"What do others think?": Task-Oriented Conversational Modeling with Subjective Knowledge

Chao Zhao [1] **Spandana Gella** [2] **Seokhwan Kim** [2] **Di Jin** [2]
Devamanyu Hazarika [2] **Alexandros Papangelis** [2] **Behnam Hedayatnia** [2]
Mahdi Namazifar [2] **Yang Liu** [2] **Dilek Hakkani-Tur** [2]
zhaochao@cs.unc.edu {sgella,seokhwk,djinamzn}@amazon.com
{dvhaz,papangea,behnam,mahdinam,yangliud,hakkanit}@amazon.com
[1] UNC Chapel Hill [2] Amazon, Alexa

Abstract

Task-oriented Dialogue (TOD) Systems aim to build dialogue systems that assist users in accomplishing specific goals, such as booking a hotel or a restaurant. Traditional TODs rely on domain-specific APIs/DBs or external factual knowledge to generate responses, which cannot accommodate subjective user requests (e.g.,"*Is the WIFI reliable?*" or "*Does the restaurant have a good atmosphere?*"). To address this issue, we propose a novel task of subjective-knowledge-based TOD (SK-TOD). We also propose the first corresponding dataset, which contains subjective knowledge-seeking dialogue contexts and manually annotated responses grounded in subjective knowledge sources. When evaluated with existing TOD approaches, we find that this task poses new challenges such as aggregating diverse opinions from multiple knowledge snippets. We hope this task and dataset can promote further research on TOD and subjective content understanding. The code and the dataset are available at https://github.com/alexa/dstc11-track5.

1 Introduction

Task-oriented Dialogue (TOD) Systems aim to build dialogue systems that assist users in accomplishing specific goals, such as booking a hotel or a restaurant. Most solutions of TOD are 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 a limited range of scenarios within the scope of APIs/DBs. To further enlarge the model's ability of task-oriented assistance, recent works (Dimitrakis et al., 2018; Kim et al., 2020, 2021; Feng et al., 2020, 2021; Majumder et al., 2022) incorporate unstructured textual information retrieved from the Internet into dialogue modeling. Most of these works focus on factual knowledge sources such as frequently asked questions (FAQs) of online products or government service guides. We refer to these models as Fact-TOD models.

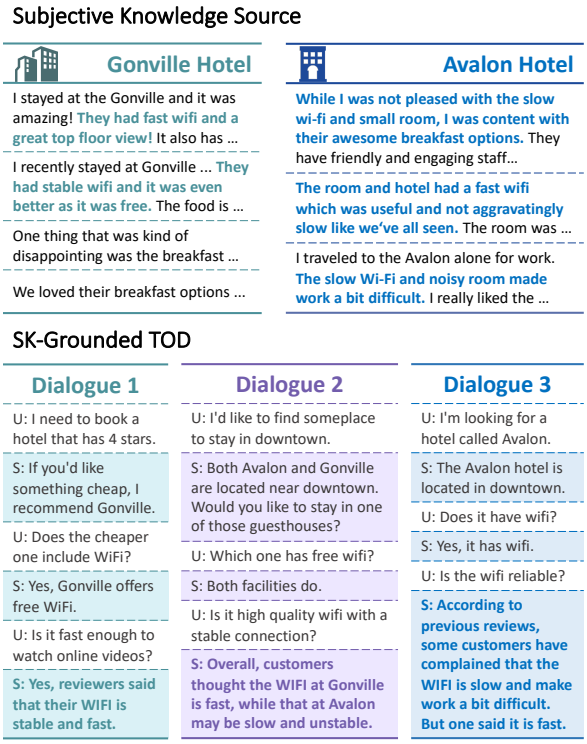

Figure 1: Examples of the SK-TOD task. The top part shows two hotels and their customer reviews. The bottom part shows three dialogue sessions between the system (denoted by S) and three users (denoted by U). The last user utterance is a subjective question about the WIFI quality of the hotel(s). The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

However, in many TOD tasks, users care about not only factual information but subjective insights as well, such as the experiences, opinions, and preferences of other customers. For instance, when booking a hotel or a restaurant, users often inquire about subject aspects 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. While subjective knowledge has been specifically studied in other NLP problems such as opinion mining (Liu and Zhang, 2012) and question answering (Bjerva et al., 2020), incorporating it into TOD has not received significant attention.

In this work, we argue that it is important to enable the TOD model to leverage subjective knowledge for more effective task-oriented assistance. To this end, we propose a novel task of subjective-knowledge-based task-oriented dialogue (SK-TOD). SK-TOD focuses on responding to user requests that seek subjective information by incorporating user reviews as subjective knowledge. Figure 1 shows three examples of such requests, where customers ask about the WiFi quality of various hotels. User reviews are valuable resources for subjective information because even for the same aspect of a product or service, customers may have different opinions and leave either positive or negative reviews. As a result, a TOD system should consider multiple reviews to provide a comprehensive representation of user opinions. Ideally, the system's response should include both positive and negative opinions, along with their respective proportions (as exemplified in Dialogue 3). This two-sided response has been recognized as more credible and valuable for customers (Kamins et al., 1989; Lee et al., 2008; Baek et al., 2012), thereby fostering trust in the TOD system.

Incorporating subjective knowledge into TOD introduces two unique challenges. Firstly, unlike in Fact-TOD where selecting a few relevant knowledge snippets suffices, the SK-TOD model must consider all relevant knowledge snippets. In other words, both precision and recall matter during this process. Secondly, the model needs to aggregate these knowledge snippets into a concise response that can faithfully reflect the diversity and proportion of opinions expressed. Conquering these challenges requires a large-scale dataset with subjective-knowledge-grounded responses, which, to our best knowledge, is not publicly available.

To facilitate the research in subjective-knowledge-grounded TOD, we have collected a large-scale dataset, 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). We evaluate the performance of strong baselines on the SK-TOD task. Results show that there is a significant gap between human-generated and machine-generated responses, particularly in terms of the faithfulness of the sentiment proportion. To address this issue, we propose a model that incorporates review understanding into SK-TOD. We experimentally demonstrate that responses generated by this model more effectively capture the sentiment proportion. Our contributions are three-fold:

- We introduce a novel task of subjective-knowledge-based TOD (SK-TOD);
- We create and release a large-scale, human-annotated dataset designed for this task;
- We propose a new model and conduct extensive experiments on the proposed task.

2 Related Work

2.1 Knowledge-Grounded Dialogue

Knowledge-grounded response generation is popular in the open-domain dialogue. Numerous external knowledge sources have been explored, from structured knowledge such as fact tables (Moghe et al., 2018; Liu et al., 2018) and knowledge graphs (Zhang et al., 2020a; Moon et al., 2019; Tuan et al., 2019), to unstructured knowledge such as Wikipedia articles (Vougiouklis et al., 2016; Zhou et al., 2018; Dinan et al., 2018), news articles (Majumder et al., 2020), web pages (Long et al., 2017; Galley et al., 2019; Komeili et al., 2022), narratives (Xu et al., 2021; Gopalakrishnan et al., 2019), user reviews and comments (Moghe et al., 2018; Ghazvininejad et al., 2018), and so on. Grounding on external knowledge makes the response more informative and meaningful when compared with models that solely rely on the dialog context.

Regarding task-oriented dialogues, previous works have primarily focused on domain-specific APIs and databases to support the dialogue response (Levin et al., 2000; Singh et al., 2002; Williams and Young, 2007; Eric et al., 2017; Wu et al., 2019), which can only support a limited scope of user queries. Later works ground task-oriented dialogues to web pages (Penha et al., 2019; Chen et al., 2022), government service documents (Saeidi et al., 2018; Feng et al., 2020, 2021), and FAQ knowledge snippets (Kim et al., 2020, 2021). Different from these works where factual knowledge is utilized, we apply subjective knowledge to generate the response and ground in multiple

knowledge snippets. While Majumder et al. (2022) also explored grounding TOD in user reviews, they did not consider the diversity of opinions.

2.2 Subjective Content Understanding

Besides being used as external knowledge sources in dialogue systems, subjective content, especially user reviews, has been studied in various non-conversational NLP tasks. For example, opinion mining (Pontiki et al., 2016; Jiang et al., 2019) focuses on extracting opinions and sentiments from user reviews. Opinion summarization (Chu and Liu, 2019; Zhao and Chaturvedi, 2020; Bražinskas et al., 2020; Angelidis et al., 2021) is used to distill multiple opinions into concise summaries. Subjective question answering (McAuley and Yang, 2016; Bjerva et al., 2020) have been proposed to answer questions based on user reviews. Explainable recommendation (Ni et al., 2019) aims to generate review-based explanations for the items recommended by a recommendation system. Table 1 provides detailed comparisons between SK-TOD and these subjective-content-based benchmarks. Generally, SK-TOD requires creating a response that is appropriate to the dialogue context. It also requires grounding in multiple subjective knowledge and explicitly considers the diversity of opinions and the proportion of sentiments.

3 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 for the last user utterance U_t. The dialogue involves one or more entities, denoted as $\mathcal{E} = \{e_1, \cdots, e_m\}$. Alongside the dialogue, we have a subjective knowledge source $\mathcal{B} = \{(e_1, \mathcal{R}_1), (e_2, \mathcal{R}_2), \cdots\}$ containing all the entities and their corresponding customer reviews. Each entity e is associated with multiple reviews $\mathcal{R} = \{R_1, R_2, \cdots\}$. Each review can be divided into segments $[K_1, K_2, \cdots]$, such as paragraphs, sentences, or sub-sentential units. In this work, we regard each review sentence as a knowledge snippet.

The SK-TOD task aims to identify whether U_t is a subjective knowledge-seeking request and, if it is, to select the relevant knowledge snippets $\mathcal{K}^+$ from the knowledge source and finally generate a response S_t grounded on $\mathcal{K}^+$.

4 Data Collection and Statistics

We ground the data collection in MultiWOZ (Budzianowski et al., 2018; Eric et al., 2020). We select dialogues from the domains of hotels and restaurants. The data collection is conducted by a group of crowd workers through Amazon Mechanical Turk (AMT). To control the data quality, we only choose workers that are pre-qualified. More details can be found in Appendix A.

4.1 Annotation Guideline

Dialogues in MultiWOZ are collected based on single or multiple entities as the back-end database. To create a subjective knowledge source to support the SK-TOD task, we first collect multiple user reviews for each entity. To control the review collection, we provide the reviewer's persona, as well as the aspects and sentiments of reviews to workers. We then ask workers to write a review with all the given information included. After collecting the reviews, we also annotate the aspect and sentiment information for each review sentence. Overall, we select 33 hotels and 110 restaurants from MultiWOZ, and collect 10 reviews for each entity. On average, each review contains 5.6 sentences and 56.71 tokens. More details about the review collection can be found in Appendix A.

After obtaining the reviews, we go back to the dialogue data to create the subjective user request. Following a similar procedure in Kim et al. (2020), for each dialogue, we provide an aspect that users are interested in (e.g., WIFI-quality of the hotel) and then ask the worker to insert a subjective user request into the dialogue. Workers are requested to carefully select the insertion position and write an utterance to maintain coherence and naturalness in the dialogue flow. Finally, we use the partial dialog until this newly inserted turn as an instance in our data. Utterances that come after the insertion position are removed from the dialogue instance.

So far, we've collected the dialogue context C and the subjective knowledge source $\mathcal{B}$. The final step is to ground the dialogue in the knowledge source. We first ask workers to identify entities that are relevant to the subjective user request as gold entities. We then align the user request and review sentences of the gold entities by matching their aspect. For example, if the aspect of a user request is about the "WIFI quality" of a hotel, all review sentences discussing the "WIFI quality" of that specific hotel will be considered relevant knowledge

	Size	Manual	Dial	TOD	Query	Aspect	Senti	Mul-Knwl	Senti-%
Semeval/MAMS (2016; 2019)	5K/22K	✓	✗	n/a	✗	✓	✓	✗	n/a
Space (2021)	1K	✓	✗	n/a	✗	✓	✓	✓	✗
Yelp/Amazon (2019; 2020)	200/180	✓	✗	n/a	✗	✗	✓	✓	✗
Justify-Rec (2019)	1.3M	✗	✗	n/a	✗	✓	✗	✓	✗
AmazonQA (2016)	309K	✗	✗	n/a	✓	✗	✗	✗	n/a
SubjQA (2020)	10K	✗	✗	n/a	✓	✓	✓	✗	n/a
Holl-E (2018)	9K	✓	✓	✗	✗	✗	✗	✓	✗
Foursquare (2018)	1M	✗	✓	✗	✗	✗	✗	✓	n/a
SK-TOD (Ours)	20K	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison between SK-TOD and other benchmarks based on the subjective content. We consider if the dataset is manually annotated, dialogue-based, task-oriented, and query-focused. We also list if it considers aspect and sentiment, multiple knowledge snippets (Mul-Knwl), and the proportion of two-sided sentiments (Senti-%).

snippets. [1] Finally, we provide the dialogue context C and all related knowledge snippets $\mathcal{K}^+$ and ask workers to generate a natural and faithful response. We explicitly instruct workers to consider the diversity and proportion of opinions in all relevant knowledge snippets during response creation. Detailed instructions can be found in Appendix A.

4.2 Quality Control

To ensure the quality of our dataset, we took great care in selecting pre-qualified workers and designing annotation interfaces. We further conducted a human verification task on the entire dataset to identify invalid instances. The annotation showed that 81.89% of subjective-knowledge-seeking user turns are valid, with an Inter-Annotator Agreement (IAA) score of 0.9369 in Gwet's gamma. For agent response turns, 96.78% were valid, with an IAA score of 0.9497 in Gwet's gamma. Any invalid instances were filtered out or manually corrected before finalizing the dataset. We paid workers an average of $13.82/hr for data annotation and $14.77/hr for data verification. Both exceed the local living minimum wage. The details of our payment settings are elaborated on in Appendix A.

4.3 Data Statistics

We collected a total of 19,696 instances consisting of subjective user requests and subjective-knowledge-grounded responses. 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). On average, each instance requires 3.88 subjective knowledge snippets. To help identify the subjective knowledge-seeking user request, we also randomly sample another 18,383 dialogues with non-subjective user requests from the original MultiWOZ dataset.

[1]Note that the aspect information is only used to build the dataset but is not included in the problem formulation of SK-TOD, which means it is not available for model training. The goal of SK-TOD is to handle user requests with arbitrary aspects, and therefore we do not define a taxonomy of aspects in the task like what is done in dialogue state tracking.

	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 2: Basic statistics of our dataset.

We split the dataset into training (75%), validation (10.8%), and test (14.2%) sets. Table 2 presents the detailed statistics of each subset. Both the 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' ability to generalize to arbitrary aspects.

5 Subjective-Knowledge-Grounded TOD

In this section, we describe the method for SK-TOD. As shown in Figure 2, we follow the pipeline introduced by Kim et al. (2020) which comprises four sequential sub-tasks: knowledge-seeking turn detection (KTD), entity tracking (ET), knowledge selection (KS), and response generation (RG). We elaborate on each subtask below.

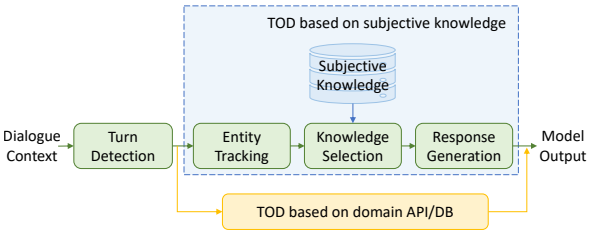

Figure 2: The pipeline architecture of SK-TOD.

5.1 Knowledge-Seeking Turn Detection

The goal of KTD is to identify the user request that requires 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.

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 seeking subjective knowledge. 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.

5.2 Entity Tracking

The goal of ET is to identify the entities $\mathcal{E} = \{e_1, \cdots, e_m\}$ that are relevant to the user request. It can help to reduce the number of candidates during the knowledge selection step.

We adopt a word-matching-based method used by Jin et al. (2021) to extract relevant entities. It first normalizes entity names in the knowledge source using a set of heuristic rules. Then a fuzzy n-gram matching is performed between the normalized entity and all dialogue turns. To find the entities that are relevant to the last user request, we choose the last dialogue turn in which the entities are detected and use these entities as the output of ET. We leave the tracking of aspects being questioned over multiple turns as future work.

5.3 Knowledge Selection

The goal of KS is to select the knowledge snippets that are relevant to the user's request. The inputs are the dialogue context C and a set of knowledge snippets candidates $\mathcal{K}$, which is a combination of all 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 $\mathcal{K}^+$.

To select relevant knowledge snippets, we calculate the relevance score between the dialogue context C and a knowledge snippet $K \in \mathcal{K}$. We regard it as a pairwise text scoring problem and consider two popular approaches: bi-encoder (Mazaré et al., 2018) and cross-encoder (Wolf et al., 2019). Generally, the bi-encoder approach is more efficient while the cross-encoder approach is more accurate.

For the bi-encoder approach, we encode C and K separately using the same pre-trained encoder and obtain two representations, h_C and h_K. Following Reimers and Gurevych (2019), we use the concatenation of h_C, h_K, and $|h_C - h_K|$ as features and apply a classifier to obtain the probability of relevance. That is,

$$\begin{aligned} h_C = \text{Enc}(C), \quad h_K = \text{Enc}(K) \\ P(C, K) = \text{softmax}\left(\text{FFN}\left(h_c, h_K, |h_C - h_K|\right)\right). \end{aligned} \tag{2}$$

For the cross-encoder approach, we encode the concatenation of C and K to obtain a contextualized representation. That is,

$$\begin{aligned} h &= \text{Enc}(C, K) \\ P(C, K) &= \text{softmax}\left(\text{FFN}\left(h\right)\right). \end{aligned} \tag{3}$$

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

5.4 Response Generation

The goal of RG is to create an utterance S_t that addresses the user's request. This response is generated based on the dialogue context C and the set of relevant knowledge snippets $\mathcal{K}^+$. To accomplish this, we concatenate $\mathcal{K}^+$ and C as the input and use a pre-trained generation model to generate the response. We consider both the decoder-only model, such as GPT-2 (Radford et al.), and the encoder-decoder model, such as BART (Lewis et al., 2020).

The model is trained to maximize the generation probability $p(S_T \mid C, \mathcal{K}^+)$.

To accurately capture the diversity and proportion of opinions, the model needs to understand the sentiment polarity of each knowledge snippet, which is challenging due to the lack of direct supervision. To address this issue, we apply a state-of-the-art aspect-based sentiment analysis (ABSA) model (Zhang et al., 2021) to predict the sentiment $Z = [z_1, \cdots, z_i, \cdots]$ for each knowledge snippet $K_i \in \mathcal{K}^+$. Then we incorporate the sentiment information into RG by maximizing $p(S_T \mid C, \mathcal{K}^+, Z)$.

More specifically, we first convert the predicted z_i into a natural language description using templates, and then append it to the end of the corresponding K_i as the enhanced input of RG. For example, given the knowledge snippet as "*The ambience was so fun.*", the ABSA model detects the aspect-based sentiment as ("ambience", "positive"). We first convert the sentiment into a natural language "*ambience is great.*" and then enhance the knowledge snippet as "*The ambience was so fun. ambience is great.*". We refer to Appendix B for more details.

6 Experiments on Sub-Tasks

We first conduct experiments on each individual subtask. To avoid any error accumulation from upstream tasks, we use the gold output of the previous task as the input to the current target task. The detailed experimental setup can be found in Appendix C.

6.1 Knowledge-Seeking Turn Detection

Setting We conduct experiments using various pre-trained language models, including BERT [2] (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), and DeBERTa (He et al., 2021).

Evaluation We report the precision, recall, F_1 score, and accuracy score.

Results Table 3 shows the results of the KTD task. All models achieve similar and near-perfect performance, which is in line with the findings of Kim et al. (2020). It demonstrates that it is feasible to identify the user requests that require subjective knowledge, allowing them to be explicitly addressed by an SK-TOD component. However, this KTD classifier's performance may be specific to this dataset or similar domains, and its generalizability to unseen domains or knowledge types requires further exploration in future works.

[2]We use the base version of all pre-trained models.

	Acc	P	R	F
BERT	99.67	99.75	99.61	99.68
RoBERTa	99.74	99.86	99.64	99.75
ALBERT	99.49	99.64	99.36	99.50
DeBERTa	99.71	99.86	99.57	99.71

Table 3: Results of KTD task. Models are evaluated using Accuracy, Precision, Recall, and F_1. All models achieve similar and near-perfect performance.

6.2 Entity Tracking

Setting We follow the setting of Jin et al. (2021) to run the ET method.

Evaluation We report the instance-level accuracy score. An instance is regarded as accurate only if the predicted entities match exactly with the gold entities.

Results The fuzzy n-gram matching method achieves an instance-level accuracy of 92.18%. We further analyzed the type of errors. For 1.8% of the instances, there is at least one gold entity missing from the predicted entities. For 7.6% of the instances, the predicted entities contain at least one spurious entity. The latter error case can be further reduced by using model-based matching approaches, which we leave as future work.

6.3 Knowledge Selection

Setting We fine-tune the KS models following the same setting as in the KTD task. Additionally, we compare them with traditional information retrieval (IR) baselines, such as TF-IDF (Manning et al., 2008) and BM25 (Robertson et al., 2009).

Evaluation Knowledge selection can be viewed as either a classification task or a retrieval task. For classification, we use precision, recall, and F_1 measures. We calculate these measures at both the instance level and the snippet level. For the instance level, we first calculate $P/R/F_1$ for each instance, and then take the average over all instances as the final scores. For the snippet level, instead of computing $P/R/F_1$ for each instance, we calculate these scores for all <C, K> pairs in the entire dataset. Regarding retrieval evaluation, we use mean-average-precision (mAP) as the metric, which is not dependent on a specific threshold value

	Instance-level			Snippet-level			mAP
	P	R	F	P	R	F	
IR Baselines							
TF-IDF	34.61	70.33	40.46	23.81	65.00	34.85	45.97
BM25	31.38	40.95	32.21	31.14	32.42	31.77	45.42
Bi-encoder							
BERT	56.66	70.06	59.31	58.87	74.69	65.84	71.59
RoBERTa	60.98	83.06	66.47	54.40	**85.38**	66.46	77.25
ALBERT	70.21	78.74	70.43	**63.13**	78.90	70.14	81.62
DeBERTa	**71.46**	**83.18**	**72.44**	62.64	83.50	**71.58**	**83.43**
Cross-encoder							
BERT	85.18	86.01	83.33	82.40	83.82	83.11	90.06
RoBERTa	81.59	83.62	80.53	82.20	80.77	81.48	88.98
ALBERT	**86.18**	87.29	84.22	**83.56**	84.78	84.16	90.50
DeBERTa	86.07	**87.64**	**84.6**	82.70	85.71	**84.18**	**91.84**
SEEN	88.80	93.45	89.93	90.83	89.90	90.37	95.70
UNSEEN	82.68	80.47	78.03	69.98	78.29	73.90	87.07

Table 4: Results of the KS task. Models are evaluated using instance-level and snippet-level classification measures, as well as mAP, a retrieval-based measure. DeBERTa achieves the best performance among all evaluation measures.

and can reflect the overall ranking positions of all relevant knowledge snippets. Since the total number of the relevant knowledge snippets can vary for each instance, we do not include top-K-based measures like Precision@K or Recall@K, which are commonly used in other Fact-TOD and knowledge-grounded open-domain dialogue tasks.

Results Table 4 shows the results of the KS task. Firstly, when comparing our models with IR baselines, all of the trained models outperform the baselines, indicating that the KS model can benefit from the annotated training data. We then compare bi-encoder models and cross-encoder models, and as expected, cross-encoder models outperform bi-encoder models by a large margin. When comparing the performance of different pre-trained models, there is a notable difference among the models under the bi-encoder setting. The variance becomes smaller when applying the cross-encoder architecture. DeBERTa achieves the best performance on all measures in both the bi-encoder and cross-encoder settings.

Finally, we compare the performance between the seen subset and the unseen subset. At the bottom of Table 4, we list the performance of DeBERTa on both the seen and unseen test subsets. The results reveal a large gap between the performance of the two subsets, indicating that one of the challenges for the KS model is to generalize from seen aspects to unseen aspects.

	BLEU	R-1	R-2	R-L	MT	BS	Len
EXT	2.89	23.17	6.53	18.33	9.62	30.83	14.93
GPT2	9.04	33.9	13.52	26.73	16.27	39.73	22.66
DialoGPT	9.19	33.6	13.62	26.81	16.15	39.72	22.05
BART	10.8	36.35	15.04	28.57	17.96	41.12	24.02
$\text{BART}_{\text{ABSA}}$	10.78	36.30	15.36	28.47	18.06	**41.75**	23.66
T5	10.72	36.50	**15.57**	28.81	**18.33**	40.84	25.36
T5_{ABSA}	**10.97**	**36.66**	15.51	**28.88**	18.15	40.94	24.75

Table 5: Results of RG task. Models are evaluated using BLEU, ROUGE (R-1, R-2, R-L), METEOR (MT), and BertScore (BS). We also listed the average length (Len) of the generated response. Encoder-decoder models such as BART and T5 achieve better performance compared with GPT2-based models.

6.4 Response Generation

Setting we experiment with decoder-only generation models such as GPT-2 (Radford et al.) [3] and DialoGPT (Zhang et al., 2020c), as well as encoder-decoder models such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). We also include two ABSA-enhanced models, namely $\text{BART}_{\text{ABSA}}$ and T5_{ABSA}. During decoding, we use beam-search with top-K sampling (Fan et al., 2018). We set the beam size as 5 and sample from the top 50 tokens. We also compare with a random extractive baseline (EXT), where the response is created by randomly selecting a relevant knowledge snippet.

Evaluation Following the evaluation of other generation tasks, We employ several automatic evaluation metrics, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), as well as BERTScore (Zhang et al., 2020b), to evaluate the quality of the generated responses compared to the reference responses. We also conduct a human evaluation, where we ask crowd workers to evaluate the quality of responses.

Results As presented in Table 5, machine-generated responses significantly outperform the extractive responses. Encoder-decoder models achieve better performance across all automatic measures compared to GPT-based models, indicating that they are more suitable for this task. They

[3] We use the base-version of all pre-trained models.

also tend to generate longer responses. There is no clear difference in automatic measures when comparing BART and T5. For ABSA-enhanced models, $\text{BART}_{\text{ABSA}}$ achieves the best performance on BertScore, while T5_{ABSA} achieves the best score on BLEU and ROUGE.

Human Evaluation To obtain a more reliable assessment of response quality, we also conduct a human evaluation on AMT. We use the same group of workers involved in the data collection process. During the evaluation, we show the dialogue context, the oracle knowledge snippets, and all responses (both the reference and the generated responses) to the workers. We randomly sample 240 instances from the test set for evaluation. For each instance, we ask three independent workers to compare the responses based on three measures:

- Appropriateness: whether the response is fluent and naturally connected to the dialogue context.
- Aspect Accuracy: whether the response provides relevant and useful information to the aspect that the user queried.
- Sentiment Accuracy: whether the sentiment proportion provided by the response is consistent with that of the subjective knowledge.

For sentiment accuracy, we first ask workers to annotate the sentiment label of each knowledge snippet, and then evaluate each response. All three measures are evaluated using a 5-Point Likert scale. The system-level score is computed as the average score over all instances and workers for each system. The compensation for workers was set at \$0.25 for the tasks of appropriateness and aspect accuracy, and \$0.4 for the task of sentiment accuracy. The average hourly pay for the crowd workers was \$15.25/hr, \$14.40/hr, and \$14.85/hr for each evaluation task, exceeding the local living minimum wage.

Table 6 shows the results of human evaluation for response generation. The inter-annotator agreement scores for each task are 0.7270, 0.7535, and 0.6239 in Gwet's gamma, respectively. The results show that machine-generated responses are comparable to the references in terms of appropriateness and aspect accuracy. Moreover, incorporating ABSA can improve the model's performance in sentiment accuracy. However, there is still a large gap in sentiment accuracy between the best model-generated responses and the references, indicating that faithfully aggregating sentiment information from multiple knowledge snippets is still a challenging task for current models.

	Approp.	**Asp-Acc**	**Senti-Acc**
EXT	2.65	3.32	3.13
GPT2	4.55	4.54	3.20
BART	4.55	4.68	3.56
$\text{BART}_{\text{ABSA}}$	4.58	4.66	3.80
T5	4.40	4.63	3.87
T5_{ABSA}	4.49	4.67	3.98
Reference	4.70	4.77	4.50

Table 6: Results of human evaluation for RG.

Dialogue History:
U: Are there any hotels (not guesthouses) that have 0 stars?
S: Yes, actually. Cityroomz is a moderately priced 0 star hotel.
U: Do they have free wifi?
S: Yes, they do have free wifi! Would you like me to book a stay for you?
U: Does the Cityroomz have strong water pressure in the shower?

Reviews:
Review 1: However, I noticed that when taking a shower there was very little water pressure which made the shower disappointing.
Review 2: The water pressure is not good and it comes out really slow.
Review 3: Everything else was a disaster, especially the lack of cleanliness in the room when we checked in, and the incredibly low water pressure in the bathroom when we took our morning showers.
Review 4: I encounter a problem when trying to shower and washing hands.
Review 5: There is no luggage storage service or anything like that, and the water pressure is much to be desired.

Model Outputs:
GPT-2: The Cityroomz gets some mixed reviews on their water pressure. Some guests found the water pressure to be very good while others found it to be disappointing. Would you like to know more about them?
BART: The Cityroomz hotel has very low water pressure according to most guests. Only one person had a good experience with it.
$\text{BART}_{\text{ABSA}}$: The Cityroomz has a lot of complaints about their water pressure being a bit low. Would you like me to find another place with better ratings?
T5: The Cityroomz has mixed reviews about their water pressure. Some guests found the water pressure to be very low while others found it to be disappointing. Would you like to know more about them?
T5_{ABSA}: The Cityroomz has a lot of reviews that say the water pressure is very low and disappointing. Do you want to look at some other places?

Reference:
No, guests consistently complain about the water pressure, unfortunately. Will that be okay or should I do another search?

Table 7: Sampled output of different RG models.

Qualitative Analysis Table 7 shows an example of responses generated by various systems. In this example, all the reviews express negative opinions about water pressure. However, responses generated by GPT-2 and BART include positive opinions. T5 correctly mentions the negative opinions but the generated response is not natural and coherent. By incorporating the ABSA model, both BART and T5 correctly generate responses with all negative opinions.

7 Experiments on End-2-End Evaluation

In the previous section, we use gold information as input for each module to avoid error accumulation.

	KS		RG		
	Macro-F	**mAP**	**BLEU**	**R-L**	**BS**
RG	-	-	10.80	28.52	41.12
+KS	84.60	91.84	10.20	27.78	40.64
+ET+KS	83.47	90.45	10.29	27.80	40.56
+KTD+ET+KS	83.46	90.45	10.27	27.79	40.55

Table 8: Results of the end-to-end evaluation. We start from RG with gold knowledge as input. We then gradually add components (KS, ET, and KTD) to the pipeline to replace the gold input with the predicted one.

	KTD	KS		RG		
	Acc	**Macro-F**	**mAP**	**BLEU**	**R-L**	**BS**
Fact-TOD	87.62	59.55	76.69	6.15	23.25	33.16
SK-TOD	99.71	84.60	91.84	10.80	28.57	41.12

Table 9: Comparison between models trained on Fact-TOD and SK-TOD training data.

Dialogue History:
U: I am looking for a guesthouse to stay at that has free wifi.
S: Ok. Were you hoping for a particular price range or part of town?
U: I don't care about the part of town or the price range, but it needs to be a three star hotel.
S: The Hamilton Lodge would be perfect! It is a moderately priced 3 star guesthouse on the north side of town. Does that sound okay to you?
U: Does it have comfortable beds?

Reviews:
Review 1: The bed really needed a new mattress, too.
Review 2: That being said, the room was spacious and the bed was very comfortable.
Review 3: We both slept extremely well every night we were there.

Model Outputs:
Fact-TOD: Yes, the Hamilton Lodge has a new mattress. Would you like to make a reservation?
SK-TOD: The Hamilton Lodge has really comfortable beds according to most guests, but one guest did say that the bed needed a new mattress.

Reference:
There are some mixed reviews on the beds. Some say they're very comfortable while others were not impressed. Would you like me to find another place with better reviews?

Table 10: Sampled outputs from the Fact-TOD model and the SK-TOD model, respectively.

In this section, we evaluate the entire pipeline in an end-to-end manner, where the input of each subtask is predicted by the previous component. We gradually add KS, ET, and KTD to the pipeline, and list the performance of KS and RG in Table 8.

The results show that errors introduced during KS can decrease the quality of response generation. However, ET and KTD do not have a significant impact on the performance of downstream tasks. It is because ET and KTD results include fewer noisy predictions compared to the KS results.

8 Comparison with Fact-TOD

One difference between SK-TOD and Fact-TOD is that responses in SK-TOD are grounded on subjective knowledge instead of factual knowledge. In this section, we investigate whether a Fact-TOD model can ground on subjective knowledge to address subjective requests. To this end, we retrain our KTD (DeBERTa), KS (DeBERTa cross-encoder), and RG (BART) models using the FAQ-grounded TOD data provided by Kim et al. (2020). We then apply it to the test set of SK-TOD without further training. We compare the results of each sub-task using the Fact-TOD models with the results of models trained on SK-TOD.

As shown in Table 9, for all tasks, there is a significant performance gap between models trained on Fact-TOD and on SK-TOD. We also provide sampled outputs by the Fact-TOD model and the SK-TOD model in Table 10. By checking the model output, we observe that the Fact-TOD model tends to only ground on and copy information from a single knowledge snippet. This behavior indicates that it is difficult to apply the Fact-TOD model to the SK-TOD task directly, as the Fact-TOD model lacks the ability to effectively aggregate information from multiple knowledge snippets, especially when there are diverse and contradictory opinions. The results also highlight that compared to Fact-TOD, SK-TOD faces new challenges in terms of subjective content understanding and dialogue modeling when integrating subjective knowledge into the responses.

9 Conclusion

In this paper, we have introduced SK-TOD: a novel task focused on subjective-knowledge-based task-oriented dialogue response generation. We create and release a large-scale, manually-annotated dataset for this task. Incorporating subjective knowledge requires models to accurately identify all relevant knowledge snippets and faithfully aggregate the information into concise and contextually appropriate responses, which brings unique challenges to this task. Experiments with strong baselines show that there is a significant performance gap between human-generated and machine-generated responses, particularly in faithfully capturing the diversity and proportion of opinions present in the subjective knowledge. We hope this task together with the provided dataset can promote future research on knowledge-grounded TOD systems and subjective content understanding.

Limitations

The dataset we collected contains two domains, restaurants and hotels. However, to evaluate the model's ability to generalize across different domains, it would be beneficial to include more domains in the dataset. Additionally, to address privacy and copyright concerns, we used crowdsourcing to collect review data, resulting in fewer and shorter reviews than those found in real-world scenarios. This limitation can be mitigated by sampling informative and reliable reviews from real-world data. Regarding the model, we did not investigate more complex models, such as large language models and novel architectures. However, we provide a strong baseline method that will serve as a benchmark for more advanced methods by the research community.

Ethical Considerations

To build our dataset, we collected the dialogue data by augmenting MultiWOZ 2.1, which is a publicly available English dialogue dataset under MIT license. Additionally, we collected the review data using crowd-sourcing, where we provided crowd workers with the reviewer's persona, as well as the aspects and sentiments of reviews. This controlled review collection process helps to exclude offensive or harmful content from the reviews. It also helps to avoid privacy or copyright issues when making the dataset publicly available. Our dataset is available under the CDLA-Sharing 1.0 license.

References

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

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.

Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5480–5494, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169.

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.

Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul Crook, and William Yang Wang. 2022. KETOD: Knowledge-enriched task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2581–2593, Seattle, United States. Association for Computational Linguistics.

Eric Chu and Peter Liu. 2019. Meansum: a neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR.

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.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

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.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at dstc7. In *AAAI Dialog System Technology Challenges Workshop*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, William B. Dolan, Jianfeng Gao, Wen tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *AAAI*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. 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)*, pages 6280–6285, Hong Kong, China. 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 Conversational 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, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tür. 2021. "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.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

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.

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on speech and audio processing*, 8(1):11–23.

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*, pages 7871–7880.

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.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, Melbourne, Australia. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yinong Long, Jianan Wang, Zhen Xu, Zongsheng Wang, Baoxun Wang, and Zhuoran Wang. 2017. A knowledge enhanced generative conversational service agent. In *Proceedings of the 6th Dialog System Technology Challenges (DSTC6) Workshop*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141, Online. Association for Computational Linguistics.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, pages 625–635.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2322–2332.

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 Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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)*, pages 188–197.

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.

Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. Introducing mantis: a novel multi-domain information seeking dialogues dataset. *arXiv preprint arXiv:1912.04639*.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Satinder Singh, Diane Litman, Michael Kearns, and Marilyn Walker. 2002. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *Journal of Artificial Intelligence Research*, 16:105–133.

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)*, pages 1855–1865, Hong Kong, China. Association for Computational Linguistics.

Pavlos Vougiouklis, Jonathon Hare, and Elena Simperl. 2016. A neural network approach for knowledge-driven response generation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3370–3380, Osaka, Japan. The COLING 2016 Organizing Committee.

Jason D Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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*.

Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. Enhancing dialog coherence with event graph grounded content planning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3941–3947.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020a. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020c. 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*, pages 270–278.

Chao Zhao and Snigdha Chaturvedi. 2020. Weakly-supervised opinion summarization by leveraging external information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9644–9651.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

A Data Collection

In this section, we describe more details of the data collection process. The data collection is conducted by a group of crowd workers through Amazon Mechanical Turk. To control the data quality, we choose English speakers from the US, CA, and GB. Workers are eligible for the annotation only if they pass our pre-qualification tests. During data collection, we also manually validate the annotation quality in several rounds to filter out the workers with low-quality annotations.

During review collection, we provide the reviewer's persona, as well as the aspects and sentiments of reviews to workers. The persona is randomly sampled from a pre-defined set of personas. For the aspects and sentiments, we first define 26 common aspects for hotel and restaurant reviews (e.g., WIFI-quality and room-bed for hotels, food-quality and indoor-decor for restaurants). We then randomly selected the target aspects to be addressed in a review. The number of aspects is randomly chosen. To mimic the sentiment distribution of the real reviews, the sentiment of each aspect is sampled based on the actual average ratings taken from Yelp. Figure 3 shows the interface of review collection. We pay workers $1.00 per task.

During user request collection, we ask workers to select the best position to insert a user request by considering every possible position of the given dialogue. Figure 4 shows the interface of user request collection. We pay workers $0.15 per task.

During response generation, we explicitly ask workers to consider the information in all snippets to create a natural and faithful response. Figure 5 shows the interface of response generation. We pay workers $0.25 per task. Below we list the complete instructions that we provide to workers.

- Please read ALL the customer reviews carefully.
- Please read the conversation carefully.
- Write down a response to the customer to answer the question and continue the conversation.
- You must read EVERY REVIEW COMMENT carefully. Each sentence was written by different people with potentially different opinions.
- Your response MUST include your SUMMARY of ALL the review sentences.

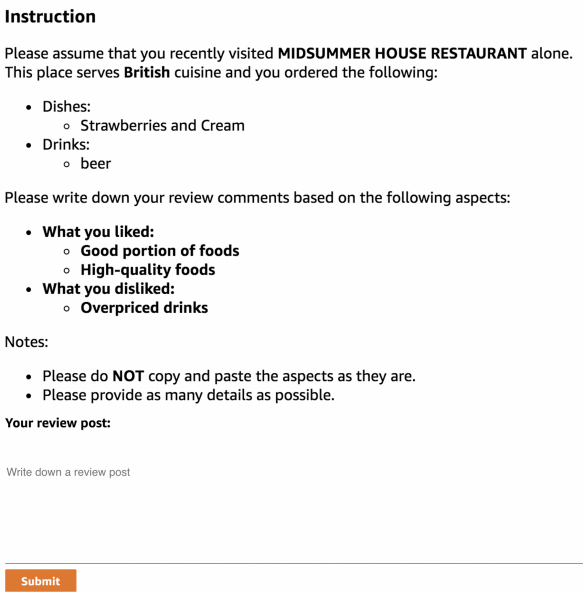

Figure 3: The interface of review collection.

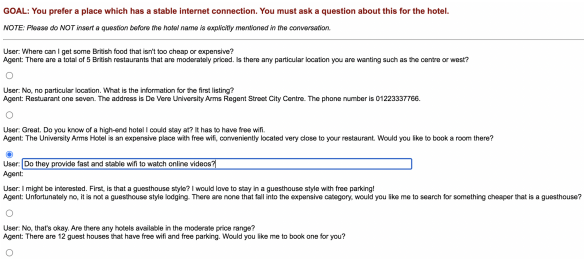

Figure 4: The interface of user request collection.

- If there's any conflict or different opinions in the reviews, your response MUST describe the minority opinion as well.
- Your response MUST be based on the contents in given review comments only.
- Please keep the way of speaking as similar as possible to the previous utterances spoken by the agent.

B Aspect Based Sentiment Analysis

To enhance the model's ability to understand the sentiment polarity of each individual knowledge snippet, we apply PGEN (Zhang et al., 2021), a state-of-the-art aspect-based sentiment analysis model, to predict the sentiment $Z = [z_1, z_2, \cdots, z_i, \cdots]$ for every knowledge snippet $[K_1, K_2, \cdots, K_i, \cdots]$ in $\mathcal{K}^+$.

PGEN converts the problem of aspect-based sentiment analysis into a sequence generation problem, where the input is the review sentence, and

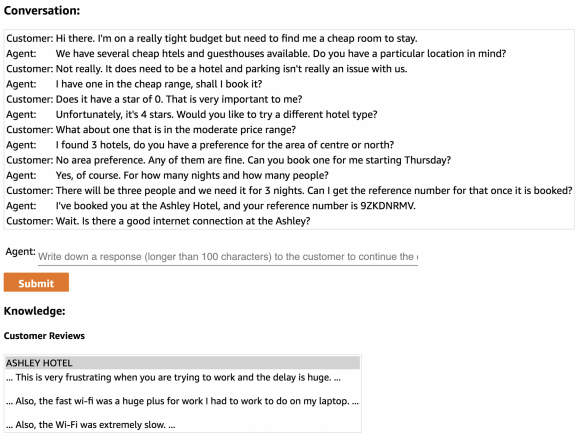

Figure 5: The interface of response generation.

the output is a natural language description of the aspect and the sentiment. For example, given the review sentence as "*The ambience was so fun.*", where the aspect term is "ambience" and the corresponding sentiment polarity is "positive", PGEN transform the aspect term and the sentiment polarity into a natural language description "ambience is great." using templates. It is transformed by keeping the aspect term unchanged and mapping the positive/neutral/negative sentiment polarities into one of the three tokens: "great", "ok", and "bad". The model is trained using a BART-base model on semeval aspect-based sentiment analysis datasets (Pontiki et al., 2015, 2016).

C Training Details

For KTD and KS, the implementation is based on Transformers (Wolf et al., 2020). During training, we use AdamW (Loshchilov and Hutter, 2018) with a learning rate of 3×10^{-5} and a batch size of 16. We apply warmup (Goyal et al., 2017) on the first 500 steps and early stopping based on the model performance on the validation set. We use a Tesla V100 GPU with 16 GB memory for training models. It takes 1 hour to train a KTD model and 5 hours to train a KS model.

During inference, we set the classification threshold as 0 for KTD, as we observe that KTD results are insensitive to the threshold. However, for the KS model, the setting of the threshold can greatly impact the precision and recall scores. We therefore choose the best threshold based on the F_1 scores on the validation set. We use a grid search between -5 to 5. The optimal thresholds for BERT, RoBERTa, ALBERT, and DeBERTa are 2.25, 1, 1.75, and 2 in the bi-encoder setting. They are 3.1, 4.6, 3.25, and 3.4 in the cross-encoder setting.

For ET model, we follow the setting of Jin et al. (2021) to identify entities. More specifically, we perform the fuzzy n-gram matching between an entity and the utterance, where n is the same as the length of the entity mention. The n-gram matching score is calculated based on the ratio of the longest common sequence between two n-grams. We set the matching threshold as 0.95.

For RG model, during training, we use AdamW with a learning rate of 3×10^{-5} and a batch size of 16. We apply the warmup on the first 500 steps and the early stopping based on the model performance (perplexity) on the development set. The model is trained on a Tesla V100 GPU with 16 GB memory for 2 hours.