

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

Anonymous ACL submission

Abstract

Task-oriented Dialogue (TOD) Systems aim to build dialogue systems to help users achieve specific goals (e.g., booking a hotel or a restaurant). Conventional TODs rely on domain-specific APIs/DBs or external factual knowledge to create 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 that are grounded in subjective knowledge sources. When evaluated with existing TOD approaches, we find that this task poses new challenges such as aggregating various 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 will be available at [url](#).

1 Introduction

Task-oriented Dialogue (TOD) Systems aim to build dialogue systems to help users to achieve specific goals (e.g., booking a hotel or a restaurant). Most solutions of TOD are based on domain-APIs (Budzianowski et al., 2018; Rastogi et al., 2020) and structured databases (Eric et al., 2017; Wu et al., 2019), which can only handle limited scenarios, restricted 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

Subjective Knowledge Source

 Gonville Hotel	 Avalon Hotel
I stayed at the Gonville and it was amazing! They had fast wifi and a great top floor view! It also has ...	While I was not pleased with the slow wi-fi and small room, I was content with their awesome breakfast options. They have friendly and engaging staff...
I recently stayed at Gonville ... They had stable wifi and it was even better as it was free. The food is ...	The room and hotel had a fast wifi which was useful and not aggravatingly slow like we’ve all seen. The room was ...
One thing that was kind of disappointing was the breakfast ...	I traveled to the Avalon alone for work. The slow Wi-Fi and noisy room made work a bit difficult. I really liked the ...
We loved their breakfast options ...	

SK-Grounded TOD

Dialogue 1	Dialogue 2	Dialogue 3
U: I need to book a hotel that has 4 stars.	U: I'd like to find someplace to stay in downtown.	U: I'm looking for a hotel called Avalon.
S: If you'd like something cheap, I recommend Gonville.	S: Both Avalon and Gonville are located near downtown. Would you like to stay in one of those guesthouses?	S: The Avalon hotel is located in downtown.
U: Does the cheaper one include WiFi?	U: Which one has free wifi?	U: Does it have wifi?
S: Yes, Gonville offers free WiFi.	S: Both facilities do.	S: Yes, it has wifi.
U: Is it fast enough to watch online videos?	U: Is it high quality wifi with a stable connection?	U: Is the wifi reliable?
S: Yes, reviewers said that their WIFI is stable and fast.	S: Overall, customers thought the WIFI at Gonville is fast, while that at Avalon may be slow and unstable.	S: According to previous reviews, some customers have complained that the WIFI is slow and make work a bit difficult. But one said it is fast.

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

service guides. We refer to these models as Fact-TOD models.

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

edge 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 is not studied.

In this work, we argue that it is important to enable the TOD model to leverage subjective knowledge during task-oriented assistance. To this end, we propose a novel task of subjective-knowledge-based task-oriented dialogue (SK-TOD). More specifically, we focus on subjective knowledge-seeking user requests and choose user reviews as external subjective knowledge sources. In Figure 1, we show three examples of such requests, where customers ask about the WiFi quality of hotels. User reviews are great resources for subjective information because even for the same aspect of the same product/service, customers may have different opinions and leave either positive or negative reviews. The subjectivity of reviews also indicates that a TOD system should consider multiple reviews to get a more comprehensive user opinion. Based on that, an ideal response should inform users of the diversity of opinions by including both positive and negative opinions as well as the proportions (like the response in Dialogue 3). Such a two-sided response has been observed as more credible and valuable for customers (Kamins et al., 1989; Lee et al., 2008; Baek et al., 2012), which can also protect the trust of users in the TOD system.

Building TOD upon subjective knowledge in this way brings in two unique challenges. First, instead of selecting the top few relevant knowledge snippets (as what is needed for Fact-TOD), the SK-TOD model needs to select all relevant knowledge snippets. In other words, both precision and recall matter during this process. Second, the model needs to aggregate these knowledge snippets into a concise response that can faithfully reflect the diversity and proportion of opinions. 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 of subjective-knowledge-grounded TOD, we collect 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 investigate the performance of strong baselines on this task. Results show that there is a large gap between human-generated and machine-generated

responses, especially in the faithfulness of the sentiment proportion. To address this issue, we further propose a model to incorporate review understanding into SK-TOD. We experimentally demonstrate that responses generated by this model can better capture the sentiment proportion. Our contributions are three-fold:

- We propose a novel task of subjective-knowledge-based TOD (SK-TOD);
- We create and release a large-scale, human-annotated dataset 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. Many 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 are merely based on the dialog context.

In the task-oriented dialogues, most works focus 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 ground TOD on user reviews, they did not consider the diversity of opinions.

	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-%).

2.2 Subjective Content Understanding

Besides being used as external knowledge sources in dialogue systems, subjective content, especially user reviews, has been studied in some non-conversational NLP tasks. For example, opinion mining (Pontiki et al., 2016; Jiang et al., 2019) aims to extract opinions and sentiments from user reviews. Opinion summarization (Chu and Liu, 2019; Angelidis et al., 2021; Brazinskas et al., 2020) is used to distill multiple opinions into a concise summary. Subjective question answering (McAuley and Yang, 2016; Bjerva et al., 2020) is proposed to answer questions based on the user reviews. Explainable recommendation (Ni et al., 2019) aims to generate review-based explanations for the items proposed 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 $C = [U_1, S_1, U_2, S_2, \dots, U_t]$ between a user and a system, where each user utterance U_i is followed by a system response utterance S_i except the last user utterance U_t . The dialogue involves a single or multiple entities $\mathcal{E} = \{e_1, \dots, e_m\}$. Along with the dialogue, we have a subjective knowledge source $\mathcal{B} = \{(e_1, \mathcal{R}_1), (e_2, \mathcal{R}_2), \dots\}$ consists of all the entities and their corresponding customer reviews. Each entity e has multiple reviews $\mathcal{R} = \{R_1, R_2, \dots\}$. Each review can be split into multiple segments $[K_1, K_2, \dots]$ 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 yes, then 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. 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. Each review contains 5.6 sentences and 56.71 tokens on average. More details about the review collection can be found in Appendix A.

After collecting reviews, we go back to the dialogue data to create the subjective user request. Following the procedure of (Kim et al., 2020), for each dialogue, we provide an aspect that users are

	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. # utterances 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.

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 the coherence and naturalness of the entire dialogue flow. Finally, we use the partial dialog until this newly inserted turn as an instance in our data. Utterances 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 last step is to ground the dialogue in 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 reviews of the gold entities by matching the aspect of user request and that of the review sentences. For example, if the aspect of a user request is about the “WIFI quality” of a hotel, all review sentences of this hotel with “WIFI quality” as the aspect will be relevant knowledge snippets.¹ Finally, we provide the dialogue context C and all related knowledge snippets \mathcal{K}^+ and ask workers to create a natural and faithful response. We explicitly ask workers to consider the diversity and proportion of opinions in all relevant knowledge snippets during response creation. Instructions can be found in Appendix A.

4.2 Data Statistics

We collected 19,696 instances with subjective user requests and subjective-knowledge-grounded responses in total. The average length of the subjective user request and the agent response is 8.75 and 24.07 tokens, respectively. While most of the instances contain a single entity, there are 1,047 in-

¹Note 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.

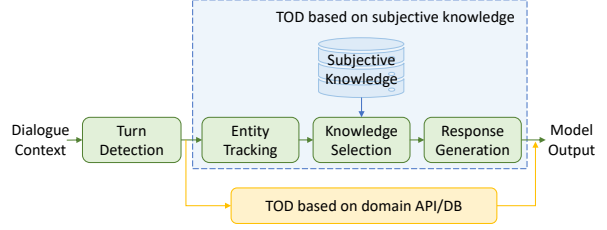


Figure 2: The pipeline architecture of SK-TOD.

stances where multiple entities are compared (like Dialogue 2 in Figure 1). Each instance requires on average 3.88 subjective knowledge snippets. To help identify the subjective knowledge-seeking user request, we randomly sample another 18,383 dialogues with non-subjective user requests from the original MultiWOZ dataset.

We split the dataset into training (75%), validation (10.8%), and test (14.2%) sets. Table 2 shows the detailed statistics of each subset. Our validation and test sets contain two subsets: the *seen* subset where the aspects of these instances are included in the training set, and the *unseen* subset where the aspects are not included in the training set. The unseen subset is designed to evaluate models’ generalizability to arbitrary aspects.

5 Subjective-Knowledge-Grounded TOD

In this section, we describe the method for SK-TOD. As shown in Figure 2, we follow the pipeline of (Kim et al., 2020) which consists of four sequential sub-tasks: knowledge-seeking turn detection (KTD), entity tracking (ET), knowledge selection (KS), and response generation (RG). The details of each subtask are described as follows.

5.1 Knowledge-Seeking Turn Detection

The goal of KTD is to identify the user request that requires to be addressed with subjective knowledge. We regard it as a binary-classification problem, where the input is the dialogue context C and the output is a binary indicator.

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

$$h = \text{Enc}(C)$$

$$P(C) = \text{softmax}(\text{FFN}(h)). \quad (1)$$

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

5.2 Entity Tracking

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

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. 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 the knowledge snippets candidates \mathcal{K} , which is a combination of all the knowledge snippets of the relevant entities in \mathcal{E} . The output $\mathcal{K}^+ \subseteq \mathcal{K}$ is a subset of relevant knowledge candidates. Note that there might be multiple knowledge snippets in \mathcal{K}^+ .

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,

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

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

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

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

5.4 Response Generation

The goal of RG is to create an utterance S_t that responds to the user’s request based on the dialogue context C and the relevant knowledge snippets \mathcal{K}^+ . We concatenate \mathcal{K}^+ and C as the input and use a pre-trained generation model to create 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 $p(S_T | C, \mathcal{K}^+)$.

To faithfully reflect 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 $Z = [z_1, \dots, z_i, \dots]$ for every knowledge snippet $K_i \in \mathcal{K}^+$. Then we incorporate the sentiment information into RG by maximizing $p(S_T | C, \mathcal{K}^+, Z)$. 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 each target task. The detailed

	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.

experimental setup can be found in Appendix C.

6.1 Knowledge-Seeking Turn Detection

Setting We explore 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 indicates that it is feasible to identify the user request that requires subjective knowledge, which can be then explicitly handled by an SK-TOD component. However, this KTD classifier may work well when restricted only to this dataset or similar, and its generalizability to unseen domains or knowledge types needs to be further explored in future works.

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 are the same as 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 that is 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 follow the setting of KTD to finetune the KS models. We compare them also with tradi-

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

	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 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.

tional information retrieval (IR) baselines such as TF-IDF (Manning et al., 2008) and BM25 (Robertson et al., 2009).

Evaluation Knowledge selection can be regarded 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 use the average over all instances as the final $P/R/F_1$. For the snippet level, instead of calculating $P/R/F_1$ for each instance, we calculate $P/R/F_1$ for all $\langle C, K \rangle$ pairs of the entire dataset. For retrieval, we use mean-average-precision (mAP) as the metric, which is insensitive to the threshold value and can reflect the overall ranking positions of all relevant knowledge snippets. Since the total number of the relevant knowledge snippets varies for each instance, we do not include top-K-based measures such as Precision@K or Recall@K which are commonly used in other Fact-TOD and knowledge-grounded open-domain dialogue problems.

Results Table 4 shows the results of KS. First, when comparing our model with IR baselines, all of the trained models outperform baselines, indicating that the KS model can benefit from the annotated training data. We then compare bi-encoder models and cross-encoder models. 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 large 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 either seen or unseen test subsets. It shows that there is 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.

6.4 Response Generation

Setting We tried GPT-2 (Radford et al.)³ and DialoGPT (Zhang et al., 2020c), two decoder-only generation models, as well as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), two encoder-decoder models. We also include BART_{ABSA} and T5_{ABSA}, two ABSA-enhanced models. 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 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 BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), as well as BERTScore (Zhang et al., 2020b) to evaluate results compared to the reference responses. We also conduct a human evaluation, where we ask crowd workers to assess the appropriateness, aspect accuracy, and sentiment accuracy of generated responses.

Results As shown in Table 5, machine-generated response is significantly better than the extractive response. Encoder-decoder models achieve better performance on all automatic measures compared with GPT-based models, indicating that they are more beneficial for this task. They also tend to generate longer responses. There is no apparent difference in automatic measures when comparing BART and T5. BART_{ABSA} achieves the best performance on BertScore while T5_{ABSA} achieves the best score on BLEU and ROUGE.

Human Evaluation While automatic measures

	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
BART _{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.

	Approp.	Asp-Acc	Senti-Acc
EXT	3.47	3.68	2.62
GPT2	3.85	4.58	2.73
BART	3.95	4.61	3.16
BART _{ABSA}	3.94	4.59	3.46
T5	3.94	4.61	3.54
T5 _{ABSA}	3.94	4.57	3.68
Reference	4.12	4.76	4.42

Table 6: Results of human evaluation for RG.

are not always reliable, we also conduct a human evaluation on AMT to better access the quality of the response. We use the same group of workers as in the data collection. 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 sample 240 instances from the test set for evaluation. For each instance, we ask three independent workers to compare the responses according to Appropriateness, Aspect Accuracy, and Sentiment Accuracy. We use the average score over all instances and workers of a system as the system-level score. More details can be found in Appendix D.

As shown in Table 6, machine-generated responses are similar to the references on appropriateness and aspect accuracy. It also shows that incorporating ABSA can improve the model performance on sentiment accuracy. However, there is still a large gap in sentiment accuracy between the best model-generated responses and the references, indicating that it is challenging for current models to faithfully aggregate sentiment information from multiple knowledge snippets.

Qualitative Analysis Table 7 shows an example of responses generated by various systems. While all reviews express negative opinions about the water pressure, the responses generated by GPT-2

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

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?
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.
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.
BART_{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.

	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.

and BART include positive opinions. T5 correctly mentions the negative opinions but the generated response is not natural and coherent. With the help of the ABSA model, both BART and T5 correctly generate the responses with all negative opinions.

7 Experiments on End-2-End Evaluation

In Section 6, we use the gold information as input for each module to avoid error accumulation. In this section, we evaluate the entire pipeline in an end-to-end manner, where the input of each sub-task 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 shows that the errors introduced during KS can decrease the quality of response generation. However, ET and KTD do not have much impact on the performance of downstream tasks. It is because the ET and KTD results include less noisy predictions compared with the KS results.

	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.

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 if a Fact-TOD model can ground on subjective knowledge to address subjective requests. To this end, we re-train our KTD (DeBERTa), KS (DeBERTa cross-encoder), and RG (BART) models using the FAQ-grounded TOD data provided by Kim et al. (2020) and then apply it to the test set of SK-TOD without further training. We compare the results of each sub-task with the results of models trained on SK-TOD.

As shown in Table 9, for all tasks, there is a large performance gap between models trained on Fact-TOD and on SK-TOD training data. By checking the model output, we further observe that the Fact-TOD model tends to only ground on and copy from a single knowledge snippet. It indicates that it is difficult to apply the Fact-TOD model to the SK-TOD task directly. It also demonstrates that compared with Fact-TOD, SK-TOD faces new challenges of subjective content understanding and dialogue modeling when incorporating subjective knowledge. We provide sampled outputs and more discussions in Appendix E.

9 Conclusion

In this paper, we propose SK-TOD: a novel task of 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 a concise response, which brings unique challenges to this task. Experiments on strong baselines show that there is a large performance gap between humans and machines, especially in faithfully depicting the diversity and proportion of opinions. We hope this task along with the dataset can promote future research on knowledge-grounded TOD and subjective content understanding.

Ethical Considerations

We collect and use the dialogue data from MultiWOZ 2.1 (Eric et al., 2020), which is a publicly available English dialogue dataset under MIT license. We collect the review data via crowdsourcing by providing the reviewer’s persona, as well as the aspects and sentiments of reviews to workers. 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 public.

During response annotation, we ground the response in multiple subjective knowledge snippets, and explicitly ask workers to consider the diversity and proportion of all snippets when creating a response. It will help to avoid the bias towards the majority of or the randomly selected opinions, which has been understudied by many subjective-content-based benchmarks (See Table 1 for more details).

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 Brazinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *ACL*.
- 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, Chinnadurai 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 *EM-SASW@ 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.

707	Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In <i>Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing</i> , pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.	763
708		764
709		
710		
711		
712		
713		
714	Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 8118–8128, Online. Association for Computational Linguistics.	
715		
716		
717		
718		
719		
720		
721	Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at dstc7. In <i>AAAI Dialog System Technology Challenges Workshop</i> .	
722		
723		
724		
725	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 <i>AAAI</i> .	
726		
727		
728		
729	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 <i>INTERSPEECH</i> , pages 1891–1895.	
730		
731		
732		
733		
734		
735	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. <i>arXiv preprint arXiv:1706.02677</i> .	
736		
737		
738		
739		
740	Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In <i>International Conference on Learning Representations</i> .	
741		
742		
743		
744	Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 6280–6285, Hong Kong, China. Association for Computational Linguistics.	
745		
746		
747		
748		
749		
750		
751		
752		
753	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 <i>Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)</i> , pages 119–127.	
754		
755		
756		
757		
758		
759		
760	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. <i>Journal of advertising</i> , 18(2):4–10.	765
761		766
762		767
	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 <i>Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue</i> , pages 278–289, 1st virtual meeting. Association for Computational Linguistics.	768
		769
		770
		771
		772
	Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papanagelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tür. 2021. “how robust ru?”: Evaluating task-oriented dialogue systems on spoken conversations. In <i>2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)</i> , pages 1147–1154. IEEE.	773
		774
		775
		776
		777
		778
		779
	Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In <i>Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.	780
		781
		782
		783
		784
		785
	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 <i>International Conference on Learning Representations</i> .	786
		787
		788
		789
		790
	Jumin Lee, Do-Hyung Park, and Ingo Han. 2008. The effect of negative online consumer reviews on product attitude: An information processing view. <i>Electronic commerce research and applications</i> , 7(3):341–352.	791
		792
		793
		794
		795
	Esther Levin, Roberto Pieraccini, and Wieland Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. <i>IEEE Transactions on speech and audio processing</i> , 8(1):11–23.	796
		797
		798
		799
	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 <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 7871–7880.	800
		801
		802
		803
		804
		805
		806
		807
	Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In <i>Text Summarization Branches Out</i> , pages 74–81, Barcelona, Spain. Association for Computational Linguistics.	808
		809
		810
		811
	Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In <i>Mining text data</i> , pages 415–463. Springer.	812
		813
		814
	Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In <i>Proceedings of the 56th Annual Meeting of the Association</i>	815
		816
		817
		818

819	for Computational Linguistics (Volume 1: Long Pa-	knowledge graphs. In <i>Proceedings of the 57th An-</i>	875
820	pers), pages 1489–1498, Melbourne, Australia. As-	annual Meeting of the Association for Computational	876
821	sociation for Computational Linguistics.	Linguistics, pages 845–854, Florence, Italy. Associ-	877
822	Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man-	ation for Computational Linguistics.	878
823	dar Joshi, Danqi Chen, Omer Levy, Mike Lewis,		
824	Luke Zettlemoyer, and Veselin Stoyanov. 2019.	Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019.	879
825	Roberta: A robustly optimized bert pretraining ap-	Justifying recommendations using distantly-labeled	880
826	proach. <i>arXiv preprint arXiv:1907.11692</i> .	reviews and fine-grained aspects. In <i>Proceedings</i>	881
827	Yinong Long, Jianan Wang, Zhen Xu, Zongsheng	of the 2019 conference on empirical methods in	882
828	Wang, Baoxun Wang, and Zhuoran Wang. 2017. A	natural language processing and the 9th interna-	883
829	knowledge enhanced generative conversational ser-	tional joint conference on natural language process-	884
830	vice agent. In <i>Proceedings of the 6th Dialog System</i>	ing (EMNLP-IJCNLP), pages 188–197.	885
831	<i>Technology Challenges (DSTC6) Workshop</i> .		
832	Ilya Loshchilov and Frank Hutter. 2018. Decoupled	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-	886
833	weight decay regularization. In <i>International Con-</i>	Jing Zhu. 2002. Bleu: a method for automatic eval-	887
834	ference on Learning Representations.	uation of machine translation. In <i>Proceedings of the</i>	888
835	Bodhisattwa Prasad Majumder, Harsh Jhamtani, Tay-	40th Annual Meeting of the Association for Compu-	889
836	lor Berg-Kirkpatrick, and Julian McAuley. 2022.	tational Linguistics, July 6-12, 2002, Philadelphia,	890
837	Achieving conversational goals with unsupervised	PA, USA, pages 311–318. ACL.	891
838	post-hoc knowledge injection. In <i>Proceedings of the</i>		
839	60th Annual Meeting of the Association for Compu-	Gustavo Penha, Alexandru Balan, and Claudia Hauff.	892
840	tational Linguistics (Volume 1: Long Papers), pages	2019. Introducing mantis: a novel multi-domain in-	893
841	3140–3153, Dublin, Ireland. Association for Com-	formation seeking dialogues dataset. <i>arXiv preprint</i>	894
842	putational Linguistics.	<i>arXiv:1912.04639</i> .	895
843	Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni,	Maria Pontiki, Dimitrios Galanis, Haris Papageor-	896
844	and Julian McAuley. 2020. Interview: Large-scale	giou, Ion Androutsopoulos, Suresh Manandhar, Mo-	897
845	modeling of media dialog with discourse patterns	hammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan	898
846	and knowledge grounding . In <i>Proceedings of the</i>	Zhao, Bing Qin, Orphée De Clercq, et al. 2016.	899
847	2020 Conference on Empirical Methods in Natural	Semeval-2016 task 5: Aspect based sentiment anal-	900
848	Language Processing (EMNLP), pages 8129–8141,	ysis. In <i>International workshop on semantic evalua-</i>	901
849	Online. Association for Computational Linguistics.	tion, pages 19–30.	902
850	Christopher D Manning, Prabhakar Raghavan, and Hin-	Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou,	903
851	rich Schütze. 2008. <i>Introduction to information re-</i>	Suresh Manandhar, and Ion Androutsopoulos. 2015.	904
852	trieval. Cambridge university press.	Semeval-2015 task 12: Aspect based sentiment anal-	905
853	Pierre-Emmanuel Mazaré, Samuel Humeau, Martin	ysis. In <i>Proceedings of the 9th international work-</i>	906
854	Raison, and Antoine Bordes. 2018. Training mil-	shop on semantic evaluation (SemEval 2015), pages	907
855	lions of personalized dialogue agents. In <i>Proceed-</i>	486–495.	908
856	ings of the 2018 Conference on Empirical Methods	Alec Radford, Jeffrey Wu, Rewon Child, David Luan,	909
857	in Natural Language Processing, pages 2775–2779,	Dario Amodei, Ilya Sutskever, et al. Language mod-	910
858	Brussels, Belgium. Association for Computational	els are unsupervised multitask learners.	911
859	Linguistics.		
860	Julian McAuley and Alex Yang. 2016. Addressing	Colin Raffel, Noam Shazeer, Adam Roberts, Katherine	912
861	complex and subjective product-related queries with	Lee, Sharan Narang, Michael Matena, Yanqi Zhou,	913
862	customer reviews. In <i>Proceedings of the 25th Inter-</i>	Wei Li, Peter J Liu, et al. 2020. Exploring the limits	914
863	national Conference on World Wide Web, pages 625–	of transfer learning with a unified text-to-text trans-	915
864	635.	former. <i>J. Mach. Learn. Res.</i> , 21(140):1–67.	916
865	Nikita Moghe, Siddhartha Arora, Suman Banerjee, and	Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara,	917
866	Mitesh M. Khapra. 2018. Towards exploiting back-	Raghav Gupta, and Pranav Khaitan. 2020. Towards	918
867	ground knowledge for building conversation sys-	scalable multi-domain conversational agents: The	919
868	tems. In <i>Proceedings of the 2018 Conference on</i>	schema-guided dialogue dataset. In <i>Proceedings of</i>	920
869	<i>Empirical Methods in Natural Language Processing,</i>	the AAAI Conference on Artificial Intelligence, vol-	921
870	<i>Brussels, Belgium, October 31 - November 4, 2018,</i>	ume 34, pages 8689–8696.	922
871	pages 2322–2332.	Nils Reimers and Iryna Gurevych. 2019. Sentence-	923
872	Seungwhan Moon, Pararth Shah, Anuj Kumar, and Ra-	BERT: Sentence embeddings using Siamese BERT-	924
873	jen Subba. 2019. OpenDialKG: Explainable conver-	sational networks. In <i>Proceedings of the 2019 Confer-</i>	925
874	sational reasoning with attention-based walks over	ence on Empirical Methods in Natural Language Processing	926
		and the 9th International Joint Conference on Natu-	927
		ral Language Processing (EMNLP-IJCNLP), pages	928
		3982–3992, Hong Kong, China. Association for	929
		Computational Linguistics.	930

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.
- 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.,

Instruction

Please assume that you recently visited **MIDSUMMER HOUSE RESTAURANT** alone. This place serves **British** cuisine and you ordered the following:

- Dishes:
 - Strawberries and Cream
- Drinks:
 - beer

Please write down your review comments based on the following aspects:

- What you liked:
 - Good portion of foods
 - High-quality foods
- What you disliked:
 - Overpriced drinks

Notes:

- Please do **NOT** copy and paste the aspects as they are.
- Please provide as many details as possible.

Your review post:

Write down a review post

Submit

Figure 3: The interface of review collection.

GOAL: You prefer a place which has a stable Internet connection. You must ask a question about this for the hotel.

NOTE: Please do NOT insert a question before the hotel name is explicitly mentioned in the conversation.

User: Where can I get some British food that isn't too cheap or expensive?
Agent: There are a total of 5 British restaurants that are moderately priced. Is there any particular location you are wanting such as the centre or west?

User: Great. Do you know of a high-end hotel I could stay at? It has to have free wifi.
Agent: The University Arms hotel is an expensive place with free wifi, conveniently located very close to your restaurant. Would you like to book a room there?

User: (Do they provide fast and stable wifi to watch online videos?)
Agent:

User: I might be interested. First, is that a guesthouse style? I would love to stay in a guesthouse style with free parking!
Agent: Unfortunately no, it is not a guesthouse style lodging. There are none that fall into the expensive category, would you like me to search for something cheaper that is a guesthouse?

User: No, that's okay. Are there any hotels available in the moderate price range?
Agent: There are 12 guest houses that have free wifi and free parking. Would you like me to book one for you?

Figure 4: The interface of user request collection.

2021), a state-of-the-art aspect-based sentiment analysis model, to predict the sentiment $Z = [z_1, z_2, \dots, z_i, \dots]$ for every knowledge snippet $[K_1, K_2, \dots, K_i, \dots]$ 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 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).

Conversation:

Customer: Hi there. I'm on a really tight budget but need to find me a cheap room to stay.
 Agent: We have several cheap hotels and guesthouses available. Do you have a particular location in mind?
 Customer: Not really. It does need to be a hotel and parking isn't really an issue with us.
 Agent: I have one in the cheap range, shall I book it?
 Customer: Does it have a star of 0. That is very important to me?
 Agent: Unfortunately, it's 4 stars. Would you like to try a different hotel type?
 Customer: What about one that is in the moderate price range?
 Agent: I found 3 hotels, do you have a preference for the area of centre or north?
 Customer: No area preference. Any of them are fine. Can you book one for me starting Thursday?
 Agent: Yes, of course. For how many nights and how many people?
 Customer: There will be three people and we need it for 3 nights. Can I get the reference number for that once it is booked?
 Agent: I've booked you at the Ashley Hotel, and your reference number is 9ZKDNRMV.
 Customer: Wait. Is there a good internet connection at the Ashley?

Agent: Write down a response (longer than 100 characters) to the customer to continue the

Submit

Knowledge:

Customer Reviews

ASHLEY HOTEL

... This is very frustrating when you are trying to work and the delay is huge. ...

... Also, the fast wi-fi was a huge plus for work I had to work to do on my laptop. ...

... Also, the Wi-Fi was extremely slow. ...

Figure 5: The interface of response generation.

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.

D Human Evaluation

We ask workers to compare these responses according to the following 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 accordant with that of the subjective knowledge.

For Appropriateness and Aspect Accuracy, we only show the dialogue context and the responses. For Sentiment Accuracy, we further show the oracle knowledge snippets. To increase the annotation quality, we first ask workers to annotate the sentiment label of each knowledge snippet, and then evaluate the sentiment accuracy of each response. All three measures are evaluated using the 5-Point Likert scale. We pay \$0.25 for the task of Appropriateness and Aspect Accuracy and \$0.4 for the task of Sentiment Accuracy. Figure 6 shows the interface of evaluating Sentiment Accuracy. The interfaces of Appropriateness and Aspect Accuracy are similar to that of Sentiment Accuracy.

E Comparison with Fact-TOD

In Table 10 and 11, we provide two responses generated by the Fact-TOD model and the SK-TOD model, respectively. It can be found that responses generated by the Fact-TOD only ground on a single knowledge snippet, indicating that the Fact-TOD model lacks the ability to aggregate information from multiple knowledge snippets, especially when there are diverse and contradictory opinions. We can also find that Fact-TOD tends to copy from knowledge snippets but fail to distinguish between facts and opinions. It is because when grounding on factual knowledge, it is sufficient to ground on a few knowledge snippets and copy from the factual knowledge text. However, when grounding on subjective knowledge, it is essential to aggregate information from multiple snippets and handle diverse and contradictory opinions.

Instructions

In this HIT, we provide a conversation between a customer and a virtual agent which ends with a customer input, the entire user reviews that are relevant to the customer's input, as well as 5 responses which were created by multiple different agents.

This HIT asks you to read conversation and score the sentiment accuracy of different agent responses based on the provided user reviews. Your task is to score how accurate each agent's response reflects the opinions that you obtained from the reviews. During evaluation, please consider:

- The sentiment (positive or negative) of each individual user review;
- The overall sentiment by consider all user reviews (for example, whether all users have positive / negative opinions; or a majority of users have positive/negative opinions; or it's 50/50?);

Please score the responses on a scale of 1 - 5. A score of 1 means that the response is completely opposite from the overall sentiment of the user reviews, and a score of 5 means that the response completely and accurately reflects the overall sentiment of the user reviews.

Please complete the task by taking the following steps:

- Read the conversation.
- Read the reviews that are from other users.
- Evaluate the sentiment of each review (positive / negative / neutral).
- Evaluate the overall sentiment of all reviews (from all positive to all negative).
- Read the final responses spoken by different agents.
- Examine how accurate is each response to the overall sentiment of the user reviews.
- Select the sentiment accuracy score for each response.
- Click the submit button.

Conversation

User: I want a hotel that has free parking.
Agent: There are several, got a price range, or area of town you'd like to stay in?
User: I want to stay in the northern part of town.
Agent: Ashley hotel is in the north in the moderate price range with free parking and internet. Would you like to reserve a room?
User: Yes, please reserve a room.
Agent: I'd be happy to help with your request, but I'll need to know what day you're arriving, how many are staying and for how many nights?
User: Never mind. I want to know is this a guesthouse. If not I need one.
Agent: No its not but the Hamilton Lodge is one. would you like a room?
User: Sure. Can I get the postcode, price range, and phone number? I also need an attraction in the same area of town.
Agent: The postcode for Hamilton Lodge is c041da and their phone is 01223385664. They're in the moderate price range.
User: It is important to me to enjoy my room as I will be in it working for some of my stay. Do you know if the room is nice with a great view?

Reviews

Here we show multiple reviews (1,2,3,...) that are from different customers. Please evaluate the sentiment of each single review sentence by selecting the most suitable radio button under "Review Sentiment Annotation".

Please only consider the sentiment of the aspect that are relevant to the last user request in the conversation.

Review Knowledge	Review Sentiment Annotation		
1. Hamilton Lodge: I enjoyed working no my speech while looking at the views.	<input type="radio"/> Positive	<input type="radio"/> Negative	<input type="radio"/> Neutral
2. Hamilton Lodge: I had a great view of the mountains.	<input type="radio"/> Positive	<input type="radio"/> Negative	<input type="radio"/> Neutral
3. Hamilton Lodge: The view was surprisingly good and the room and bathroom were pretty clean.	<input type="radio"/> Positive	<input type="radio"/> Negative	<input type="radio"/> Neutral
4. Hamilton Lodge: The price is steep and not good at all, but the location makes it very convenient and the views in the rooms were very nice.	<input type="radio"/> Positive	<input type="radio"/> Negative	<input type="radio"/> Neutral

Which option do you think can best describe the overall sentiment when considering **all review knowledge**?

☐ 1. All reviews have positive opinions
☐ 2. The reviews have mixed opinions. Most reviews have positive opinions while others have negative opinions
☐ 3. The reviews have mixed opinions. half of reviews have positive opinions while the other half have negative opinions
☐ 4. The reviews have mixed opinions. Most reviews have negative opinions while others have positive opinions
☐ 5. All reviews have negative opinions

Responses

From 1-5, select the sentiment accuracy score for each response. 1: completely opposite; 5: completely accurate.

When there are multiple restaurants or hotels involved, please consider if the agent makes correct recommendations based on the sentiment proportion. A good agent should recommend the restaurants or hotels that are with more positive reviews compared with negative reviews. 1: completely wrong recommendation; 5: completely correct recommendation;

You should **NOT** consider other factors when evaluating the response such as length, fluency, and naturalness.

Responses	Sentiment Accuracy Scores
Agent 0: The view was surprisingly good and the room and bathroom were pretty clean.	<input type="radio"/>
Agent 1: Yes. At Hamilton Lodge, a view of the mountains is possible depending on your room. People have said it was very nice.	<input type="radio"/>
Agent 2: The Hamilton Lodge has a really nice view of the mountains. Would you like to know more about them?	<input type="radio"/>
Agent 3: The Hamilton Lodge has really good views according to most guests. Only one person said it wasn't very good.	<input type="radio"/>
Agent 4: Yes, the rooms at the Hamilton Lodge have a great view of the mountains. Would you like me to book a room for you?	<input type="radio"/>
Agent 5: Yes, the rooms at the Hamilton Lodge are nice with a great view of the mountains. Would you like me to book a room for you?	<input type="radio"/>
Agent 6: The Hamilton Lodge has really nice views of the mountains. Would you like me to make a reservation for you?	<input type="radio"/>

Submit

Figure 6: The interface of evaluating the sentiment accuracy of responses.

U: Hi! I'm looking for a place to stay with 3 stars that includes wifi.

S: I can help you with that. I have several options that fit this criteria. Do you prefer a particular area of town?

U: Just near Cambridge , and should be a guesthouse.

S: Bridge guest house is a guesthouse in the south in the moderate price range with free wifi and parking. Would you like to reserve a room?

U: Certainly. There will be 5 of us for 4 nights starting on Tuesday.

S: I was able to book for 5 people for 4 nights starting on Tuesday. The reference number is 97VS1TDE. Is there anything else I can help you with?

U: I prefer a big room so do you know how big the room is?

Review 1: The room was smaller than expected and I was surprised that they didn't have any luggage storage service for the amount we paid.

Review 2: Our room was bright and cheerful with lots of space and a newly refurbished bathroom.

Fact-TOD: The Bridge Guest House's room is smaller than expected and I was surprised that they don't have any luggage storage services. Do you need anything else?

SK-TOD: According to the reviews I have on hand for that location, the opinions are mixed. Half found the rooms smaller than expected and half found them bright and cheerful with lots of space and a newly refurbished bathroom.

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?

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.

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.

Table 10: The first sampled output from the Fact-TOD model and the SK-TOD model, respectively.

Table 11: The second sampled output from the Fact-TOD model and the SK-TOD model, respectively.