Task-Oriented Conversational Modeling with Subjective Knowledge Track in DSTC11

Seokhwan Kim[†], **Spandana Gella**[†], **Chao Zhao**[‡], **Di Jin**[†],
Alexandros Papangelis[†], **Behnam Hedayatnia**[†], **Yang Liu**[†], **Dilek Hakkani-Tur**[†]
[†] Amazon Alexa AI [‡]University of North Carolina at Chapel Hill
{seokhwk,sgella,djinamzn,papangea,behnam,yangliud,hakkanit}@amazon.com
zhaochao@cs.unc.edu

Abstract

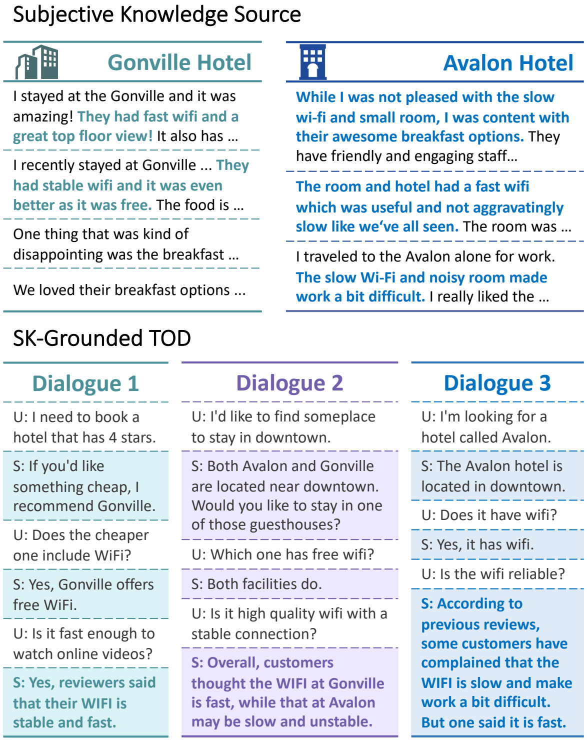
Conventional Task-oriented Dialogue (TOD) Systems rely on domain-specific APIs/DBs or external factual knowledge to create responses. In DSTC11 track 5, we aims to provide a new challenging task to accommodate subjective user requests (e.g.,"*Is the WIFI reliable?*" or "*Does the restaurant have a good atmosphere?*") into TOD. We release a benchmark dataset, which contains subjective knowledge-seeking dialogue contexts and manually annotated responses that are grounded in subjective knowledge sources. The challenge track received a total of 48 entries from 14 participating teams.

1 Introduction

Task-oriented Dialogue (TOD) Systems aim to build dialogue systems to help users to achieve specific goals (e.g., booking a hotel or a restaurant). Most solutions of TOD are either based on domain-APIs (Budzianowski et al., 2018; Rastogi et al., 2020) and structured databases (Eric et al., 2017; Wu et al., 2019), which can only handle limited scenarios. Recent works incorporate unstructured factual information (Dimitrakis et al., 2018; Kim et al., 2020, 2021b; Feng et al., 2020, 2021; Majumder et al., 2022) into dialogue modeling, which further enlarges the model's ability of task-oriented assistance.

However, in many TOD tasks, users care about not only factual information but subjective information as well, such as the experiences, opinions, and preferences of other customers. For example, when booking a hotel or a restaurant, users may ask questions like *"Is the WIFI reliable?"* or *"Does the restaurant have a good atmosphere?"*. To respond to such user requests, an agent needs to seek information from subjective knowledge sources such as online customer reviews.

In this track, we focus on enabling the TOD model to leverage subjective knowledge during task-oriented assistance. To this end, we proposed a novel task of subjective-knowledge-based task-oriented dialogue (SK-TOD) (Zhao et al., 2023). More specifically, we focus on subjective knowledge-seeking user requests and choose user reviews as external subjective knowledge sources. In Figure 1, we show three examples of such requests, where customers ask about the WiFi quality of hotels. User reviews are great resources for subjective information because even for the same aspect of the same product/service, customers may



Gonville Hotel:

I stayed at the Gonville and it was amazing! They had fast wifi and a great top floor view! It also has ...

I recently stayed at Gonville ... They had stable wifi and it was even better as it was free. The food is ...

One thing that was kind of disappointing was the breakfast ...

We loved their breakfast options ...

Avalon Hotel:

While I was not pleased with the slow wi-fi and small room, I was content with their awesome breakfast options. They have friendly and engaging staff...

The room and hotel had a fast wifi which was useful and not aggravatingly slow like we've all seen. The room was ...

I traveled to the Avalon alone for work. The slow Wi-Fi and noisy room made work a bit difficult. I really liked the ...

Dialogue 1:

U: I need to book a hotel that has 4 stars.
S: If you'd like something cheap, I recommend Gonville.
U: Does the cheaper one include WiFi?
S: Yes, Gonville offers free WiFi.
U: Is it fast enough to watch online videos?
S: Yes, reviewers said that their WIFI is stable and fast.

Dialogue 2:

U: I'd like to find someplace to stay in downtown.
S: Both Avalon and Gonville are located near downtown. Would you like to stay in one of those guesthouses?
U: Which one has free wifi?
S: Both facilities do.
U: Is it high quality wifi with a stable connection?
S: Overall, customers thought the WIFI at Gonville is fast, while that at Avalon may be slow and unstable.

Dialogue 3:

U: I'm looking for a hotel called Avalon.
S: The Avalon hotel is located in downtown.
U: Does it have wifi?
S: Yes, it has wifi.
U: Is the wifi reliable?
S: According to previous reviews, some customers have complained that the WIFI is slow and make work a bit difficult. But one said it is fast.

Figure 1: Examples of the SK-TOD task. The top part shows two hotels and their customer reviews. The bottom part shows three dialogue sessions between the system (denoted by S) and three users (denoted by U). The last user utterance is a subjective question about the WIFI quality of the hotel(s). The system needs to retrieve information from the relevant subjective knowledge, which is highlighted in the review text.

Proceedings of The Eleventh Dialog System Technology Challenge, pages 274–281
September 11, 2023 ©2023 Association for Computational Linguistics

have different opinions and leave either positive or negative reviews.The subjectivity of reviews also indicates that a TOD system should consider multiple reviews to get a more comprehensive user opinion. Based on that, an ideal response should inform users of the diversity of opinions by including both positive and negative opinions as well as the proportions (like the response in Dialogue 3). Such a two-sided response has been observed as more credible and valuable for customers (Kamins et al., 1989; Lee et al., 2008; Baek et al., 2012), which can also protect the trust of users in the TOD system.

Building TOD upon subjective knowledge in this way brings in two unique challenges. First, instead of selecting the top few relevant knowledge snippets (as what is needed for Fact-TOD), the SK-TOD model needs to select all relevant knowledge snippets. Second, the model needs to aggregate these knowledge snippets into a concise response that can faithfully reflect the diversity and proportion of opinions. To facilitate the research of subjective-knowledge-grounded TOD, we released a large-scale dataset for this track, which contains 19,696 subjective knowledge-seeking dialogue contexts and manually annotated responses that are grounded on 143 entities and 1,430 reviews (8,013 sentences).

This paper provides an overview of the track. The rest of the paper is organized as follows. We first introduce the problem formulation (Section 2) and data statistics (Section 3). Then we introduce the baseline approach (Section 4) and the evaluation metrics (Section 5). Finally, we report the participants and results (Section 6), and we close the paper with the conclusion (Section 7).

2 Problem Formulation

Formally, we have a dialogue context $C = [U_1, S_1, U_2, S_2, \cdots, U_t]$ between a user and a system, where each user utterance U_i is followed by a system response utterance S_i except the last user utterance U_t. The dialogue involves single or multiple entities $\mathcal{E} = \{e_1, \cdots, e_m\}$. Along with the dialogue, we have a subjective knowledge source $\mathcal{B} = \{(e_1, \mathcal{R}_1), (e_2, \mathcal{R}_2), \cdots\}$ consisting of all the entities and their corresponding customer reviews. Each entity e has multiple reviews $\mathcal{R} = \{R_1, R_2, \cdots\}$. Each review can be split into multiple segments $[K_1, K_2, \cdots]$ such as paragraphs, sentences, or sub-sentential units. In this track, we regard each review sentence as a knowledge snippet.

	Train	Val	Test
# instances	14768	2129	2799
# seen instances	14768	1471	1547
# unseen instances	0	658	1252
# multi-entity instances	412	199	436
Knowledge Snippets			
Avg. # snippets per instance	3.80	4.07	4.21
Avg. # tokens per snippet	14.68	15.49	14.5
Dialogue			
Avg. # uttrances per instance	9.29	9.44	9.36
Avg. # tokens per request	8.65	8.94	9.12
Avg. # tokens per response	24.18	23.61	23.86

Table 1: Basic statistics of our dataset.

The SK-TOD task contains the following three sub-tasks.

Knowledge-Seeking Turn Detection aims to identify the user request that requires to be addressed with subjective knowledge. We regard it as a binary-classification problem, where the input is the dialogue context C and the output is a binary indicator.

Knowledge Selection is then used to select the knowledge snippets that are relevant to the user's request. The inputs are the dialogue context C and the knowledge snippets candidates $\mathcal{K}$, which is a combination of all the knowledge snippets of the relevant entities in $\mathcal{E}$. The output $\mathcal{K}^+ \subseteq \mathcal{K}$ is a subset of relevant knowledge candidates. Note that there might be multiple knowledge snippets in $\mathcal{K}^+$.

Response Generation is to create an utterance S_t that responds to the user's request based on the dialogue context C and the relevant knowledge snippets $\mathcal{K}^+$.

3 Data

We ground the data collection in MultiWOZ (Budzianowski et al., 2018; Eric et al., 2020) and select dialogues from the domains of hotels and restaurants. The data collection is conducted by a group of crowd workers through Amazon Mechanical Turk as described in (Zhao et al., 2023).

The dataset contains 19,696 instances with subjective user requests and subjective-knowledge-grounded responses in total. The average length of the subjective user request and the agent response is 8.75 and 24.07 tokens, respectively. While most of the instances contain a single entity, there are 1,047 instances where multiple entities are compared (like Dialogue 2 in Figure 1). Each instance requires on average 3.88 subjective knowledge snip-

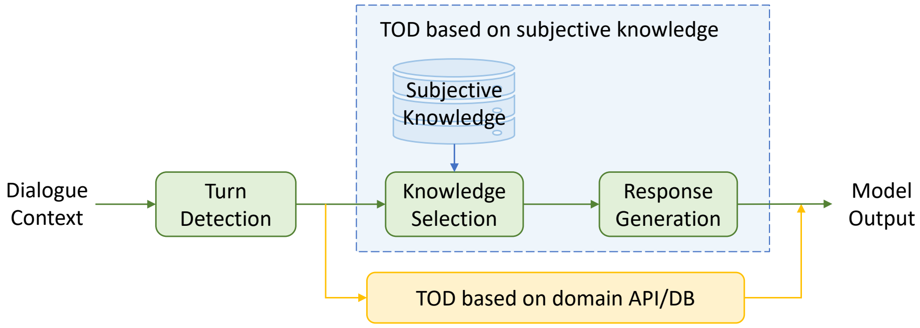


Figure 2: The pipeline architecture of SK-TOD.

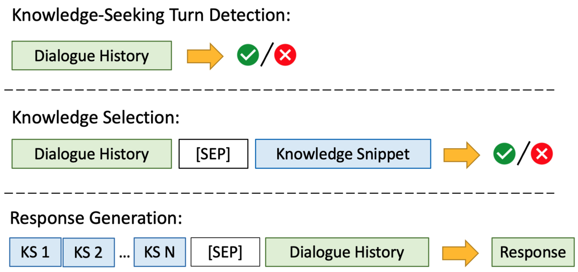


Figure 3: The Illustration of baselines for each subtask.

pets. To help identify the subjective knowledge-seeking user request, we randomly sample another 18,383 dialogues with non-subjective user requests from the original MultiWOZ dataset.

We split the dataset into training (75%), validation (10.8%), and test (14.2%) sets. Table 1 shows the detailed statistics of each subset. Our validation and test sets contain two subsets: the *seen* subset where the aspects of these instances are included in the training set, and the *unseen* subset where the aspects are not included in the training set. The unseen subset is designed to evaluate models' generalizability to arbitrary aspects.

4 Baseline

In this section, we describe the baseline method for SK-TOD. Figure 3 shows the pipeline of the baseline method which consists of three sequential sub-tasks. The details of each subtask are described as follows. They are illustrated in Figure 3.

4.1 Knowledge-Seeking Turn Detection

We employ a pre-trained language model (e.g., BERT (Devlin et al., 2019)) to encode C and adopt the hidden state of the first token as its representation. Then we apply a classifier to obtain the probability that the current user request is a subjective knowledge-seeking request. That is,

$$\begin{aligned} h &= \text{Enc}(C) \\ P(C) &= \text{softmax}\left(\text{FFN}\left(h\right)\right). \end{aligned} \tag{1}$$

The model is finetuned with the binary cross-entropy loss.

4.2 Knowledge Selection

Given a dialogue history, we first adopt a word-matching-based method used by Jin et al. (2021) to extract relevant entities. To select relevant knowledge snippets, we calculate the relevance score between the dialogue context C and a knowledge snippet $K \in \mathcal{K}$ of the corresponding entities. We regard it as a pairwise text scoring problem and consider the cross-encoder (Wolf et al., 2019) approach. We encode the concatenation of C and K to obtain the contextualized BERT representation. That is,

$$\begin{aligned} h &= \text{Enc}(C, K) \\ P(C, K) &= \text{softmax}\left(\text{FFN}\left(h\right)\right). \end{aligned} \tag{2}$$

During training, we use all relevant knowledge snippets to construct positive (C, K) pairs. Due to the large size of irrelevant knowledge snippets, we randomly sample the same number of irrelevant snippets to build negative pairs. We optimize the model using the binary cross-entropy loss. During inference, we predict the relevance probability of all knowledge snippets in the candidates. Since both precision and recall matter during KS, instead of selecting the top few results, we use a threshold to determine the relevance, which is estimated from the validation set.

4.3 Response Generation

We apply T5 (Raffel et al., 2020), a sequence-to-sequence pre-trained model as the generation model. The model receives the concatenation of dialogue context and the selected knowledge snippets as input, and has the target response as output. We train the model by maximizing $p(S_T \mid C, \mathcal{K}^+, Z)$. During the test, we generate the system response using beam-search with top-K sampling (Fan et al., 2018).

5 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 by comparing to the ground-truth labels and responses. For Knowledge-Seeking Turn Detection and Knowledge Selection, we report the precision, recall, F_1 score, and accuracy score. For Response Generation, following the evaluation of other generation tasks, we employ BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) to evaluate results compared to the reference 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:

$$S_p(X) = \frac{\sum_{x_i \in X} \left(s(x_i) \cdot f_1(x_i) \cdot \tilde{f}_1(x_i) \right)}{\sum_{x_i \in X} \tilde{f}_1(x_i)},$$

$$S_r(X) = \frac{\sum_{x_i \in X} \left(s(x_i) \cdot f_1(x_i) \cdot \tilde{f}_1(x_i) \right)}{\sum_{x_i \in X} f_1(x_i)},$$

$$S_f(X) = \frac{2 \cdot S_p(X) \cdot S_r(X)}{S_p(X) + S_r(X)}, \tag{3}$$

where $f_1(x)$ and $\tilde{f}_1(x)$ are the reference and prediction for the knowledge-seeking turn detection task, respectively, and $s(x)$ 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:

$$S_{overall}(e) = \frac{1}{|M|} \sum_{i=1}^{|M|} \frac{1}{rank_i(e)}, \tag{4}$$

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.

Based on the overall objective score, we selected the finalists to be manually evaluated by the following two crowd sourcing tasks:

- Appropriateness: whether the response is fluent and naturally connected to the dialogue context.
- Accuracy: whether the sentiment proportion provided by the response is accordant with that of the subjective knowledge.

For Appropriateness, we only show the dialogue context and the responses. For Accuracy, we further show the oracle knowledge snippets. To increase the annotation quality, we first ask workers to annotate the sentiment label of each knowledge snippet, and then evaluate the accuracy of each response. Both measures are evaluated using the 5-Point Likert scale.

In both tasks, we assigned each instance to three crowd workers and took their average as the final human evaluation score for the instance. Those scores were then aggregated over the entire test set following Equation 3, 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.

6 Results

We received 48 entries in total submitted from 14 participating teams. To preserve anonymity, the teams were identified by numbers from 1 to 14.

6.1 Objective Evaluation Results

Table 2 shows the objective evaluation results of the best entry from each team selected based on the overall score (Equation 4). The full results including all the submitted entries are available on the track repository[1].

[1] https://github.com/alexa/dstc11-track5

Table 2: Objective evaluation results of the best entries from each team. Bold denotes the best score for each metric; and * indicates the finalists selected for the human evaluations.

	Task 1: Detection			Task 2: Selection				Task 3: Generation				
Team-Entry	P	R	F	P	R	F	EM	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	0.9982	0.9979	0.9980	0.7901	0.7877	0.7889	0.3906	0.1004	0.1748	0.3520	0.1430	0.2753
1 - 0	0.9977	0.7702	0.8693	0.7872	0.6048	0.6841	0.3090	0.0873	0.1512	0.3061	0.1248	0.2400
2 - 3*	0.9940	0.9986	0.9963	0.8093	0.7858	0.7974	0.4156	0.0984	0.1774	**0.3658**	0.1509	0.2875
3 - 1	0.9986	0.9971	0.9979	0.7567	0.8242	0.7890	0.3966	0.0874	0.1641	0.3425	0.1372	0.2702
4 - 1	0.9982	0.9986	0.9984	0.8183	0.8442	0.8311	0.4895	0.1003	0.1731	0.3510	0.1422	0.2737
5 - 0	0.9986	0.9968	0.9977	0.7741	0.8557	0.8128	0.4675	0.0963	0.1719	0.3470	0.1379	0.2692
6 - 0*	0.9968	**0.9996**	0.9982	0.8039	**0.8775**	0.8391	0.5547	0.1017	**0.1894**	0.3629	0.1478	0.2804
7 - 4*	0.9979	0.9993	**0.9986**	0.8183	0.8506	0.8342	0.5314	0.1075	0.1744	0.3585	0.1459	0.2794
8 - 0*	0.9979	0.9982	0.9980	0.8240	0.8141	0.8190	0.5130	0.1029	0.1764	0.3587	0.1479	0.2822
9 - 1	**0.9995**	0.7691	0.8693	0.8385	0.6398	0.7258	0.3873	0.0788	0.1532	0.3141	0.1233	0.2406
10 - 0	0.9950	0.9986	0.9968	0.7955	0.7936	0.7946	0.4177	0.1035	0.1791	0.3598	0.1473	0.2812
11 - 0	0.9989	0.9982	0.9986	0.6540	0.5391	0.5910	0.3071	0.0980	0.1614	0.3368	0.1386	0.2713
12 - 2*	0.9986	0.9986	0.9986	0.7538	0.8227	0.7868	0.4291	0.0961	0.1715	0.3572	0.1467	0.2798
13 - 3*	0.9964	0.9982	0.9973	**0.8590**	0.8449	**0.8519**	**0.6432**	**0.1081**	0.1819	0.3652	0.1528	0.2872
14 - 0*	0.9979	0.9989	0.9984	0.7856	0.8035	0.7944	0.4183	0.1066	0.1748	0.3599	**0.1577**	**0.2899**

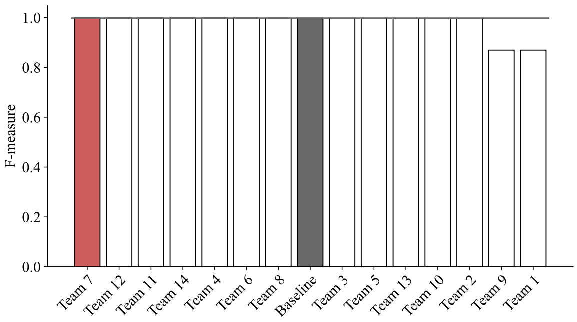


Figure 4: Knowledge-seeking turn detection performance (F-measure) from different entries. The horizontal line indicates the baseline performance.

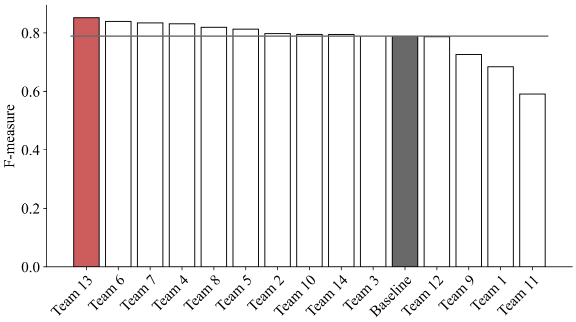


Figure 5: Knowledge selection performance (F-measure) from different entries. The horizontal line indicates the baseline performance.

Table 3: Human evaluation results. Bold indicates the best score for each metric.

Rank	Team	Entry	Accuracy	Appropriateness	Average
Ground-truth			2.9189	3.6422	3.2806
1	6	0	2.9095	**3.6596**	**3.2846**
2	8	0	2.9005	3.6535	3.2770
3	13	3	**2.9100**	3.6321	3.2710
4	2	3	2.8908	3.6487	3.2697
5	7	4	2.9046	3.6348	3.2697
6	12	2	2.8856	3.6518	3.2687
7	14	0	2.8912	3.6427	3.2670
Baseline			2.8715	3.6348	3.2531

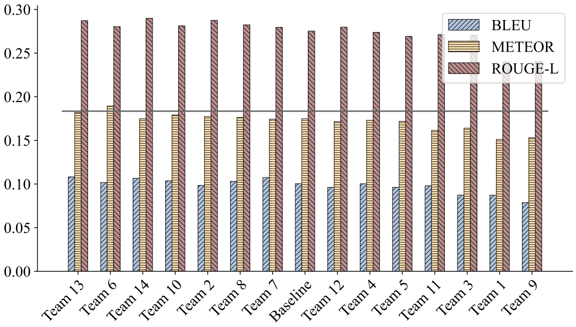


Figure 6: Knowledge-grounded generation performance from different entries. The horizontal line indicates the average of the baseline scores.

Figure 4 shows that the majority including the baseline achieved near-perfect knowledge-seeking turn detection results with F-measure surpassing 0.99. This can be attributed to the characteristics of this data, where knowledge-seeking and non-knowledge-seeking turns are easily distinguishable from each other. For knowledge selection, the majority of the teams submitted the improved results over the baseline (Figure 5), while only half of the teams achieved higher average scores than the baseline for response generation (Figure 6). Team 13 was determined to be the best team based on the overall objective scores (Equation 4), derived from the highest F-measure and exact matching accuracy for knowledge selection, along with the averaged response generation scores.

6.2 Human Evaluation Results

We selected 7 finalists to be manually evaluated, corresponding to the best entry from each of the top half teams in the overall objective score (Equation 4). Table 3 shows the official ranking of the finalists based on the human evaluation results. Team 3 achieved the highest accuracy score, which aligns with their knowledge selection results from the objective evaluation. On the other hand, Team 6 achieved the highest appropriateness ratings, which were even higher than the scores for the reference responses. This led them to be the final winner based on the average score between accuracy and appropriateness.

To further analyze the correlation between each automatic metric and the final human evaluation results, 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 7. Consistent with the findings from our previous challenge tracks (Kim et al., 2021a, 2022), the knowledge selection metrics have the highest correlation with the accuracy ratings in the human evaluation. Among the response generation metrics, METEOR shows better alignment with the human ratings, particularly as it is the only metric with a positive correlation with appropriateness ratings. Nonetheless, none of the automatic evaluation metrics showed a strong correlation with the appropriateness ratings. This suggests a new research direction aimed at developing more reliable metrics for this task.

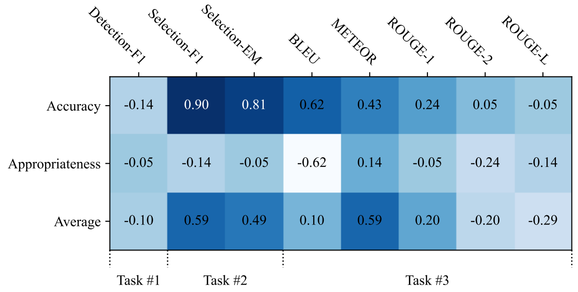


Figure 7: Correlations between the objective and human evaluation metrics in Spearman's ρ. The higher score of a pair of metrics has, the stronger correlation they have.

7 Conclusion

We presented the official evaluation results of the Task-Oriented Conversational Modeling with Subjective Knowledge Track in DSTC11 This challenge track addressed the new conversational modeling tasks to accommodate subjective user requests into task-oriented dialogue systems. A total of 14 teams participated with an overall number of 48 entries submitted. One notable thing is that some teams have used the large language models (LLMs) such as OpenAI's ChatGPT and GPT-4 for enhancing data augmentation and refining response ranking. While these methods were expected to help to improve our task performances, it was not a decisive factor to achieve elevated scores and win the benchmark in both automatic and human evaluations. This suggests a future research direction exploring more effective strategies for utilizing LLMs for the tasks.

References

Hyunmi Baek, JoongHo Ahn, and Youngseok Choi. 2012. Helpfulness of online consumer reviews: Readers' objectives and review cues. *International Journal of Electronic Commerce*, 17(2):99–126.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 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*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eleftherios Dimitrakis, Konstantinos Sgontzos, Panagiotis Papadakos, Yannis Marketakis, Alexandros Papangelis, Yannis Stylianou, and Yannis Tzitzikas. 2018. On finding the relevant user reviews for advancing conversational faceted search. In *EMSASW@ ESWC*, pages 22–31.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Di Jin, Seokhwan Kim, and Dilek Hakkani-Tur. 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 Conver-*

sational Question Answering (DialDoc 2021), pages 119–127.

Michael A Kamins, Meribeth J Brand, Stuart A Hoeke, and John C Moe. 1989. Two-sided versus one-sided celebrity endorsements: The impact on advertising effectiveness and credibility. *Journal of advertising*, 18(2):4–10.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 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*, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.

Seokhwan Kim, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, and Dilek Hakkani-Tur. 2021a. Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access track in dstc9.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tür. 2021b. "how robust ru?": Evaluating task-oriented dialogue systems on spoken conversations. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1147–1154. IEEE.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Behnam Hedayatnia, Karthik Gopalakrishnan, and Dilek Hakkani-Tür. 2022. Knowledge-grounded task-oriented dialogue modeling on spoken conversations track at dstc10. In *Proceedings of the AAAI-22 Workshop on Dialog System Technology Challenges (DSTC10)*.

Jumin Lee, Do-Hyung Park, and Ingoo Han. 2008. The effect of negative online consumer reviews on product attitude: An information processing view. *Electronic commerce research and applications*, 7(3):341–352.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2022. Achieving conversational goals with unsupervised post-hoc knowledge injection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3140–3153, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

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

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local memory pointer networks for task-oriented dialogue. In *International Conference on Learning Representations*.

Chao Zhao, Spandana Gella, Seokhwan Kim, Di Jin, Devamanyu Hazarika, Alexandros Papangelis, Behnam Hedayatnia, Mahdi Namazifar, Yang Liu, and Dilek Hakkani-Tur. 2023. "what do others think?": Task-oriented conversational modeling with subjective knowledge.