓ | | | 0.0014 | 0.6946 | 0.5683 | 0.2692 | 0.3653 | 0.7257 | 0.9810 | 0.8342 |
| Baseline | ✓ | ✓ | | | 0.0039 | 0.7052 | 0.6130 | 0.3097 | 0.4115 | 0.7302 | 0.9716 | 0.8338 |

Table 4: Data augmentation methods by the top-4 teams

| | | Value | Dialogue | Speech |
|---|---|---|---|---|
| Rank | Team | Substitution | Generation | Simulation |
| 1 | A11 | ✓ | ✓ | ✓ |
| 2 | A01 | ✓ | | |
| 3 | A07 | ✓ | | |
| 4 | A10 | | ✓ | |

fit 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. Table 4 summarizes the data augmentation methods introduced by the top 4 teams, which are categorized into: value substitution from the training utterances with the new ontology contents, multi-turn or whole dialogue-level synthetic data generation, and speech/ASR simulation. While the other teams focused on a single approach only, Team A11 used all three data augmentation methods to generate a huge amount of training data and achieved the best performance.

In addition, the model ensemble also helped to boost the performance from the single-model results. Figure 4 shows the performance gains by the model ensemble from all three teams 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.

## Task 2 Results

Table 5 shows the objective evaluation results of the best entry from each team selected based on the overall score (Equation 1). Most entries show the improved performance from the baseline models in all three tasks. Figure 5 shows that 14 out of 16 teams achieved higher knowledge-seeking turn detection performance than the baseline (Jin et al. 2021). For knowledge selection, the majority of the teams submitted the improved results over the baselines (Figure 6). 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.

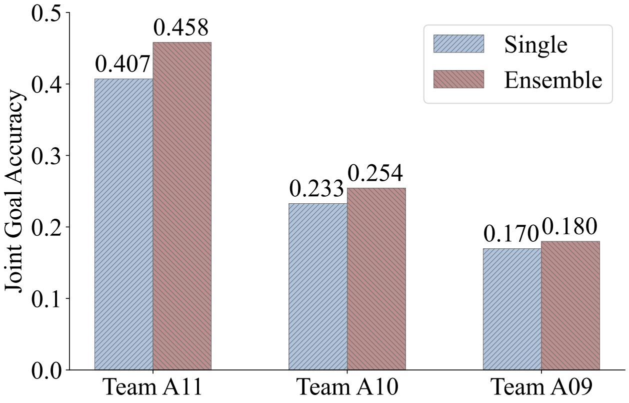

Figure 4: Comparisons between single-model and ensemble-based DST performance

Participating teams also significantly improved the response generation task. As shown in Figure 7, 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 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 (Equation 1). Table 6 shows the official ranking of the finalists based on the human evaluation

Table 5: Task 2 objective evaluation results of the best entries from each team. 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.

| | Task #1: Detection | | | Task #2: Selection | | | Task #3: Generation | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Team-Entry | P | R | F | MRR@5 | R@1 | R@5 | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-1 | ROUGE-2 | ROUGE-L |
| DSTC9 | 0.9017 | 0.7116 | 0.7954 | 0.5230 | 0.4583 | 0.6252 | 0.1153 | 0.0516 | 0.0186 | 0.0075 | 0.1215 | 0.1443 | 0.0412 | 0.1143 |
| Knover | 0.8967 | 0.6735 | 0.7692 | 0.5574 | 0.4950 | 0.6472 | 0.1248 | 0.0615 | 0.0290 | 0.0153 | 0.1364 | 0.1519 | 0.0519 | 0.1229 |
| B01 - 0 | 0.8589 | 0.9180 | 0.8875 | 0.2038 | 0.1840 | 0.2293 | 0.3017 | 0.2072 | 0.1213 | 0.0681 | 0.3514 | 0.3527 | 0.1730 | 0.3389 |
| B02 - 3* | 0.8689 | 0.9414 | 0.9037 | 0.7233 | 0.6929 | 0.7660 | 0.3732 | 0.2775 | 0.1957 | 0.1317 | 0.4387 | 0.4196 | 0.2392 | 0.4115 |
| B03 - 2 | 0.1862 | 0.0952 | 0.1260 | 0.0000 | 0.0000 | 0.0000 | 0.0063 | 0.0008 | 0.0000 | 0.0000 | 0.0076 | 0.0099 | 0.0010 | 0.0090 |
| B04 - 1* | 0.8814 | 0.9575 | 0.9179 | 0.7891 | 0.7481 | 0.8407 | 0.3380 | 0.2483 | 0.1733 | 0.1118 | 0.4070 | 0.3930 | 0.2201 | 0.3868 |
| B05 - 0 | 0.8900 | 0.8770 | 0.8835 | 0.5569 | 0.4499 | 0.7094 | 0.1553 | 0.0826 | 0.0430 | 0.0213 | 0.1689 | 0.1952 | 0.0735 | 0.1616 |
| B06 - 1 | 0.8439 | 0.7438 | 0.7907 | 0.5575 | 0.4887 | 0.6506 | 0.1184 | 0.0533 | 0.0218 | 0.0094 | 0.1285 | 0.1488 | 0.0440 | 0.1177 |
| B07 - 3* | 0.9020 | 0.8624 | 0.8817 | 0.7223 | 0.6527 | 0.8204 | 0.3335 | 0.2460 | 0.1722 | 0.1101 | 0.4172 | 0.3814 | 0.2143 | 0.3915 |
| B08 - 0 | 0.8413 | 0.9780 | 0.9045 | 0.7430 | 0.7028 | 0.8030 | 0.3961 | 0.3002 | 0.2172 | 0.1453 | 0.4526 | 0.4433 | 0.2613 | 0.4372 |
| B08 - 3* | 0.8503 | **0.9810** | 0.9109 | 0.7479 | 0.7097 | 0.8049 | **0.4013** | **0.3047** | **0.2231** | **0.1525** | **0.4597** | **0.4487** | **0.2657** | **0.4405** |
| B09 - 1 | 0.8248 | 0.8682 | 0.8459 | 0.3258 | 0.3067 | 0.3509 | 0.1269 | 0.0536 | 0.0227 | 0.0129 | 0.1365 | 0.1577 | 0.0444 | 0.1241 |
| B10 - 0 | 0.8818 | 0.9502 | 0.9147 | 0.7926 | 0.7428 | 0.8654 | 0.1427 | 0.0875 | 0.0405 | 0.0190 | 0.1919 | 0.2106 | 0.0906 | 0.2009 |
| B10 - 1* | 0.8793 | 0.9707 | 0.9228 | **0.8332** | **0.7933** | **0.8866** | 0.1617 | 0.0990 | 0.0486 | 0.0237 | 0.2096 | 0.2295 | 0.0981 | 0.2187 |
| B11 - 0 | 0.8774 | 0.7335 | 0.7990 | 0.4606 | 0.4131 | 0.5311 | 0.1106 | 0.0499 | 0.0229 | 0.0105 | 0.1257 | 0.1425 | 0.0436 | 0.1154 |
| B12 - 1* | 0.8697 | 0.9385 | 0.9028 | 0.7327 | 0.6915 | 0.7901 | 0.3327 | 0.2351 | 0.1568 | 0.0984 | 0.4039 | 0.3849 | 0.2003 | 0.3742 |
| B13 - 1 | 0.8876 | 0.8902 | 0.8889 | 0.5716 | 0.5716 | 0.5716 | 0.1452 | 0.0754 | 0.0321 | 0.0150 | 0.1580 | 0.1783 | 0.0630 | 0.1420 |
| B14 - 0* | **0.9214** | 0.9268 | **0.9241** | 0.6464 | 0.6204 | 0.6847 | 0.2705 | 0.1818 | 0.0936 | 0.0298 | 0.3167 | 0.3324 | 0.1543 | 0.3184 |
| B15 - 1 | 0.8688 | 0.8433 | 0.8559 | 0.3737 | 0.3135 | 0.4710 | 0.0496 | 0.0244 | 0.0133 | 0.0074 | 0.0835 | 0.1069 | 0.0389 | 0.1036 |
| B16 - 3* | 0.8852 | 0.8697 | 0.8774 | 0.7493 | 0.7105 | 0.7976 | 0.3083 | 0.2097 | 0.1268 | 0.0701 | 0.3547 | 0.3641 | 0.1781 | 0.3523 |

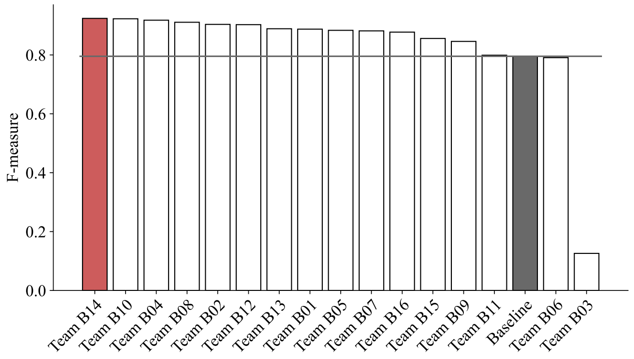

Figure 5: knowledge-seeking turn detection performance (F-measure) from different entries. The horizontal line indicates the baseline performance.

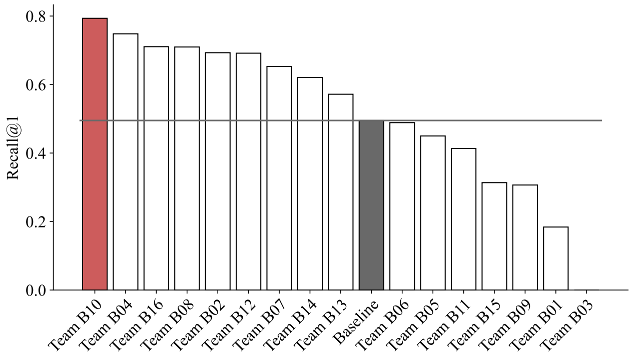

Figure 6: knowledge selection performance (Recall@1) from different entries. The horizontal line indicates the baseline performance.

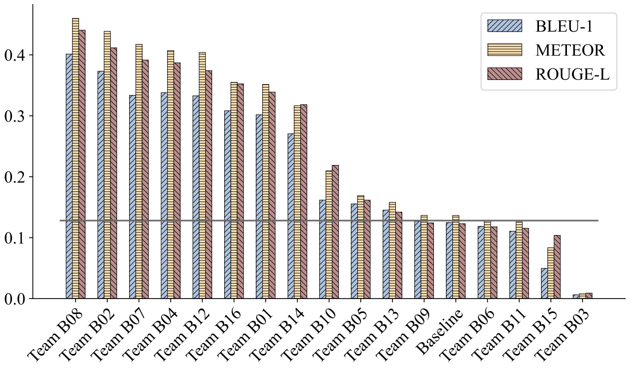

Figure 7: Knowledge-grounded generation performance from different entries. The horizontal line indicates the average of the baseline scores.

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 scores from the objective evaluation.

Consistently with our DSTC9 track results (Kim et al. 2021a), the best team on the knowledge selection task again ended up with the final winner after the human evaluation. To better understand the importance of each task and its metrics towards end-to-end performance, we calculated the Spearman's rank correlation coefficient (Spearman 1961) of the ranked lists of all the entries in every pair of objective and human evaluation metrics, as shown in Figure 8. While the detection and selection scores were relatively well-aligned with the end-to-end results, we observed the negative correlations between the automatic generation met-

Table 6: Human evaluation results. Bold indicates the best score for each metric 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 |

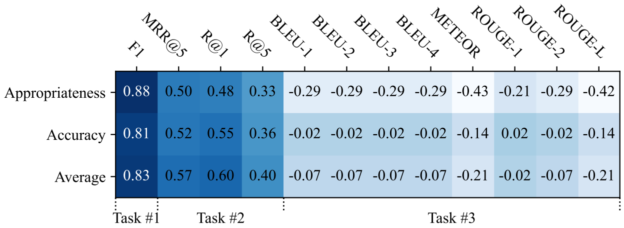

Figure 8: Correlations between the objective and human evaluation metrics in Spearman's $\rho$. The higher score of a pair of metrics has, the stronger correlation they have.

rics and the final human evaluation scores. It indicates that the more similar responses to the reference speech transcripts in terms of lexical overlaps are not always more preferred by human evaluators. We suppose this suggests a new research direction towards more reliable metrics in both objective and human evaluations for spoken conversations.

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, as summarized in Table 7. In addition, all the top-4 teams for task 2 utilized the augmented data to train their models, similar to the best entries for Task 1. Finally, model ensembles further improved performance, as shown in Figure 9.

Table 7: Proposed methods by the top-4 teams for Task 2

| | | Entity | Data | Model |
|---|---|---|---|---|
| Rank | Team | Tracking | Augmentation | Ensemble |
| 1 | B10 | | ✓ | ✓ |
| 2 | B04 | ✓ | ✓ | |
| 3 | B08 | ✓ | ✓ | ✓ |
| 4 | B14 | ✓ | ✓ | |

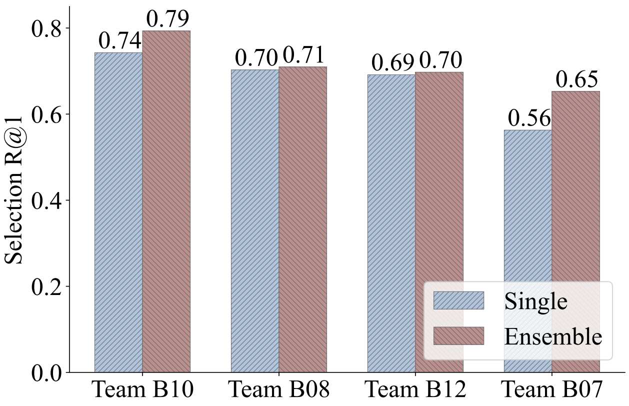

Figure 9: Comparisons between single-model and ensemble-based knowledge selection performance

## 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.

## References

Baevski, A.; Zhou, H.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv:2006.11477.

Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H.; Wu, W.; Guo, Z.; Liu, Z.; and Xu, X. 2021. PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning. arXiv:2006.16779.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gasic, M. 2018. MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 5016–5026.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

El Asri, L.; Schulz, H.; Sarma, S. K.; Zumer, J.; Harris, J.; Fine, E.; Mehrotra, R.; and Suleman, K. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 207–219.

Eric, M.; Goel, R.; Paul, S.; Sethi, A.; Agarwal, S.; Gao, S.; and Hakkani-Tur, D. 2019. MultiWOZ 2.1: Multi-Domain

Dialogue State Corrections and State Tracking Baselines. *arXiv preprint arXiv:1907.01669*.

Gopalakrishnan, K.; Hedayatnia, B.; Wang, L.; Liu, Y.; and Hakkani-Tur, D. 2020. Are neural open-domain dialog systems robust to speech recognition errors in the dialog history? an empirical study. *arXiv preprint arXiv:2008.07683*.

Gunasekara, C.; Kim, S.; D'Haro, L. F.; Rastogi, A.; Chen, Y.-N.; Eric, M.; Hedayatnia, B.; Gopalakrishnan, K.; Liu, Y.; Huang, C.-W.; Hakkani-Tür, D.; Li, J.; Zhu, Q.; Luo, L.; Liden, L.; Huang, K.; Shayandeh, S.; Liang, R.; Peng, B.; Zhang, Z.; Shukla, S.; Huang, M.; Gao, J.; Mehri, S.; Feng, Y.; Gordon, C.; Alavi, S. H.; Traum, D.; Eskenazi, M.; Beirami, A.; Eunjoon; Cho; Crook, P. A.; De, A.; Geramifard, A.; Kottur, S.; Moon, S.; Poddar, S.; and Subba, R. 2020. Overview of the Ninth Dialog System Technology Challenge: DSTC9. arXiv:2011.06486.

Hakkani-Tür, D.; Béchet, F.; Riccardi, G.; and Tur, G. 2006. Beyond ASR 1-best: Using word confusion networks in spoken language understanding. *Computer Speech & Language*, 20(4): 495–514.

He, H.; Lu, H.; Bao, S.; Wang, F.; Wu, H.; Niu, Z.; and Wang, H. 2021. Learning to Select External Knowledge with Multi-Scale Negative Sampling. arXiv:2102.02096.

Heafield, K. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, 187–197.

Heck, M.; van Niekerk, C.; Lubis, N.; Geishauser, C.; Lin, H.-C.; Moresi, M.; and Gasic, M. 2020. TripPy: A Triple Copy Strategy for Value Independent Neural Dialog State Tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 35–44.

Henderson, M.; Gašić, M.; Thomson, B.; Tsiakoulis, P.; Yu, K.; and Young, S. 2012. Discriminative spoken language understanding using word confusion networks. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, 176–181.

Henderson, M.; Thomson, B.; and Williams, J. D. 2014a. The Second Dialog State Tracking Challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 263–272. Philadelphia, PA, U.S.A.: Association for Computational Linguistics.

Henderson, M.; Thomson, B.; and Williams, J. D. 2014b. The third Dialog State Tracking Challenge. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, 324–329.

Huang, C.-W.; and Chen, Y.-N. 2019. Adapting pretrained transformer to lattices for spoken language understanding. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 845–852. IEEE.

Huang, C.-W.; and Chen, Y.-N. 2020. Learning Spoken Language Representations with Neural Lattice Language Modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3764–3769.

Jin, D.; Gao, S.; Kim, S.; Liu, Y.; and Hakkani-Tur, D. 2021. Towards Zero and Few-shot Knowledge-seeking Turn Detection in Task-orientated Dialogue Systems. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, 281–288. Online: Association for Computational Linguistics.

Jin, D.; Kim, S.; and Hakkani-Tur, D. 2021. Can I Be of Further Assistance? Using Unstructured Knowledge Access to Improve Task-oriented Conversational Modeling. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, 119–127. Online: Association for Computational Linguistics.

Jindal, A.; Ghosh Chowdhury, A.; Didolkar, A.; Jin, D.; Sawhney, R.; and Shah, R. R. 2020. Augmenting NLP models using Latent Feature Interpolations. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6931–6936. Barcelona, Spain (Online): International Committee on Computational Linguistics.

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

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

Kim, S.; Eric, M.; Gopalakrishnan, K.; Hedayatnia, B.; Liu, Y.; and Hakkani-Tur, D. 2020. Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 278–289.

Kim, S.; Eric, M.; Hedayatnia, B.; Gopalakrishnan, K.; Liu, Y.; Huang, C.-W.; and Hakkani-Tur, D. 2021a. Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access Track in DSTC9. arXiv:2101.09276.

Kim, S.; Liu, Y.; Jin, D.; Papangelis, A.; Gopalakrishnan, K.; Hedayatnia, B.; and Hakkani-Tur, D. 2021b. " How Robust ru?": Evaluating Task-Oriented Dialogue Systems on Spoken Conversations. *arXiv preprint arXiv:2109.13489*.

Ladhak, F.; Gandhe, A.; Dreyer, M.; Mathias, L.; Rastrow, A.; and Hoffmeister, B. 2016. LatticeRnn: Recurrent Neural Networks Over Lattices. In *Interspeech 2016*, 695–699.

Masumura, R.; Ijima, Y.; Asami, T.; Masataki, H.; and Higashinaka, R. 2018. Neural confnet classification: Fully neural network based spoken utterance classification using word confusion networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6039–6043. IEEE.

Mehri, S.; Eric, M.; and Hakkani-Tur, D. 2020. DialoGLUE: A Natural Language Understanding Benchmark for Task-Oriented Dialogue. arXiv:2009.13570.

Namazifar, M.; Malik, J.; Li, L. E.; Tur, G.; and Tür, D. H. 2021. Correcting Automated and Manual Speech Transcription Errors using Warped Language Models. *arXiv preprint arXiv:2103.14580*.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210. IEEE.

Peng, B.; Li, C.; Zhang, Z.; Zhu, C.; Li, J.; and Gao, J. 2020. RADDLE: An Evaluation Benchmark and Analysis Platform for Robust Task-oriented Dialog Systems. *arXiv preprint arXiv:2012.14666*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Rastogi, A.; Zang, X.; Sunkara, S.; Gupta, R.; and Khaitan, P. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8689–8696.

Sawhney, R.; Thakkar, M.; Agarwal, S.; Jin, D.; Yang, D.; and Flek, L. 2021. HypMix: Hyperbolic Interpolative Data Augmentation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9858–9868. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Spearman, C. 1961. The proof and measurement of association between two things.

Tur, G.; Deoras, A.; and Hakkani-Tür, D. 2013. Semantic parsing using word confusion networks with conditional random fields. In *Interspeech 2013*, 2579–2583. Citeseer.

Tur, G.; Wright, J.; Gorin, A.; Riccardi, G.; and Hakkani-Tür, D. 2002. Improving spoken language understanding using word confusion networks. In *Seventh International Conference on Spoken Language Processing*.

Velikovich, L. 2016. Semantic model for fast tagging of word lattices. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, 398–405. IEEE.

Wang, L.; Fazel-Zarandi, M.; Tiwari, A.; Matsoukas, S.; and Polymenakos, L. 2020. Data Augmentation for Training Dialog Models Robust to Speech Recognition Errors. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, 63–70.

Wen, T.-H.; Vandyke, D.; Mrkšić, N.; Gasic, M.; Barahona, L. M. R.; Su, P.-H.; Ultes, S.; and Young, S. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 438–449.

Weng, Y.; Miryala, S. S.; Khatri, C.; Wang, R.; Zheng, H.; Molino, P.; Namazifar, M.; Papangelis, A.; Williams, H.; Bell, F.; and Tur, G. 2020. Joint Contextual Modeling for ASR Correction and Language Understanding. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6349–6353.

Zang, X.; Rastogi, A.; Sunkara, S.; Gupta, R.; Zhang, J.; and Chen, J. 2020. MultiWOZ 2.2: A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020*, 109–117.