KETOD: Knowledge-Enriched Task-Oriented Dialogue

Anonymous ACL submission

Abstract

Existing studies in dialogue system research mostly treat task-oriented dialogue and chit-chat as separate domains. Towards building a human-like assistant that can converse naturally and seamlessly with users, it is important to build a dialogue system that conducts both types of conversations effectively. In this work, we investigate how task-oriented dialogue and knowledge-grounded chit-chat can be effectively integrated into a single model. To this end, we create a new dataset, KETOD (Knowledge-Enriched Task-Oriented Dialogue), where we naturally enrich task-oriented dialogues with chit-chat based on relevant entity knowledge. We also propose two new models, SimpleToDPlus and Combiner, for the proposed task. Experimental results on both automatic and human evaluations show that the proposed methods can significantly improve the performance in knowledge-enriched response generation while maintaining a competitive task-oriented dialog performance. We believe our new dataset will be a valuable resource for future studies. The code and the dataset will be made publicly available.

1 Introduction

Dialogue systems have achieved substantial progress (Zhang et al., 2020; Hosseini-Asl et al., 2020a; Tao et al., 2021) due to recent success in language model pre-training (Radford et al., 2019; Raffel et al., 2020; Lewis et al., 2020). One major type of dialogue being studied is task-oriented dialogue (TOD) (Wen et al., 2017a; Budzianowski et al., 2018; Rastogi et al., 2020; Hosseini-Asl et al., 2020a), where the system aims to collect user intents/goals to complete certain tasks (e.g. restaurant-booking). In most of TOD systems, the system responses are concise and templated, as we only focus on the success of task completion but not providing a natural and engaging conversational experience. The latter is the target of another

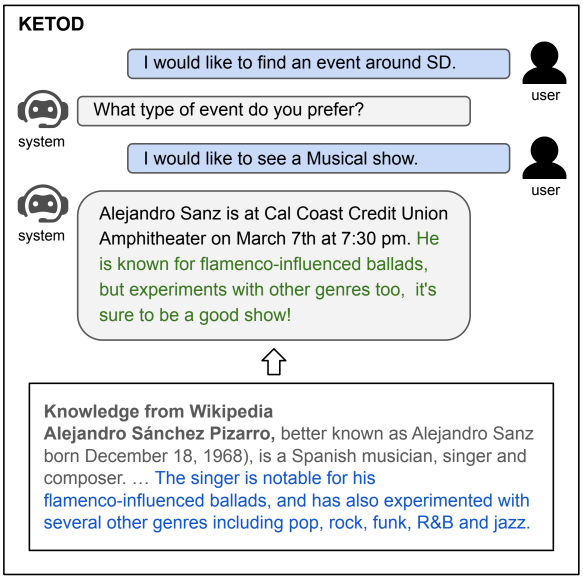

Figure 1: An example from the KETOD dataset: the green text is our enriched chit-chat based on the entity knowledge of *Alejandro Sanz* in the original TOD. Such knowledge-grounded chit-chat makes the dialogue more natural and engaging.

kind of popularly studied dialogue - knowledge-grounded chit-chat (Ghazvininejad et al., 2018; Zhang et al., 2018; Tuan et al., 2019; Dinan et al., 2019). Knowledge-grounded chit-chat enables dialog systems to access external knowledge so that they can provide more engaging and knowledgeable conversations and in the same time reduce hallucinations (Shuster et al., 2021).

Existing studies mostly focus on one specific type of dialogue, either task-oriented dialogue or knowledge-grounded chit-chat. However, the ultimate goal of Conversational AI is a human-like, unified system capable of conversing with the users naturally and seamlessly among all kinds of dialogues. Current TOD systems can hardly make interesting and engaging conversations only with templated functional responses. Few previous works like ACCENTOR (Sun et al., 2021) have studied the combination of TOD and chit-chat, but their chit-chat augmentation is largely limited to simple general responses like ‘you’re welcome’, ‘sounds good to me’. In this work, we propose to enrich TOD with knowledge-grounded chit-chat, as one step further towards the ultimate goal of building a human-like, unified system (See Figure 1 for an example). We believe that the proposed knowledge-enriched TOD system can conduct more social, natural, and engaging conversations.

To this end, we propose a new dataset, KETOD (Knowledge-Enriched Task-Oriented Dialogue). In order to obtain natural and high-quality knowledge-grounded chit-chat, we design the dataset construction framework by augmenting existing TODs and using the relevant entity knowledge to make the chit-chat enrichment. Specifically, for a given TOD, 1) extracting the entities from the dialogue states and actions; 2) retrieving the knowledge associated with the entities from external knowledge sources; 3) asking the human annotators to enrich the system responses with chit-chat using the retrieved knowledge. We demonstrate that the knowledge-enriched dialogues constructed with the proposed framework are consistently preferred by human judges across all axes of engagingness, interestingness, knowledge, and humanness.

We propose two models, and study the challenges and insights of our new dataset. The first model is an end-to-end language model that jointly learns and generates both the TOD results (dialogue states and actions) and the knowledge-enriched responses. The second model is a pipeline that first generates the TOD results, then uses another response generation model to generate the knowledge-enriched responses. We run comprehensive experiments to demonstrate the improvement over the baselines, and show that our models can generate better knowledge-enriched responses while maintaining competitive performance on the TOD tasks. To summarize, we make the following major contributions:

- We propose the task of combining TOD and knowledge-grounded chit-chat.
- We construct a new large-scale dataset, KETOD, with high-quality, manually annotated dialogue responses enriched with knowledge-grounded chit-chat. We will release the dataset upon acceptance of the paper.
- We propose two models for our dataset, and carry comprehensive experiments to study the challenges and insights. We believe our dataset should be a valuable resource for building a human-like conversational assistant.

2 Related Work

Task-oriented dialogue. Task-oriented dialogue (TOD) has been one of the most popular types of dialogue in the research community. There have been many works on building each component of the TOD system, such as dialogue state tracking, action prediction, and response generation (Wen et al., 2015, 2017b; Mrksic et al., 2017; Zhong et al., 2018; Eric et al., 2020; Liu et al., 2018; Peng et al., 2017; Zhou et al., 2017). Later works begin to investigate building end-to-end systems (Bordes et al., 2017; Liu et al., 2018, 2017; Xu et al., 2020). Most recent works on TOD also apply such language model pre-training style methods on building end-to-end systems (Hosseini-Asl et al., 2020a; Peng et al., 2020; Su et al., 2021), achieving top performances on various datasets. Popular datasets in TOD include the DSTC challenge series (Williams et al., 2016), MultiWOZ (Budzianowski et al., 2018), SGD (Rastogi et al., 2020), etc. As the primary goal of TOD is the successful completion of the functional tasks, the system responses are mostly concise and templated.

Chit-chat dialogue. Another type of popular studied dialogue is chit-chat, with the goal of making a natural and engaging conversation. Apart from the ‘pure’ simple chit-chat that mostly covers plain and general responses, more works focus on knowledge groundings to achieve better specificity and engagingness, such as using user profiles (Zhang et al., 2018), social media contexts (Sordoni et al., 2015), or knowledge graphs (Tuan et al., 2019; Moon et al., 2019), etc. In this work, our enriched chit-chat is grounded on open-domain knowledge, similar as the Topical-Chat (Gopalakrishnan et al., 2019) and the WOW dataset (Dinan et al., 2019), where the system converses with the users about certain topics involving entity knowledge in an open-ended setting. In contrast, their datasets specifically focus on knowledge-grounded chit-chat, while our dataset combines TOD and such chit-chat.

Combination of task-oriented dialogue and chit-chat. ACCENTOR (Sun et al., 2021) proposes to combine TOD with chit-chat by prepending or appending chit-chat to the TOD system responses. But their chit-chat is mostly general responses like ’sounds good!’, ’you’re welcome’. FusedChat (Young et al., 2021) proposes to insert chit-chat turns into TOD as well as re-writing

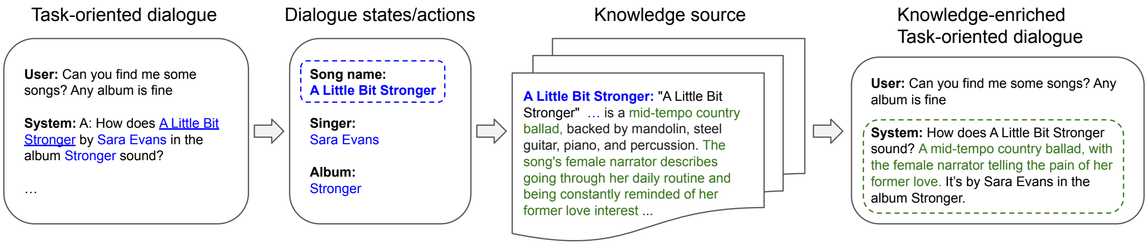

Figure 2: The pipeline of dataset construction: for each task-oriented dialogue, we first extract all the entities from the dialogue states and actions. Then we retrieve the knowledge associated with each entity from external knowledge sources (Wikipedia). At last, we ask human annotators to enrich the TOD system responses with chit-chat grounded on the retrieved knowledge.

TOD turns, but their chit-chat is still mostly general responses or based on commonsense knowledge. Kim et al. (2020) propose to insert additional turns into TOD, where the system needs to respond based on the knowledge from domain FAQs. The DSTC10 task 2 (Kim et al., 2021) is based on the dataset from (Kim et al., 2020) with a similar focus. HyKnow (Gao et al., 2021) also proposes to insert turns into TOD grounded on knowledge from unstructured documents. These datasets focus on the challenge of detecting those turns requiring external knowledge and selecting the knowledge to generate the responses. In contrast, our dataset focuses on injecting knowledge-grounded chit-chats into the original TOD responses, to make the dialogue more natural and engaging. Our dataset poses more challenges in selecting knowledge based on the dialogue context and generating the responses with both the correct TOD information and the chit-chat seamlessly.

3 The KETOD Dataset

3.1 Dataset Construction

In this section, we describe our framework to construct the KETOD dataset. We start from existing TOD datasets and employ human annotators to augment the functional system responses with knowledge-grounded chit-chat. The proposed approach is demonstrated to give natural, contextual-relevant knowledge enrichment, and meanwhile easy to scale to different datasets. Figure 2 gives an overview of the dataset construction pipeline.

Data preparation. We build upon the SGD dataset (Rastogi et al., 2020), with TOD spanning 16 domains, such as `Restaurant`, `Wheather`, etc. Given each TOD, to obtain the knowledge relevant to the dialogue context, we first extract all the entities from the dialogue states and actions. We exclude the domains `Alarm`, `Banks`, and `Payment` as there are mostly no entities involved in these domains; Also, to simplify the human annotation process in the next step, we remove the dialogues with over 10 entities involved.

Knowledge retrieval. For each entity, we use the concatenation of the domain name and entity name as the query to retrieve Wikipedia articles. We use the DrQA retriever (Chen et al., 2017) to retrieve the top 2 Wikipedia articles and take the first 2 paragraphs of each article as the knowledge candidates associated with each entity. Then we break the retrieved articles into sentences, with each sentence as one knowledge snippet.

Response enrichment. In this step, we employ human annotators to enrich the system responses in the original TOD based on the dialogue context and the retrieved knowledge. For each TOD, we present to the annotators the full dialogue, as well as all the knowledge snippets associated with the entities in the dialogue. The annotators can click on each entity name to see the associated knowledge snippets in an expanded textbox. See Appendix A for our annotation interface.

The annotation process is as follows: 1) Read the full dialog first to have an overall story in mind, as well as the relevant knowledge snippets, then to decide how many turns to enrich with chit-chat and which turn(s) to enrich; If there is no way to make a natural chit-chat enrichment, skip the example. 2) After deciding the turn(s) to enrich with the chit-chat, select the knowledge snippets used to make the enrichment (at most 3 snippets for each turn); 3) Rewrite the system response to enrich with chit-chat grounded on the selected knowledge snippets; The functional information in the original response should be maintained, while may be rephrased to make the enriched response more natural.

To ensure the dataset quality, we first interview the annotators to select the appropriate hires through a few test examples. Then we launch a training session for all the annotators to learn the task and the annotation interface. We launch the official batches after the annotators can well-master the task. During annotation, we specifically emphasize the contextualization of the knowledge-grounded chit-chat - the enrichment should be contextualized closely on the dialogue context, but not a plain restatement of the knowledge snippets.

3.2 Dataset Statistics and Analysis

We end up with 5,324 dialogues with enriched system responses. We make the split of 4,247/545/532 as the train/dev/test set. Table 1 shows the statistics of the KETOD dataset. Around 12.1% of the turns (which indicates mostly 1 or 2 turns in one dialogue) are enriched with knowledge-grounded chit-chat. This intuitively complies with our goal of making the whole dialogue natural and engaging, since too frequent chit-chat may result in redundancy and unnaturalness.

Quality assessment of the annotation. During the annotation process, around 12% of the dialogues cannot be enriched with any turns and thus discarded. It takes around 100 seconds for the annotators to finish each dialogue. To assess the quality of the annotation, we sample 5% of the annotated dialogues and distribute them to linguistics to check: 1) If the chit-chat enrichment is relevant and natural; 2) If the knowledge snippets are accurately selected corresponding to the enrichment. We end up with a correct rate of 87.0%.

Justification of the chit-chat enrichment. To demonstrate that our proposed knowledge-enriched TOD can be more natural and engaging, we conduct human evaluations to compare KETOD dialogues and their corresponding original TOD dialogues without chit-chat enrichment (SGD). We follow (Li et al., 2019) to make pairwise comparisons of the full dialogues over the following four axes: engagingness, interestingness, knowledge, and humanness. The results in Figure 3 show the superiority of KETOD over all axes.

Dialogues	5,324
Vocabulary	27k
All turns	52,063
Turns enriched with chit-chat	6,302
All entities	4,639
All knowledge snippets	33,761
Avg. # turns per dialogue	9.78
Avg. # tokens in enriched responses	28.07
Avg. # entities per dialogue	4.98
Avg. # knowledge snippets per dialogue	70.50

Table 1: General statistics of KETOD.

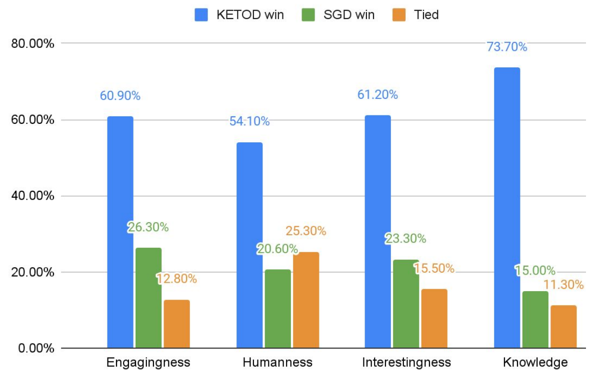

Figure 3: Results of pairwise comparison of KETOD vs SGD.

4 Approaches

In this section, we will describe the proposed two models for the KETOD dataset.

4.1 Overview and Formulations

For each dialogue turn, denote the dialogue context (history) as C, belief states as B, database search results as D, actions as A, the knowledge snippets used for chit-chat enrichment as K, the response as T. Then we formulate the problem as: given the dialogue context C and a knowledge source (Wikipedia in this dataset), the target is to generate the belief states B, actions A, and the response T, which may be enriched with chit-chat grounded on the knowledge based on the context. The goal of the optimization on KETOD is two-folded: 1) Optimizing the generation of knowledge-enriched responses; 2) Maintaining the task performances;

In this work, we propose the following modeling framework on KETOD: 1) given the dialogue context, generate the belief states and actions; 2) extract the entities in the belief states and actions, then use these entities to retrieve knowledge candidates (similar as in the dataset construction process); 3) conditioned on the dialogue context, use a knowledge selection model to select knowledge snippets from the knowledge candidates retrieved; 4) generate the knowledge-enriched response conditioned on both the dialogue context and the selected knowledge snippets.

Based on the above general framework, we propose two architectural approaches, **SimpleToD-Plus** and **Combiner**, respectively in §4.3 and §4.4.

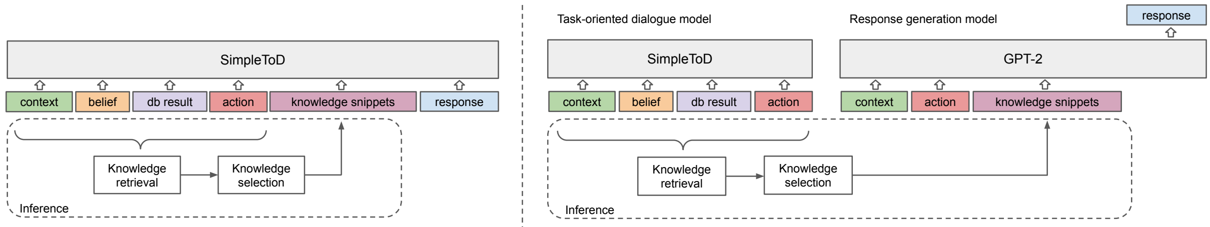

Figure 4: Illustration of the models. *Left*: the SimpleToDPlus model; *Right*: the Combiner model;

4.2 Knowledge Selection

After the generation of belief states and actions, we retrieve the knowledge snippet candidates from Wikipedia using the entities in the belief states and actions. The average number of knowledge snippets candidates retrieved for each dialogue is around 70. It is impractical to input all of them into the models due to the large amount. As we have the annotation for the ground truth knowledge snippets used for each chit-chat enrichment, we train a knowledge selection model to select the top knowledge snippets most appropriate for chit-chat enrichment. Specifically, we concatenate the dialogue context with each knowledge snippet as the input. Then we use BERT (Devlin et al., 2019) to train a simple classifier to rank all the knowledge snippets candidates. We take the top 3 ones as the knowledge selection results. We use the same knowledge selection model for both architectures.

4.3 SimpleToDPlus

SimpleToD (Hosseini-Asl et al., 2020b) is a recent popular approach on TOD, which uses one single language model to sequentially generate the belief states, actions, and responses. It has achieved strong performances in all the above functional tasks. In this work, we propose its extension, **SimpleToDPlus**, to generate knowledge-enriched responses for TOD. The left part of Figure 4 shows the overview of SimpleToDPlus. We formulate the training sequence as:

$$[C, B, D, A, K, \text{<chitchat>}, T] \qquad (1)$$

Where <chitchat> is a tag to indicate the decision of whether to enrich the response with knowledge grounded chit-chat or not. If the response is not enriched, we insert the tag <nochitchat>. Since the number of the gold knowledge snippets varies from 1 to 3 (as in the dataset construction), to be compatible with inference time, here we first run the knowledge selection model on all training instances. Then we construct the knowledge snippets K as the merge of the gold knowledge snippets and the knowledge selection model results, truncated to 3 ones. If the response is not enriched with chit-chat, i.e., no gold knowledge snippets, we still put 3 snippets from the knowledge selection model ranking results here during training.

In the inference time, we first sequentially generate the belief states and actions. Then we extract the entities from the generated belief states and actions, and apply the same process of knowledge retrieval as in dataset construction. Next, we run the knowledge selection model on the retrieved knowledge candidates and take the top 3 knowledge snippets as the model input followed by the generated actions. At last, the model generates the decision to make chit-chat enrichment or not, followed by the final response.

Since the knowledge-enriched response is conditioned on the entity knowledge from the belief states and actions, we need to directly include the entities in the actions and responses during generation, instead of generating a delexicalized result first and then lexicalizing in the post-process as in the original SimpleToD. To simplify, we use the oracle database search results for all the experiments.

4.4 Combiner

SimpleToDPlus models all the generations in an end-to-end manner. In **Combiner**, we use a pipeline of a TOD model followed by a response generation model to separate the TOD part (belief states, actions) with the generation of knowledge-enriched responses. The goal is to study whether an independent model can better learn each task with less interference from the other. The overview of the architecture is shown on the right of Figure 4.

For the TOD model, we use SimpleToD to generate the belief states and actions, with the training sequence as:

$$[C, B, D, A] \qquad (2)$$

We find that including the knowledge-enriched responses during training degrades the task per-

Models	Joint GA	Avg GA	Act-Slot F1	BLEU-4$_{aug}$	BLEU-4$_{orig}$	BLEU-4$_{all}$
SimpleToD-ref	27.6	54.2	67.6	-	-	-
SimpleToD	23.7	50.1	62.7	4.8	10.7	10.0
SimpleToDPlus	**28.6**	**52.2**	**66.9**	6.3	**11.7**	**11.0**
Combiner	24.5	51.5	64.5	**6.5**	9.9	9.5

Table 2: Main experiment results: Both SimpleToDPlus and Combiner outperform the baseline. Overall SimpleToDPlus obtains better response generation performance while maintaining competitive TOD performance.

formance, indicating the disturbance from the ungrounded knowledge in the responses.

For the response generation model, we use GPT-2 (Radford et al., 2019) with the concatenation of the dialogue context, actions, and the knowledge snippets as the prompt:

$$T = \text{GPT-2}(C, A, K) \quad (3)$$

We use the same way of constructing the merged knowledge snippets during training, and the same process of knowledge retrieval and selection during inference as in SimpleToDPlus.

5 Experimental Results

Baseline model. We use SimpleToD (Hosseini-Asl et al., 2020b) as our baseline model, i.e., with the training sequence as [C, B, D, A, T], without the injection of knowledge snippets. Therefore the knowledge-grounded chit-chat in the responses T do not have any knowledge groundings - we aim to show the necessity of knowledge grounding for our task, as well as the effectiveness of our proposed models to incorporate knowledge.

Experimental setups and evaluations. Check Appendix B for details of model training and parameter settings. For the TOD performances, we evaluate the belief states with joint goal accuracy (Joint GA) and average goal accuracy (Avg GA), and the actions with act-slot F1, same as (Sun et al., 2021). For the automatic evaluations of response generation, we use three BLEU-4 scores: BLEU-4$_{aug}$ for evaluating the responses enriched with knowledge; BLEU-4$_{orig}$ for evaluating the responses not enriched with knowledge; BLEU-4$_{all}$ for evaluating all responses;

5.1 Main Results

Performance on response generation. Table 2 shows our main experiment results. For the performances on response generation, we can see that both of our proposed models, SimpleToDPlus and Combiner, improve on the knowledge-enriched response generation (BLEU-4$_{aug}$) over the SimpleToD baseline. Since in the baseline, we do not include the knowledge snippets in the input, the generated responses are mostly enriched with random knowledge or frequent knowledge in the training data. The improvements demonstrate the necessity of knowledge grounding and the effectiveness of the proposed knowledge enrichment methods. Combiner performs slightly better on knowledge-enriched responses than SimpleToDPlus but falls short on the responses without knowledge-enrichment (i.e., original TOD responses). This is partially because of its pipeline nature - a separated response generation module can better learn the knowledge enrichment without the disturbance of other tasks, but the error cascading from the generated actions degrades the performance of the TOD responses part.

Performances on belief states and actions. To better study how the knowledge enrichment affects the TOD performances, we first train SimpleToD on our dataset without the knowledge enrichment, i.e., replace all the knowledge-enriched responses with the original responses in SGD. We name it as SimpleToD-ref in Table 2, serving as a reference of the original TOD performances. The SimpleToD baseline gives largely degraded performances due to the disturbance from the ungrounded knowledge in the responses during training. Therefore in Combiner, we do not include the responses in the training sequences of the TOD model (specified in section 4.4), and obtain better scores. SimpleToDPlus achieves the best TOD performances, which are nearly competitive with SimpleToD-ref. This is partially due to the enhancement of language modeling ability brought by the training on the responses grounded on the input knowledge.

Human evaluations. In order to get the more comprehensive measure of the response generation performances, we conduct human evaluations for both dialogue-level pairwise comparison and turn-level factualness evaluation. For dialogue-level pairwise comparison, we randomly sample 200 dialogues

Metrics	SimpleToDPlus win (%)	Combiner win (%)	Tied (%)
Engagingness	47.8	24.5	27.8
Interestingness	34.5	19.0	46.5
Knowledge	29.5	26.3	44.3
Humanness	43.3	23.8	33.0

Table 3: Human evaluation of SimpleToDPlus vs. Combiner.

Metrics	SimpleToDPlus win (%)	Gold win (%)	Tied (%)
Engagingness	16.8	60.5	22.8
Interestingness	12.0	51.0	37.0
Knowledge	14.5	44.8	40.8
Humanness	17.3	58.0	24.8

Table 4: Human evaluation of SimpleToDPlus vs. Gold.

from the test set and apply the same process as in dataset evaluation (3.2). For each model, we construct the full dialogue results by concatenating the generated response for each turn given the gold dialogue context. Table 3 shows the results of pairwise comparison between the SimpleToDPlus model and the Combiner model, demonstrating SimpleToDPlus is more performant. Table 4 shows the results of pairwise comparison between SimpleToDPlus and the gold reference, indicating there is still a large room for further improvements. See Appendix C for the human evaluation results of comparing both methods to the baseline. For turn-level factualness evaluation, we randomly sample one turn with chit-chat enrichment from each dialogue, and present both the generated response and the selected knowledge snippets to the annotators. The annotators are asked to check whether the chit-chat in the responses are factually correct based on the knowledge snippets. SimpleToDPlus and Combiner obtain the factualness correct rate of 64.2% and 66.1%, respectively. In summary, Combiner achieves better factualness of knowledge enrichment since its independent response generation model can better focus on the learning of knowledge groundings. But its error cascading due to the pipeline nature may degrade the overall consistency and human-likeness of the generated dialogue.

As we have two optimization goals on KETOD 1) Optimizing the generation of knowledge-enriched responses; 2) Maintaining the task performances, we consider SimpleToDPlus as a better model regarding the overall performances. We will use the results of SimpleToDPlus for the ablations and other analyses in the rest of the experiments.

	BLEU-4_{aug}	BLEU-4_{all}
Given gold TOD results, decision, and knowledge		
SimpleToD	6.5	13.1
SimpleToDPlus	9.7	14.6
Combiner	14.6	15.1
Given gold TOD results		
SimpleToD	6.3	12.8
SimpleToDPlus	7.4	14.0
Combiner	9.6	13.9

Table 5: Analysis of different inference stages: we provide the models with gold results up to certain stages, and investigate the performances for the inferences on following stages.

	BLEU-4_{aug}	BLEU-4_{all}	Knowledge selection recall (%)
Gold	9.7	14.6	100.0
BERT selection	7.8	14.4	52.7
TF-IDF selection	6.6	13.7	14.1

Table 6: SimpleToDPlus response generation performance with varying knowledge selection strategies.

5.2 Ablations and Analysis

Analysis of different inference stages. There are several inference stages for this task - the TOD results (belief states and actions), the selection of knowledge snippets, and the final response generation, where each stage is conditioned on previous results. Therefore the errors accumulate through all the stages leading to the final performances. Here we run another two sets of experiments to study such error accumulations and compare the two models. Specifically, first, we feed the models with the gold TOD results, chit-chat decisions, and knowledge snippets, to solely test the abilities to generate the knowledge-enriched responses; Second, we feed the models with the gold TOD results to test the following stages of knowledge selection and the response generation. The results are shown in Table 5. Compared with the full inference results in Table 2, we can see that the Combiner model largely outperforms SimpleToDPlus if provided with more gold results for previous stages. However, it gradually falls behind SimpleToDPlus when moving to fully end-to-end inference due to the error cascading of its pipeline nature.

Importance of knowledge selection strategies. To demonstrate the importance of the knowledge selection strategies (and their subsequent recall performance), we run SimpleToDPlus with 1) gold knowledge snippets; 2) predicted knowledge snippets (with BERT); 3) knowledge snippets selected by heuristics (we use TF-IDF matching between the current dialogue turn and the knowledge snip-

Gold	Knowledge snippets	Predicted	Knowledge snippets
User: Sure, that sounds right. Could you help me find some songs? **System:** I found 10 songs for you. One song is Supercut by Lorde from the album Melodrama. **User:** What other songs did you have? I like Battle Beast. **System:** They're quite a well-known power metal band. I found 1 song for you. It is called Eden by Battle Beast from the album No More Hollywood Endings?	**Battle Beast** Battle Beast is a Finnish heavy/power metal band established in Helsinki in 2005. The initial line-up consisted of guitarists…	**User:** Sure, that sounds right. Could you help me find some songs? **System:** I have 10 songs for you. Would you like to listen to Supercut by Lorde from the album Melodrama? **User:** What other songs did you have? I like Battle Beast. **System:** I found 1 song by Battle Beast from the album No More Hollywood Endings. It is a metal song by Battle Beast.	**Battle Beast** Battle Beast is a Finnish heavy/power metal band established in Helsinki in 2005. The initial line-up consisted of guitarists…
User: I would like to find an event around SD. **System:** What type of event do you prefer? **User:** I would like to see a Musical shows. **System:** Alejandro Sanz is at Cal Coast Credit Union Amphitheater on March 7th at 7:30 pm. He is known for flamenco-influenced ballads, but experiments with other genres too, so it's sure to be a good show!	**Alejandro Sánchez Pizarro…** The singer is notable for his flamenco-influenced ballads, and has also experimented with several other genres including pop, rock, funk, R&B and jazz.	**User:** I would like to find an event around SD. **System:** What type of event would you like? Music, Sports, or something else? **User:** I would like to see a Musical shows. **System:** I have 10 events that you might like. One is Alejandro Sanz at Cal Coast Credit Union Amphitheater. He is a Spanish singer-songwriter and musician.	**Alejandro Sánchez Pizarro,** is a Spanish musician, singer and composer…

Figure 5: Case studies: two examples of comparing the generation from SimpleToDPlus (right) with the gold reference (left), together with the knowledge snippets selected. Overall our model can mostly generate reasonable knowledge enrichment, but still falls short on engagingness and consistency compared to the gold references.

	BLEU-4$_{aug}$	BLEU-4$_{all}$	Enrichment decision F1 (%)
Gold decision	9.7	14.6	100.0
Predicted decision	8.0	14.1	58.7

Table 7: SimpleToDPlus response generation performance using (1) the gold set of turns to enrich with chit-chat, and (2) the predicted set of turns.

	All	Hotels	Movies	Restaurant	Music
BLEU-4$_{aug}$	6.3	7.1	5.2	5.1	7.7
BLEU-4$_{all}$	11.0	10.3	12.2	14.0	12.3

Table 8: Domain breakdown of SimpleToDPlus response generation performances.

pets). To eliminate the influences brought by other inference stages, we feed the model with gold TOD results (dialogue states and actions). The results are shown in Table 6. There exists a certain level of variance for knowledge selection, e.g., when recommending a song for the user, you may talk about its genre, its singer, or the album.

Learning *when* to inject knowledge-enriched chit-chat. In all models, we use the special token '<chitchat>' and '<nochitchat>' to indicate the decision to inject knowledge enrichment for the responses. To study the effect of the chit-chat injection decision-making accuracy on the overall dialogue tasks, we run SimpleToDPlus (1) with the ground-truth information of turns to enrich with chit-chat, and (2) with the predicted decisions, using the gold TOD results. Table 7 shows the performance gap, which highlights the importance of knowing *when* to inject knowledge-enriched chit-chat. While such decisions are conditioned on the dialogue history, e.g., we may tend to not enrich a turn if many of the previous turns are enriched to avoid redundancy, there also exists some variance. In a real system, we may consider specifying the turns to make the chit-chat enrichment instead of letting the model make the decision.

Domain analysis. We investigate the model performance for each domain in Table 8. We observe that the performance differences may depend on the variance of the enriched knowledge. Domains with larger variance on selected knowledge tend to have lower automatic scores. For example, in `Hotels` domain, mostly the chit-chat is about the locations since there are mostly location entities involved in this domain. But for the `restaurants` domain, the enriched knowledge can be about the food, the restaurant, as well as the location. The selected knowledge shows more diversity and variance.

We provide case studies in Figure 5 to compare the predicted results with the gold references.

6 Conclusion

In this work, we propose to combine task-oriented dialogue with knowledge-grounded chit-chat, and construct a new dataset named KETOD, with manually composed knowledge-enriched system responses. We conduct comprehensive experiments on our new dataset to study the insights and challenges. We believe that our proposed task is an important step towards the ultimate goal to build a unified, human-like conversational AI. Our new dataset KETOD, annotated by experts, will greatly facilitate the research in this direction.

7 Ethical Considerations

Data Access and Licensing. We develop the KETOD dataset based on the publicly available SGD dataset[1] (Rastogi et al., 2020). The SGD dataset is publicly available under the CC-BY-SA-4.0 License.

Dataset Collection Process and Conditions. This project is approved by our Institutional Review Board (IRB). Our annotators are all U.S. based. For the annotation of our KETOD dataset, linguistics for assessing data quality, and all the human evaluations, our annotators were hired as full-time employees through a leading annotation services vendor, and were paid in accordance with a fair wage rate. During the data annotation, we instruct the annotators to skip any example that contains offensive or any unethical contents.

References

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 5016–5026. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1870–1879. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Kumar Goyal, Peter Ku, and Dilek Hakkani-Tür. 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, LREC 2020, Marseille, France, May 11-16, 2020*, pages 422–428. European Language Resources Association.

Silin Gao, Ryuichi Takanobu, Wei Peng, Qun Liu, and Minlie Huang. 2021. Hyknow: End-to-end task-oriented dialog modeling with hybrid knowledge management. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1591–1602. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5110–5117. AAAI Press.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1891–1895. ISCA.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020a. A simple language model for task-oriented dialogue. *CoRR*, abs/2005.00796.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020b. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tür. 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, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, pages 278–289. Association for Computational Linguistics.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tur. 2021. "how robust r u?": Evaluating task-oriented dialogue systems on spoken conversations. *CoRR*, abs/2109.13489.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. *CoRR*, abs/1909.03087.

Bing Liu, Gökhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry P. Heck. 2017. End-to-end optimization of task-oriented dialogue model with deep reinforcement learning. *CoRR*, abs/1711.10712.

Bing Liu, Gökhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry P. Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2060–2069. Association for Computational Linguistics.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 845–854. Association for Computational Linguistics.

Nikola Mrksic, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve J. Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1777–1788. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. SOLOIST: few-shot task-oriented dialog with A single pre-trained auto-regressive model. *CoRR*, abs/2005.05298.

Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2231–2240. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8689–8696. AAAI Press.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3784–3803. Association for Computational Linguistics.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 196–205. The Association for Computational Linguistics.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2021. Multi-task pre-training for plug-and-play task-oriented dialogue system. *CoRR*, abs/2109.14739.

Kai Sun, Seungwhan Moon, Paul A. Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chat to enhance task-oriented dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1570–1583. Association for Computational Linguistics.

Chongyang Tao, Changyu Chen, Jiazhan Feng, Ji-Rong Wen, and Rui Yan. 2021. A pre-training strategy for zero-resource response selection in knowledge-grounded conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4446–4457. Association for Computational Linguistics.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1855–1865. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Peihao Su, David Vandyke, and Steve J. Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1711–1721. The Association for Computational Linguistics.

Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve J. Young. 2017a. Latent intention dialogue models. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3732–3741. PMLR.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve J. Young. 2017b. 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, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 438–449. Association for Computational Linguistics.

Jason D. Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue Discourse*, 7(3):4–33.

Haotian Xu, Haiyun Peng, Haoran Xie, Erik Cambria, Liuyang Zhou, and Weiguo Zheng. 2020. End-to-end latent-variable task-oriented dialogue system with exact log-likelihood optimization. *World Wide Web*, 23(3):1989–2002.

Tom Young, Frank Z. Xing, Vlad Pandelea, Jinjie Ni, and Erik Cambria. 2021. Fusing task-oriented and open-domain dialogues in conversational agents. *CoRR*, abs/2109.04137.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1458–1467. Association for Computational Linguistics.

Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. Mechanism-aware neural machine for dialogue response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3400–3407. AAAI Press.

Appendix A: Dataset Construction

Figure 6 shows our annotation interface to add knowledge-grounded chit-chat to TOD. The left part shows the full dialogue, where the annotators can click and expand each turn to make the chit-chat enrichment. The right part shows all the entities with the associated knowledge snippets. The annotators can click on each entity name to expand the textbox to see the knowledge snippets. We add index number to each knowledge snippet (shown in green brackets), and the annotators are asked to write down the indexes of the knowledge snippets they used for writing the knowledge grounded chit-chat. Figure 7 shows one example annotation turn using our interface.

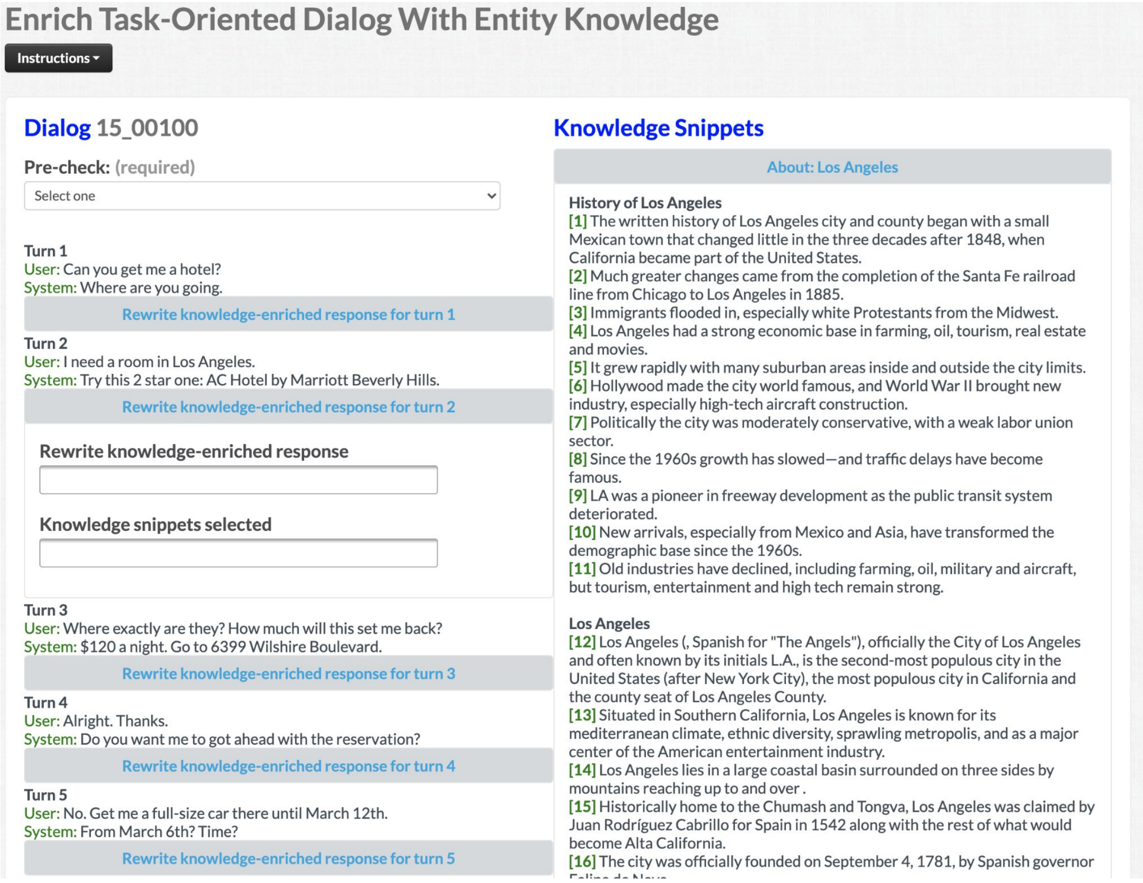

Figure 6: Our annotation interface example 1.

Appendix B: Model and Training Details

All the implementations are based on the Huggingface Transformers library[2]. For all models, we use the Adam optimizer (Kingma and Ba, 2015). For the knowledge selection model, we use BERT-base with learning rate of 3e-5 and batch size of 16. For the baseline SimpleToD model, SimpleToDPlus model, and Combiner model, we all use learning rate of 1e-4 and batch size of 16. All the experiments are done using TESLA M40 GPU cards.

Appendix C: Evaluation Details

Table 9 and 10 show the human evaluation results of SimpleToDPlus vs. SimpleToD, and Combiner vs. SimpleToD, respectively.

Metrics	SimpleToDPlus win (%)	SimpleToD win (%)	Tied (%)
Engagingness	40.0	30.3	29.8
Interestingness	31.8	19.5	48.8
Knowledge	38.0	18.3	43.8
Humanness	38.3	26.8	35.0

Table 9: Human evaluation of SimpleToDPlus vs. SimpleToD.

Metrics	Combiner win (%)	SimpleToD win (%)	Tied (%)
Engagingness	34.8	33.5	31.8
Interestingness	27.0	22.5	50.5
Knowledge	32.5	23.0	44.5
Humanness	27.8	32.5	39.8

Table 10: Human evaluation of Combiner vs. SimpleToD.

[2] https://github.com/huggingface/transformers

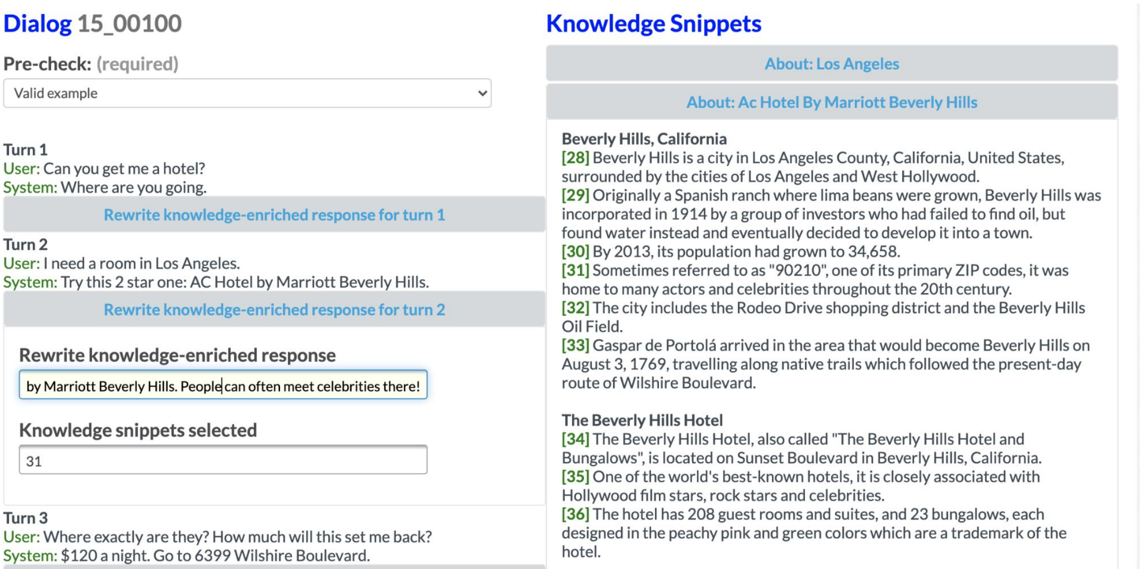

Figure 7: Our annotation interface example 2.