

# Adding Chit-Chat to Enhance Task-Oriented Dialogues

Kai Sun<sup>1\*</sup>, Seungwan Moon<sup>2</sup>, Paul Crook<sup>2</sup>, Stephen Roller<sup>3</sup>, Becca Silvert<sup>2</sup>,  
Bing Liu<sup>2</sup>, Zhiguang Wang<sup>2</sup>, Honglei Liu<sup>2</sup>, Eunjoon Cho<sup>2</sup>, and Claire Cardie<sup>1</sup>

<sup>1</sup>Cornell University

<sup>2</sup>Facebook, <sup>3</sup>Facebook AI Research

✉ ks985@cornell.edu, shanemoon@fb.com

## Abstract

Existing dialogue corpora and models are typically designed under two disjoint motives: while task-oriented systems focus on achieving functional goals (e.g., booking hotels), open-domain chatbots aim at making socially engaging conversations. In this work, we propose to integrate both types of systems by Adding Chit-Chat to ENhance Task-ORiented dialogues (ACCENTOR), with the goal of making virtual assistant conversations more engaging and interactive. Specifically, we propose a Human  $\leftrightarrow$  AI collaborative data collection approach for generating diverse chit-chat responses to augment task-oriented dialogues with minimal annotation effort. We then present our new chit-chat-based annotations to 23.8K dialogues from two popular task-oriented datasets (Schema-Guided Dialogue and MultiWOZ 2.1) and demonstrate their advantage over the originals via human evaluation. Lastly, we propose three new models for adding chit-chat to task-oriented dialogues, explicitly trained to predict user goals *and* to generate contextually relevant chit-chat responses. Automatic and human evaluations show that, compared with the state-of-the-art task-oriented baseline, our models can code-switch between task and chit-chat to be more engaging, interesting, knowledgeable, and humanlike, while maintaining competitive task performance.

## 1 Introduction

With modeling innovations, increasing computing power, and a growing number of datasets, recent years have witnessed significant improvements in the performance of both task-oriented dialogue systems and chit-chat systems (Adiwardana et al., 2020; Roller et al., 2020; Hosseini-Asl et al., 2020; Peng et al., 2020a). Most research on dialogue

\*Work done as a research intern at Facebook. The code and data are available at <https://github.com/facebookresearch/accentor>.

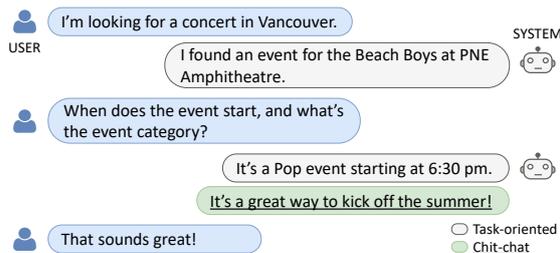


Figure 1: A sample task-oriented dialogue snippet augmented by *chit-chat*.

systems focuses on a particular type of dialogue system. Work on task-oriented dialogue systems typically aims to track user goals with higher accuracy to better achieve functional goals (Rastogi et al., 2020) with the sacrifice of not paying explicit attention to user experience, such as making the conversation more engaging, while the latter is usually the target of research on chit-chat systems (Li et al., 2019). In this work, we step forward and propose to integrate both types of systems by Adding Chit-Chat to ENhance Task-ORiented dialogues (ACCENTOR), aiming to have a virtual assistant capable not only of performing various complex tasks such as checking the weather, booking hotels, and finding restaurants, but also incorporating casual and contextually relevant chit-chat. We hypothesize that the added chit-chat can make the assistant appear more social, personable, and engaging, without being misleading or inappropriate, compared with existing task-oriented dialogue systems.

To show the feasibility of ACCENTOR and gather supervisory data for follow-up research, we propose a Human $\leftrightarrow$ AI collaborative data construction approach that can effectively add suitable chit-chat to the beginning or end of system responses in existing task-oriented dialogue datasets. Specifically, we first generate chit-chat candidates for augmentation using off-the-shelf pre-trained language models and open-domain chatbots (Section 2.1). Next, we automatically filter out candidates that

are unlikely to be of good quality using a filter model (Section 2.2). Finally, human annotators label each of the remaining candidates as good or bad, with justifications (Section 2.3). We augment the Schema-Guided Dialogue (SGD) (Rastogi et al., 2020) and MultiWOZ 2.1 (Eric et al., 2020) corpora using the proposed approach. (See Figure 1 or Appendix A.4 for examples.) We employ ACUTE-Eval (Li et al., 2019) to compare the augmented versions with the originals along four axes: engagingness, interestingness, knowledge, and humanness. We find that the augmented dialogues are consistently preferred by human judges across the four axes for both datasets (Section 4.1).

In addition, we propose and evaluate three models for adding chit-chat to task-oriented dialogues, including an end-to-end model and two code-switcher models built upon off-the-shelf task-oriented and chit-chat systems (Section 3). Compared with the baseline model trained with the original unaugmented data, our models trained with the augmented version can generate significantly higher-rated responses in terms of human preference while maintaining competitive task performance in goal tracking accuracy and action decision F1 (Section 4.2).

**Our main contributions** are: we propose (1) a data augmentation approach for generating diverse chit-chat supervisory data for task-oriented dialogues, leveraging pre-trained generative models and a custom filter model to minimize human annotation effort; (2) new versions of the popular task-oriented datasets, SGD and MultiWOZ 2.1, with newly added chit-chat annotations to 23.8K dialogues; and (3) three integrated chit-chat and task-oriented neural dialogue models for the above, substantially outperforming the state-of-the-art approach in terms of human evaluation of engagingness, interestingness, knowledge, and humanness. To our knowledge, we are the first to propose an annotated dataset and models that study explicit code-switching between full-stack task-oriented dialogues and free-form chit-chat responses.

## 2 Data Construction

In this section, we describe an approach to gather supervisory data for adding contextually relevant chit-chat to task-oriented dialogues. Our approach needs minimal annotation effort to augment suitable and diverse chit-chat add-ons that are not available in existing task-oriented datasets (Sec-

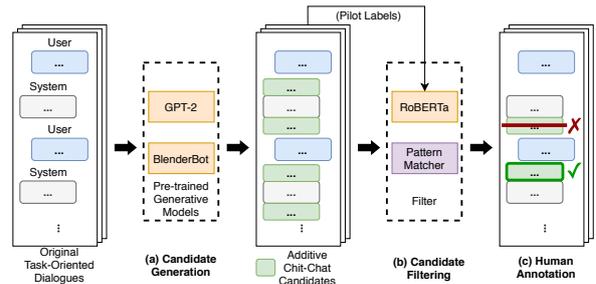


Figure 2: Data construction overview: (a) We generate diverse free-form chit-chat candidates using the state-of-the-art pre-trained generative models to augment original task-oriented dialogues, and (b) filter out bad candidates using the custom filter to minimize annotation effort. (c) Crowd workers annotate contextually relevant chit-chat augmentation with justifications.

tion 5.1). We primarily report results based on dialogues from the SGD dataset in this study, because it is the largest task-oriented dialogue dataset and is generally cleaner compared with most other task-oriented dialogue datasets. However, our approach is flexible and thus not limited to dialogues from a particular task-oriented dataset (Section 4.1). Figure 2 shows the overview of our approach.

### 2.1 Candidate Generation

Given a task-oriented dialogue  $D = \{u_1, s_1, u_2, s_2, \dots, u_n, s_n\}$ , where  $u_{1..n}$  and  $s_{1..n}$  represent user turns and system turns, respectively, we generate chit-chat candidates for augmenting  $s_i$  in two ways: (i) pass  $u_1, s_1, \dots, u_i, s_i$  to an off-the-shelf pre-trained model (a language model or a chit-chat chatbot) and let the model add tokens to the end of  $s_i$ ; (ii) pass  $u_1, s_1, \dots, u_i$  to a pre-trained model and let the model generate a turn. We regard the output of (i) and (ii) as a chit-chat candidate to be appended and prepended to  $s_i$ , respectively. If a chit-chat candidate consists of multiple sentences, we also regard each individual sentence as a chit-chat candidate. We run differently sized GPT-2 (Radford et al., 2019) and BlenderBot (Roller et al., 2020) with various decoding parameters as the pre-trained model and generate an average of 175.5 candidates for each of the dialogues from the SGD dataset. See Appendix A.1 for configuration details.

### 2.2 Candidate Filtering

We examine the quality of the model-generated candidates from Section 2.1 by performing a pilot annotation ourselves on a small proportion of the

	Appropriate Behaviours	Examples	Inappropriate Behaviours	Examples
Opinions	Express general opinions about generic, impersonal, or non-sensitive topics.	- "I love penguins." - "There's a lot of fun stuff to do."	Express strong personal opinions, or opinions on sensitive topics.	- "I love you." - "The President is an idiot."
Preferences	Express preferences when making impersonal, or non-sensitive recommendations.	- "Their latest album wasn't as good." - "Their food is good."	Express strong dispreferences, or preferences on personal or sensitive subjects.	- "I hated it, but you might like it." - "Invite her! I like her better."
Physical Actions	Use epistemic verbs to express uncertainty or opinions, or refer through hearsay to actions that it may not perform.	- "I hear it's beautiful." - "They say it tastes like chicken."	Behave as though it could act physically, or perform tasks outside of its role.	- "I haven't arrived there yet." - "I can drive you there."
Experiences	Refer to others' experiences or personify experiences it is capable of (e.g., reading).	- "That sounds like a great trip!" - "I enjoyed reading that novel."	Pretend to have experiences that it is incapable of.	- "We didn't have that when I was a kid." - "My roommate used to eat there a lot."

**Who is the virtual assistant?** This digital assistant is more than just a bot that spits out facts. It has access to a wide range of information which can express not only as factual commentaries but also as opinions and preferences. However, it is not a person and should not pretend to have real experiences or be capable of physical actions. It should be personable and personlike, without appearing counterfeit.

Table 1: The role of the virtual assistant and its appropriate/inappropriate behaviors with examples.

candidates. The annotation results show that only about  $1/10$  of the candidates are suitable. Therefore, instead of directly sending the candidates to crowd workers for annotation, we propose to build a filter model to automatically filter out candidates that are unlikely to be of good quality first to reduce potential annotation workload.

The filter is a hybrid model that consists of a RoBERTa-based binary classifier (Liu et al., 2019) and a rule-based ranker. The classifier takes as input an augmented dialogue, in which we explicitly surround the added chat candidate with a pair of special tokens to help the model locate the candidate. We train the classifier with 1.7K candidates that are labeled as good/bad from the pilot annotation. The rule-based ranker ranks each candidate based on (i) the posterior probability output by the binary classifier, (ii) whether the candidate matches a list of bad patterns (e.g., containing an URL), (iii) the frequency of appearances of the candidate among all generated candidates, (iv) the similarity to the other candidates for the dialogue, and (v) the similarity to the system response being augmented. While (i) and (ii) directly help evaluate the quality of the candidate, (iii), (iv), and (v) additionally help create more variety (e.g., punishing high-frequency candidates such as "You're welcome"). We keep the top ten candidates for each of the dialogues. We present more details in Appendix A.2.

### 2.3 Annotation

We ask annotators (crowd workers) to label each of the remaining candidates from Section 2.2 as *good* or *bad*. Additionally, to guide the annotation process, improve the potential quality, and facil-

itate the candidate distribution analysis, we also ask annotators to choose from four justifications that we come up with based on our pilot annotation experience to support their annotations. Annotators can choose one, both, or neither of the following justifications for a *bad* candidate:

- **Inappropriate:** The candidate does not fit into the context (e.g., repeating, unnatural), or it contradicts the context or the role of the assistant (Table 1). This category comprises most of the commonly found bad cases such as improper switching, providing opinions or comments that are incompatible with the context, and misusing verbal routine.
- **Misleading:** The candidate provides additional information that is false or cannot be verified immediately. For example, the underlined candidate in the two-turn dialogue "*U: I want to book a hotel room in San Diego with a check in on Thursday. A: There are over 10 hotels in San Diego. I would stay at Arlo NoMad if I were you.*" should be marked as *misleading* because "*Arlo NoMad*" is newly introduced information, which the annotator would have to look up to verify that a hotel by this name exists in San Diego, even though the information may be true.

Annotators can choose one, both, or neither of the following justifications for a *good* candidate:

- **Social:** The candidate keeps the conversation flowing smoothly by appropriately switching to relevant topics, asking casual follow up questions, or engaging in social pleasantries.

The design of this subcategory is inspired by the line of research that studies different social and discourse strategies in chit-chat dialogue systems (Yu et al., 2016).

- Useful: The candidate enhances the conversation by appropriately offering opinions, commentaries, or pertinent and truthful information. Truthfulness should be established by conversational context or real world knowledge. To reduce annotation workload, if annotators have to use external resources (e.g., Wikipedia, search engines, maps) to verify information, they are instructed to label the candidate as *misleading* instead. The design of this subcategory is inspired by the line of work on knowledge-grounded dialogue systems that study contextual knowledge injections (Dinan et al., 2019).

We instruct annotators to evaluate each candidate independently as if it were the only augmentation for its associated dialogue. We discuss the additional dimension of complexity introduced by having multiple augmentations jointly in Section 4.1.

Metric	Value
# of candidates	228, 250
# of unique candidates	68, 406
vocabulary size	10, 005
# of distinct 2-grams	59, 259
# of distinct 3-grams	131, 989
# of distinct 4-grams	195, 508
# of distinct 5-grams	239, 278
average length (in tokens)	8.7
<hr/>	
# of good candidates (%)	94, 600 (41.4)
◊ social	86, 324 (37.8)
◊ useful	7, 681 ( 3.4)
◊ social & useful	577 ( 0.3)
◊ other (good)	18 ( 0.0)
<hr/>	
# of bad candidates (%)	133, 650 (58.6)
◊ inappropriate	127, 648 (55.9)
◊ misleading	5, 800 ( 2.5)
◊ inappropriate & misleading	164 ( 0.1)
◊ other (bad)	38 ( 0.0)

Table 2: Statistics of annotated chit-chat candidates in ACCENTOR-SGD.

Annotation time per dialogue is 243s. The Fleiss’ Kappa among crowd workers is 0.52. We view the agreement score as reasonable since whether an added chit-chat candidate leads to improved quality of a conversation can be highly subjective in many scenarios. We denote our augmented version of the SGD dataset as ACCENTOR-SGD and summarize the statistics in Table 2. We observe that the four provided justification categories provide adequate coverage of the justifications for most

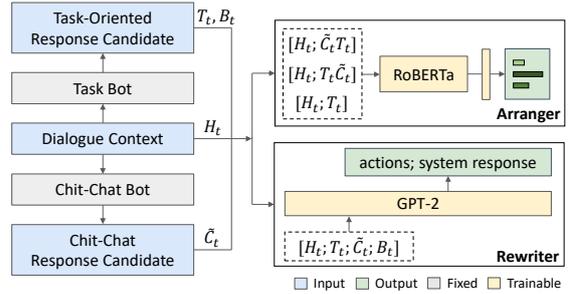


Figure 3: A diagram for the proposed code-switching models. Given the dialogue context ( $H_t$ ) and the pre-generated task-oriented and chit-chat response candidates ( $T_t, \tilde{C}_t$ ), the **Arranger** learns the optimal code-switching sequences (discriminative), while the **Rewriter** outputs free-form paraphrases (generative).

annotations. 41.4% of the candidates are good, showing the effectiveness of candidate filtering. An analysis based on linguistic features suggests that bad candidates are more personal and negative than good candidates. Specifically, 40.0% of bad candidates involve first-person pronouns, while the ratio is 26.5% for good candidates. 81.7% of good candidates have positive sentiment, measured by VADER, a lexicon and rule-based sentiment analysis tool (Hutto and Gilbert, 2014), while the ratio is 73.0% for bad candidates. Examples of the resulting dataset are presented in Appendix A.4.

## 3 Approaches

### 3.1 Task Formulations

Since oracle information (i.e., oracle belief states and oracle action decisions) is not available in practical use and the SGD dataset does not have the associated database (i.e., a table of possible entities) released, we focus on exploring the end-to-end setting in which we generate *delexicalized* task-oriented responses without using oracle information and database search results following Hosseini-Asl et al. (2020). Given dialogue history (i.e., previous turns) as context, the goal of the model for each system turn is to accurately generate belief states (i.e., a list of  $(domain, slot, value)$  triplets), action decisions (i.e., a list of  $(domain, action\_type, slot)$  triplets), and a corresponding system response that is functionally accurate *and* socially engaging.

### 3.2 Models

We re-implement SimpleTOD (Hosseini-Asl et al., 2020) as our main baseline model, which is a state-

of-the-art model in the end-to-end setting we explore. In addition, we propose an extension of SimpleTOD that incorporates chit-chat acts, as well as two new models (Arranger and Rewriter; Figure 3) that code-switch between chit-chat and task-oriented responses more explicitly.

**SimpleTOD.** It is a causal language model that models the joint probability over the concatenation of dialogue history  $H_t$ , belief states  $B_t$ , action decisions  $A_t$ , and a task-oriented response  $T_t$  for each turn  $t$ . During inference, the model takes as input  $H_t$  and generates  $B_t$ ,  $A_t$ , and  $T_t$ . We refer readers to Hosseini-Asl et al. (2020) for more details.

**SimpleTOD+.** We extend SimpleTOD by introducing to the construction of input sequences a special new dialogue action `chit-chat` and good chit-chat candidates during training. Specifically, let  $\mathbb{C}_t^+$  denote the set of good candidates for system turn  $t$ . If  $\mathbb{C}_t^+$  is empty, we construct the same training sequence as SimpleTOD. Otherwise, for each  $C_t \in \mathbb{C}_t^+$  that is labeled as a candidate to be prepended (resp. appended) to the turn, we use the concatenation of  $H_t$ ,  $B_t$ , [`chit-chat`],  $A_t$ ,  $C_t$ , and  $T_t$  (resp.  $H_t$ ,  $B_t$ ,  $A_t$ , [`chit-chat`],  $T_t$ , and  $C_t$ ) as a training sequence.

**Arranger.** This model arranges the output of an off-the-shelf task-oriented dialogue model and an off-the-shelf chit-chat model without intervening in the task. It outputs the belief states and action decisions generated by the task-oriented model without modification. To generate a response for each system turn  $t$ , this model takes as input (i) dialogue history  $H_t$ , (ii) a chit-chat response  $\tilde{C}_t$  generated by the chit-chat model based on  $H_t$ , and (iii) a task-oriented response  $T_t$  generated by the task-oriented dialogue model based on  $H_t$ . The model chooses one of the following as the response:  $\tilde{C}_t$  followed by  $T_t$ ,  $T_t$  followed by  $\tilde{C}_t$ , and  $T_t$  only. Specifically, the model encodes the concatenation of  $H_t$  and each of these three responses by a RoBERTa encoder (Liu et al., 2019) and passes the resulting representations through a linear plus softmax layer to make the choice. To train the model, we form training instances by regarding each chit-chat candidate for turn  $t$  from the training set of ACCENTOR-SGD as  $\tilde{C}_t$  and the ground-truth task-oriented response as  $T_t$  and setting the target choice based on the label (i.e., good/bad) and position (i.e., beginning/end of the response) of the candidate.

**Rewriter.** This model rewrites the output of an off-the-shelf task-oriented dialogue model and an off-the-shelf chit-chat model. It directly outputs the task-oriented model’s belief states without modification and generates action decisions and a system response by a causal language model. The causal language model differs from SimpleTOD+ in that it has two additional components  $T_t$  and  $\tilde{C}_t$  added between  $H_t$  and  $B_t$  in each training sequence, where we form  $T_t$  and  $\tilde{C}_t$  in the same way as we do for Arranger. During the inference stage, it takes as input  $H_t$ ,  $T_t$  output by the task-oriented dialogue model,  $\tilde{C}_t$  output by the chit-chat model, and  $B_t$  output by the task-oriented dialogue model, and generates action decisions and a system response. Note that since 25.4% of the annotated system turns in the training set of ACCENTOR-SGD have both good and bad chit-chat candidates,  $\mathbb{C}_t^+$  can be non-empty when  $\tilde{C}_t$  is a bad candidate, which enables the model to potentially generate a suitable chit-chat augmented response even if the output of the off-the-shelf chit-chat model is not good.

### 3.3 Implementation Details

Unless specified otherwise, for causal language models, we use the 12-layer GPT-2 (117M parameters) as the pre-trained language model (Radford et al., 2019) and fine-tune for ten epochs. We set the batch size to 36 and the learning rate to  $1 \times 10^{-3}$ . We employ the SimpleTOD baseline as the off-the-shelf task-oriented dialogue model for Arranger and Rewriter. We fine-tune a 90M parameter model (Shuster et al., 2020) on each of the good chit-chat candidates with the associated dialogue history as the context from the training set of ACCENTOR-SGD following hyperparameters employed by Roller et al. (2020) and employ the resulting model as the off-the-shelf chit-chat model in Arranger and Rewriter. We use RoBERTa<sub>BASE</sub> (Liu et al., 2019) as the pre-trained language model for Arranger and fine-tune for three epochs with a learning rate of  $2 \times 10^{-5}$  and a batch size of 24.

## 4 Experiments and Discussions

### 4.1 Data Evaluations

**ACCENTOR-SGD.** We first evaluate ACCENTOR at the dataset level, aiming to answer two questions: Q1. Are task-oriented dialogues augmented with good chit-chat more preferred by human judges than the unaugmented? Q2. Does the answer to Q1 depend on how frequently we augment system

	Joint GA		Avg GA		Act-Slot F1		BLEU-4 <sub>orig</sub>		BLEU-4 <sub>aug</sub>	
	All	Seen	All	Seen	All	Seen	All	Seen	All	Seen
SimpleTOD	29.4	79.0	46.9	90.3	61.7	88.9	12.3	17.0	8.0	9.9
SimpleTOD+	29.3	77.4	47.5	90.0	61.5	88.1	11.2	15.3	10.8	12.8
Arranger	29.4	79.0	46.9	90.3	61.7	88.9	9.6	13.5	12.2	14.4
Rewriter	29.4	79.0	46.9	90.3	60.8	86.6	10.4	14.5	11.3	13.3

Table 3: Automatic evaluation results on the test set of ACCENTOR-SGD.

responses with chit-chat? To answer these questions, we randomly sample 100 dialogues from ACCENTOR-SGD, each having at least 8 turns and enough candidates labeled as good for augmenting over 40% of system responses so that we can compare the same task-oriented dialogue with different chit-chat injection frequencies that fall into each of the following four intervals: (0.1, 0.2], (0.2, 0.3], (0.3, 0.4], and (0.4, 1]. Particularly, for the last interval, we augment all system responses that have chit-chat candidates labeled as good, while for the first three intervals, we only augment a randomly selected fraction to fit the interval. We employ ACUTE-Eval (Li et al., 2019) for evaluation, whereby we ask human evaluators to make pairwise comparisons of complete dialogues over four axes: engagingness, interestingness, knowledge, and humanness. We provide the wording of the questions in Appendix A.3.

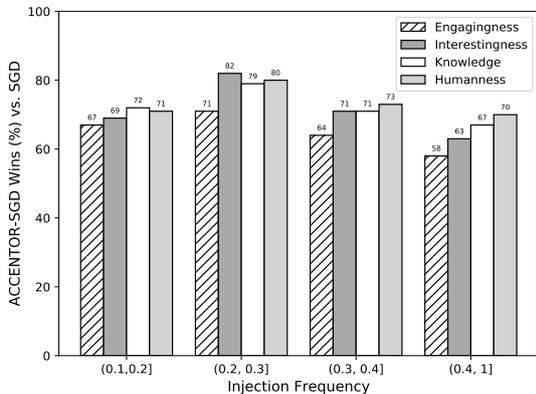


Figure 4: Comparisons between SGD and ACCENTOR-SGD with different injection frequencies at the dataset level using ACUTE-Eval.

As shown in Figure 4, the chit-chat augmented dialogues from ACCENTOR-SGD are more preferred by human judges than the originals over all ACUTE-Eval metrics, regardless of the injection frequency (all p-values < 0.05). Among different injection frequency ranges, (0.2, 0.3] is the best. We offer three hypotheses to explain this finding: (i) (0.2, 0.3] best balances being engaging and not

U:	I like to find some movies directed by Jonathan Levine.
A:	What type of movies you interested?
U:	I like comic movies which is great.
A:	① <i>That's cool.</i> There are 1 movies of your interest. What is your opinion about Long Shot?
U:	Long Shot is a good one.
A:	② <i>I agree.</i> Do you like to run the movie?
U:	I need to run the movie with subtitles.
A:	Please give confirmation details: Running Long Shot with subtitles. ③ <i>Thank you for your help.</i>
U:	Yes. Okay fine.
A:	The movie started playing. ④ <i>It's a great comedy.</i>
U:	I want to make a reservation at a restaurant around there.
A:	Do you have any choice of restaurant and city and what is a booking time you need?
U:	I need it in San Francisco at 17:30 and the restaurant is Aliment.
A:	Please give confirmation details: Table booking at Aliment in San Francisco at 5:30 pm on march 1st and booking is for 2 people. ⑤ <i>Thank you.</i>
U:	Fine. This is okay for me.
A:	Your booking is successful and location address is 786 Bush street. ⑥ <i>I hope you will enjoy the rest of your evening.</i>
U:	Thank you so much for your help.
A:	⑦ <i>You are very welcome.</i> Do you need any other help?
U:	No, thanks a lot.
A:	Have a great day.

Table 4: An augmented dialogue (with injection frequency in (0.4, 1]) that is less preferred than the unaugmented in terms of human evaluation (U: user; A: assistant; *chit-chat* is marked by circled numbers).

too talkative. (ii) There are inevitable annotation errors, and scenarios where whether a candidate is good or bad is subjective. A higher injection frequency means a higher chance of being affected by these factors. (iii) Since candidates are labeled independently, inter-candidate incompatibility may arise (e.g., expressing contradicted preferences), especially when we have a high injection frequency. Table 4 shows a real example to support our hypotheses. Specifically, ③ is labeled as good but is indeed not a suitable (or at least a questionable) candidate, supporting the hypothesis (ii). While ② and ④ are good candidates when we evaluate them separately, they may be less preferred if we assess them jointly because they convey the same meaning: “*Long Shot is a good comedy.*” Having them together may appear incompatible (i.e., repetition) or sound verbose to the user, supporting the hypothesis (i) and (iii).

**ACCENTOR-MultiWOZ.** To investigate the flexibility of our data construction approach, we aug-

		Win %			
		SimpleTOD	SimpleTOD+	Arranger	Rewriter
Loss %	SimpleTOD	–	<b>63</b> **	<b>76</b> **	<b>91</b> **
	SimpleTOD+	<b>37</b> **	–	<b>46</b>	<b>50</b>
	Arranger	<b>24</b> **	<b>54</b>	–	<b>39</b> *
	Rewriter	<b>9</b> **	<b>50</b>	<b>61</b> *	–

(a) Engagingness.

		Win %			
		SimpleTOD	SimpleTOD+	Arranger	Rewriter
Loss %	SimpleTOD	–	<b>63</b> **	<b>73</b> **	<b>70</b> **
	SimpleTOD+	<b>37</b> **	–	<b>47</b>	<b>50</b>
	Arranger	<b>27</b> **	<b>53</b>	–	<b>40</b> *
	Rewriter	<b>30</b> **	<b>50</b>	<b>60</b> *	–

(b) Interestingness.

		Win %			
		SimpleTOD	SimpleTOD+	Arranger	Rewriter
Loss %	SimpleTOD	–	<b>64</b> **	<b>77</b> **	<b>81</b> **
	SimpleTOD+	<b>36</b> **	–	<b>47</b>	<b>55</b>
	Arranger	<b>23</b> **	<b>53</b>	–	<b>45</b>
	Rewriter	<b>19</b> **	<b>45</b>	<b>55</b>	–

(c) Knowledge.

		Win %			
		SimpleTOD	SimpleTOD+	Arranger	Rewriter
Loss %	SimpleTOD	–	<b>68</b> **	<b>71</b> **	<b>82</b> **
	SimpleTOD+	<b>32</b> **	–	<b>51</b>	<b>48</b>
	Arranger	<b>29</b> **	<b>49</b>	–	<b>40</b> *
	Rewriter	<b>18</b> **	<b>52</b>	<b>60</b> *	–

(d) Humanness.

Figure 5: Human evaluation results on the test set of ACCENTOR-SGD using ACUTE-Eval (\*: p-value < 0.05, \*\*: p-value < 0.005).

	MultiWOZ 2.1 vs. ACCENTOR-MultiWOZ (Win %)	ACCENTOR-MultiWOZ (Win %)
Engagingness	<b>10</b> **	<b>90</b> **
Interestingness	<b>11</b> **	<b>89</b> **
Knowledge	<b>13</b> **	<b>87</b> **
Humanness	<b>10</b> **	<b>90</b> **

Figure 6: Comparisons between MultiWOZ 2.1 and ACCENTOR-MultiWOZ at the dataset level using ACUTE-Eval (\*\*: p-value < 0.005).

ment about 1K randomly sampled dialogues from another task-oriented dataset, MultiWOZ 2.1 (Eric et al., 2020) following the same steps as described in Section 2. Crowd workers label 30.0% of the candidates as good, which is lower compared with ACCENTOR-SGD (41.4% in Table 2). We attribute the difference to (i) the performance downgrade of the filter model since we do not re-train the model for MultiWOZ 2.1, and (ii) a higher chance of a chit-chat augmented response being too verbose to be good since the average number of tokens per system turn in MultiWOZ 2.1 is larger than that of SGD (17.3 vs. 13.1). Nevertheless, the augmented version (denoted as ACCENTOR-MultiWOZ) is significantly more preferred than the original, as shown in Figure 6, where we randomly sample 100 dialogues from ACCENTOR-MultiWOZ, augment all of their system responses that have chit-chat candidates labeled as good, and compare these augmented dialogues with the corresponding original dialogues.

## 4.2 Model Evaluations

**Automatic Evaluations.** We consider joint goal accuracy (Joint GA) and average goal accuracy (Avg GA) for evaluating belief states, act-slot F1

for evaluating action decisions, and two BLEU-4 scores (BLEU-4<sub>orig</sub>, BLEU-4<sub>aug</sub>) for evaluating system responses, where we use **original** (resp. **augmented**) system responses as references for BLEU-4<sub>orig</sub> (resp. BLEU-4<sub>aug</sub>). Table 3 summarizes the evaluation results. Since the test set of SGD contains unseen services (i.e., services not seen during training) designed to evaluate the model’s generalizability, we report the results on all services (All) and seen services only (Seen) following Rastogi et al. (2020). Our proposed models generally achieve a similar task performance level compared with the SimpleTOD baseline. Unsurprisingly, the proposed models achieve lower BLEU-4<sub>orig</sub> and higher BLEU-4<sub>aug</sub>.

**Human Evaluations.** We turn to human evaluations for a more comprehensive measure of the response generation performance. We employ the same ACUTE-Eval metrics as we do in data evaluations. We randomly sample 100 dialogues from the test set of ACCENTOR-SGD. For each sampled dialogue  $D = \{u_1, s_1, u_2, s_2, \dots, u_n, s_n\}$ , we pass  $u_1, s_1, \dots, u_i$  to each model  $\mathcal{M} \in \{\text{SimpleTOD}, \text{SimpleTOD+}, \text{Arranger}, \text{Rewriter}\}$  to obtain its system response  $s_i^{\mathcal{M}}$  for the  $i$ -th system turn ( $1 \leq i \leq n$ ). Let  $D^{\mathcal{M}}$  represent  $\{u_1, s_1^{\mathcal{M}}, \dots, u_n, s_n^{\mathcal{M}}\}$ . We ask evaluators to compare each pair of  $D^{\mathcal{M}_1}$  and  $D^{\mathcal{M}_2}$ , where  $\mathcal{M}_1, \mathcal{M}_2 \in \{\text{SimpleTOD}, \text{SimpleTOD+}, \text{Arranger}, \text{Rewriter}\}$  and  $\mathcal{M}_1 \neq \mathcal{M}_2$ . As shown in Figure 5, all of the chit-chat augmented models outperform the SimpleTOD baseline over four ACUTE-Eval metrics. Among the chit-chat augmented models, no one shows a clear win over the other two on the quantitative level. We show a full dialogue example comparing responses gener-

	SimpleTOD (Win %)	vs. Modified Arranger (Win %)
Engagingness	14** ↓10	86** ↑10
Interestingness	25** ↓2	75** ↑2
Knowledge	20** ↓3	80** ↑3
Humanness	20** ↓9	80** ↑9

Figure 7: Human evaluation results of the modified Arranger with controlled injection frequency (\*\*: p-value < 0.005, ↑/↓: increased/decreased win % compared with the original Arranger).

ated by different models along with supplementary discussions in Appendix A.5.

Considering that the injection frequency affects human evaluations (Section 4.1) and that all our models do not explicitly control the injection frequency, we experiment with controlling the injection frequency by modifying Arranger to consider including chit-chat into the current turn only when the injection frequency from the first turn to the current turn is less than 0.3. Compared with the original Arranger, the modified Arranger achieves a higher win percentage over SimpleTOD, as shown in Figure 7. We leave further exploration of injection frequency for future work.

### 4.3 Limitations and Further Discussions

**Approach.** Our proposed strategy to augment task-oriented dialogue system responses with chit-chat is simple, compared with how it emerges in human conversations, where both functionality and engagingness structurally intertwine with each other in a more complex fashion. Our proposed Rewriter model does have a modeling capability to compose both functions organically but is limited due to the dataset’s target arrangement (i.e., concatenation of two separate components). Despite the limitation, our chosen design of “code-separation” has practical merits: we can easily extend the proposed approach to an existing production-level virtual assistant system as a modularized solution, and it has minimal interference to the user-perceived *task success rate*, a core metric widely adapted in virtual assistant systems. Another limitation of our work is that we only augment responses on the system side in our dataset, and the augmentations are independent of each other, whereas in real-life situations, users are also likely to make chit-chat, and the chit-chat between the user and the system should ideally be related to each other. We leave for future research addressing these limitations.

**Evaluation.** We follow the previous literature on evaluation and regard the four ACUTE-Eval metrics as the primary measure of the response generation performance in this work. However, there is a large overlap between the desired quality measured by different human judgment categories used in ACUTE-Eval. The four ACUTE-Eval metrics favor the same dialogue 84.4% of the time in our evaluation, indicating high correlations between these metrics. We leave the study of addressing this issue for future work.

## 5 Related Work

### 5.1 Dialogue Datasets

Dialogue system research has been consistently supported by the development of new datasets. The Dialog State Tracking Challenge (DSTC) series (Williams et al., 2013; Henderson et al., 2014a,b; Williams et al., 2014; Kim et al., 2016, 2017; Moon et al., 2020) provide common testbeds for task-oriented dialogues. Following DSTC, researchers have created a variety of publicly available task-oriented dialogue datasets (El Asri et al., 2017; Shah et al., 2018; Budzianowski et al., 2018; Rastogi et al., 2020). Another line of work seeks to facilitate open-domain chatbot development with large amounts of human-created text data generated in a social context (Baumgartner et al., 2020) and supervision for a variety of desirable general qualities such as being engaging, personable, knowledgeable, and empathetic (Zhang et al., 2018; Dinan et al., 2019; Rashkin et al., 2019; Moon et al., 2019; Wang et al., 2019; Smith et al., 2020). Our work bridges the two lines. We compare ACCENTOR-SGD and ACCENTOR-MultiWOZ with relevant and representative dialogue datasets in Table 5.

Note that very few dialogue corpora contain explicit annotations for both task-oriented and chit-chat utterances. For example, task-oriented dialogue corpora constructed by Rastogi et al. (2020) and Moon et al. (2020) contain annotations for a few chit-chat dialogue acts, but they are limited to light social greetings (e.g., “Thank you!”, “Good Bye.”) typically at the end of each dialogue session. Zhao et al. (2017) propose to artificially augment task-oriented dialogues with randomly sampled utterances from a chit-chat corpus, mainly to improve the out-of-domain recovery performance. Akasaki and Kaji (2017) annotate user utterances with chat/non-chat binary labels. Still, they do not study the contextual combination of these two

Dataset	Construction Method	# Dialogues	Task-Oriented	Chit-Chat
DSTC2 (Henderson et al., 2014a)	crowdsourcing	3, 235	✓	✗
MultiWOZ 2.1 (Eric et al., 2020)	crowdsourcing	10, 438	✓	✗
Schema-Guided Dialogue (Rastogi et al., 2020)	crowdsourcing	22, 825	✓	✗
SIMMC (Moon et al., 2020)	crowdsourcing	12, 948	✓	✗
PersonaChat (Zhang et al., 2018)	crowdsourcing	10, 907	✗	✓
Wizard of Wikipedia (Dinan et al., 2019)	crowdsourcing	22, 311	✗	✓
EmpatheticDialogues (Rashkin et al., 2019)	crowdsourcing	24, 850	✗	✓
BlendedSkillTalk (Smith et al., 2020)	crowdsourcing	6, 808	✗	✓
Pushshift Reddit (Baumgartner et al., 2020)	crawling & scraping	651, 778, 198 <sup>†</sup>	✗	✓
ACCENTOR-SGD (this work)	crowdsourcing	22, 825	✓	✓
ACCENTOR-MultiWOZ (this work)	crowdsourcing	997	✓	✓

Table 5: Statistics of dialogue datasets (<sup>†</sup>: regarding each thread (i.e., a post and its comments) as a dialogue).

to make conversations more engaging, and their corpus does not contain goal labels like typical task-oriented dialogue corpora. In contrast, our work drastically increases the diversity and contextual coverage of chit-chat additions for any task-oriented dialogue corpus (e.g., “*It’s a great way to kick off the summer!*”, “*I hear it’s beautiful.*”).

Compared with other approaches of creating a high-quality dialogue corpus (e.g., via human-to-human “Wizard-of-Oz” collection (Eric et al., 2020), dialogue self-play and paraphrase (Shah et al., 2018)), the annotation cost of the proposed model-based dialogue generation approach combined with the quality control mechanisms is lower, as our work does not involve authoring new sentences by human annotators.

## 5.2 Task-Oriented Dialogue Systems

Over the past few years, neural models have achieved remarkable success in the development of the main components of task-oriented dialogue systems, including understanding user intent, tracking dialogue states, determining system actions, and generating system responses (Henderson et al., 2013; Sun et al., 2014; Wen et al., 2015; Liu and Lane, 2016; Mrkšić et al., 2017; Wen et al., 2017; Nouri and Hosseini-Asl, 2018; Heck et al., 2020; Chen et al., 2020). Recently, connecting separate components and building end-to-end task-oriented neural dialogue systems have attracted increasing interest (Bordes et al., 2017; Peng et al., 2020b). The most recent thread is to unify all components in a single end-to-end neural model by fine-tuning a pre-trained deep language model on multiple tasks, which leads to state-of-the-art performance (Hosseini-Asl et al., 2020; Peng et al., 2020a). We follow this thread and further en-

hance the ability to generate appropriate non-task-oriented add-ons, on top of the ability to achieve functional goals that existing systems are typically narrowly tailored to. A few work have studied training a dialogue model leveraging multiple chit-chat and task-oriented dialogues (Madotto et al., 2019, 2020), which allows the model to attend on a relevant task for a given user utterance and respond accordingly, thus increasing the skill coverage of the model. Our proposed models are trained on the newly collected ACCENTOR-SGD dataset with the turn-level supervision signals, allowing for contextual and flexible code-switching between chit-chat and functional tasks in a single system turn.

## 6 Conclusion

We propose adding chit-chat to enhance task-oriented dialogues (ACCENTOR) in this study. We present a general Human↔AI collaborative data construction approach for ACCENTOR, with which we create a dataset consisting of 23.8K chit-chat augmented task-oriented dialogues. We show via human evaluation that chit-chat augmented dialogues are preferred than the unaugmented. In addition, we propose three models for ACCENTOR. Evaluation results show that compared with the baseline trained on the original unaugmented data, our proposed models trained on the chit-chat augmented counterpart achieve a similar task performance level and higher human evaluation scores.

## Acknowledgements

We thank Gerald Demeunynck for helping with the data annotation process. We would also like to thank the anonymous NAACL reviewers for their constructive and insightful feedback.

## References

- Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. [Towards a human-like open-domain chatbot](#). *arXiv preprint*, cs.CL/2001.09977v3.
- Satoshi Akasaki and Nobuhiro Kaji. 2017. [Chat detection in an intelligent assistant: Combining task-oriented and non-task-oriented spoken dialogue systems](#). In *Proceedings of the ACL*, pages 1308–1319, Vancouver, Canada.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. [The pushshift reddit dataset](#). In *Proceedings of the ICWSM*, volume 14, pages 830–839, Atlanta, GA.
- Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. [Learning end-to-end goal-oriented dialog](#). In *Proceedings of the ICLR*, Toulon, France.
- 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 EMNLP*, pages 5016–5026, Brussels, Belgium.
- Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. [Schema-guided multi-domain dialogue state tracking with graph attention neural networks](#). In *Proceedings of the AAAI*, pages 7521–7528, New York, NY.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. [Wizard of Wikipedia: Knowledge-powered conversational agents](#). In *Proceedings of the ICLR*, New Orleans, LA.
- Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. [Frames: a corpus for adding memory to goal-oriented dialogue systems](#). In *Proceedings of the SIGDIAL*, pages 207–219, Saarbrücken, Germany.
- 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 LREC*, pages 422–428, Marseille, France.
- Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geisshauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. [TripPy: A triple copy strategy for value independent neural dialog state tracking](#). In *Proceedings of the SIGDIAL*, pages 35–44.
- Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014a. [The second dialog state tracking challenge](#). In *Proceedings of the SIGDIAL*, pages 263–272, Philadelphia, PA.
- Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014b. [The third dialog state tracking challenge](#). In *Proceedings of the SLT*, pages 324–329, South Lake Tahoe, NV.
- Matthew Henderson, Blaise Thomson, and Steve Young. 2013. [Deep neural network approach for the dialog state tracking challenge](#). In *Proceedings of the SIGDIAL*, pages 467–471, Metz, France.
- Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. [A simple language model for task-oriented dialogue](#). *arXiv preprint*, cs.CL/2005.00796v3.
- Clayton Hutto and Eric Gilbert. 2014. [VADER: A parsimonious rule-based model for sentiment analysis of social media text](#). In *Proceedings of the ICWSM*, volume 8, pages 216–225, Ann Arbor, MI.
- Seokhwan Kim, Luis Fernando D’Haro, Rafael E Banchs, Jason D Williams, Matthew Henderson, and Koichiro Yoshino. 2016. [The fifth dialog state tracking challenge](#). In *Proceedings of the SLT*, pages 511–517, San Diego, CA.
- Seokhwan Kim, Luis Fernando D’Haro, Rafael E Banchs, Jason D Williams, and Matthew Henderson. 2017. [The fourth dialog state tracking challenge](#). In *Dialogues with Social Robots*, pages 435–449. Springer.
- Margaret Li, Jason Weston, and Stephen Roller. 2019. [ACUTE-EVAL: Improved dialogue evaluation with optimized questions and multi-turn comparisons](#). In *NeurIPS workshop on Conversational AI*, Vancouver, Canada.
- Bing Liu and Ian Lane. 2016. [Attention-based recurrent neural network models for joint intent detection and slot filling](#). In *Proceedings of the Interspeech*, pages 685–689, San Francisco, CA.
- 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*, cs.CL/1907.11692v1.
- Andrea Madotto, Zhaojiang Lin, Yejin Bang, and Pascale Fung. 2020. [The Adapter-Bot: All-in-one controllable conversational model](#). *arXiv preprint*, cs.CL/2008.12579v2.
- Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, Jamin Shin, and Pascale Fung. 2019. [Attention over parameters for dialogue systems](#). In *NeurIPS workshop on Conversational AI*, Vancouver, Canada.
- Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. [ParlAI: A dialog research software platform](#). In *Proceedings of the EMNLP*, pages 79–84, Copenhagen, Denmark.

- Seungwhan Moon, Satwik Kottur, Paul A Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eun-joon Cho, Rajen Subba, and Alborz Geramifard. 2020. *Situated and interactive multimodal conversations*. *arXiv preprint*, cs.CL/2006.01460v1.
- 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 ACL*, pages 845–854, Florence, Italy.
- Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. *Neural belief tracker: Data-driven dialogue state tracking*. In *Proceedings of the ACL*, pages 1777–1788, Vancouver, Canada.
- Elnaz Nouri and Ehsan Hosseini-Asl. 2018. *Toward scalable neural dialogue state tracking model*. In *NeurIPS workshop on Conversational AI*, Montreal, Canada.
- Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020a. *SOLOIST: Few-shot task-oriented dialog with a single pre-trained auto-regressive model*. *arXiv preprint*, cs.CL/2005.05298v3.
- Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiu-jun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020b. *Few-shot natural language generation for task-oriented dialog*. *arXiv preprint*, cs.CL/2002.12328v1.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. *Language models are unsupervised multitask learners*. *Preprint*, available at <https://openai.com/blog/better-language-models/>.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. *Towards empathetic open-domain conversation models: A new benchmark and dataset*. In *Proceedings of the ACL*, pages 5370–5381, Florence, Italy.
- 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*, pages 8689–8696, New York, NY.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. *Recipes for building an open-domain chatbot*. *arXiv preprint*, cs.CL/2004.13637v2.
- Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. *Building a conversational agent overnight with dialogue self-play*. *arXiv preprint*, cs.CL/1801.04871v1.
- Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. *The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents*. In *Proceedings of the ACL*, pages 2453–2470.
- Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. *Can you put it all together: Evaluating conversational agents’ ability to blend skills*. In *Proceedings of the ACL*, pages 2021–2030.
- Kai Sun, Lu Chen, Su Zhu, and Kai Yu. 2014. *The SJTU system for dialog state tracking challenge 2*. In *Proceedings of the SIGDIAL*, pages 318–326, Philadelphia, PA.
- Xuwei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. *Persuasion for good: Towards a personalized persuasive dialogue system for social good*. In *Proceedings of the ACL*, pages 5635–5649, Florence, Italy.
- Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. *Semantically conditioned LSTM-based natural language generation for spoken dialogue systems*. In *Proceedings of the EMNLP*, pages 1711–1721, Lisbon, Portugal.
- Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve Young. 2017. *Latent intention dialogue models*. In *Proceedings of the ICML*, volume 70, pages 3732–3741, Sydney, Australia.
- Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. *The dialog state tracking challenge*. In *Proceedings of the SIGDIAL*, pages 404–413, Metz, France.
- Jason D Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, and Deepak Ramachandran. 2014. *The dialog state tracking challenge series*. *AI Magazine*, 35(4):121–124.
- Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. 2016. *Strategy and policy learning for non-task-oriented conversational systems*. In *Proceedings of the SIGDIAL*, pages 404–412, Los Angeles, CA.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. *Personalizing dialogue agents: I have a dog, do you have pets too?* In *Proceedings of the ACL*, pages 2204–2213, Melbourne, Australia.
- Tiancheng Zhao, Allen Lu, Kyusong Lee, and Maxine Eskenazi. 2017. *Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability*. In *Proceedings of the SIGDIAL*, pages 27–36, Saarbrücken, Germany.

## A Appendix

### A.1 Details of Candidate Generation

We summarize model configurations in Table 6, which are employed together for candidate generation in Section 2.1. Our implementation is based on ParlAI (Miller et al., 2017), and all unspecified parameters take the default values set in the interactive mode of ParlAI.

Generative Model	Beam Size	Minimum Beam Length
BlenderBot (90M)	10	1
BlenderBot (90M)	10	5
BlenderBot (90M)	10	20
BlenderBot (90M)	30	20
BlenderBot (2.7B)	10	1
BlenderBot (2.7B)	10	5
BlenderBot (2.7B)	10	20
BlenderBot (2.7B)	30	20
BlenderBot (9.4B)	10	1
BlenderBot (9.4B)	10	20
GPT-2 (117M)	10	1
GPT-2 (345M)	1	1
GPT-2 (345M)	10	1
GPT-2 (345M)	10	5
GPT-2 (345M)	10	20
GPT-2 (345M)	30	1
GPT-2 (762M)	10	1

Table 6: Employed models and decoding parameters for candidate generation.

### A.2 Details of Candidate Filtering

The ranker initially ranks each candidate according to the posterior probability output by the binary classifier. It then lowers the ranks of candidates that match a list of bad patterns. Most bad patterns are about newly introduced counterfeit information (e.g., containing an URL/email address, a phone number, time, or amount of money). The rest of the bad patterns are mainly about text genre (e.g., containing email sign-offs such as “best regards”) and format (e.g., misuse of punctuation marks). Lastly, the ranker raises the ranks of (i) uncommon candidates and (ii) candidates that are dissimilar to the other candidates for the dialogue and the system response being augmented. We measure the similarity by Levenshtein distance.

Note that we do not explore the optimal settings for candidate filtering, as it is not the primary focus of this paper. For instance, how much the rule-based ranker lowers or raises the ranks of candidates is set manually based on engineering intuition rather than rigorous analysis; we do not exhaustively investigate how much labeled data is required to obtain a good enough binary classifier; the 1.7K examples from the pilot annotation are randomly sampled. Tuning the procedure (e.g., the

number and selection of training examples) may lead to a better resulting candidate set.

### A.3 Human Evaluation Questions

- Engaging: *Who would you prefer to talk to? Which version is more likely to hold your attention and make you want to hear more?*
- Interesting: *Who would you say is more interesting? Which version arouses your curiosity or tells you something new or useful?*
- Humanlike: *Who would you say sounds more human? Which version is more natural and personable?*
- Knowledgeable: *Who would you say is more knowledgeable? Which version seems more well informed and confident in the information?*

## A.4 Example Dialogues

Utterance	Task Labels
U: I'm starting to get hungry. I want to go eat at a nice restaurant.	INFORM_INTENT [intent: FindRestaurants]
A: Sure. What city would you like to eat in?	REQUEST [slot: city]
U: Milpitas.	INFORM [city: Milpitas]
A: <u>(X) That sounds delicious.</u> Do you have a particular type of food you are craving? Mexican? American? Something else?	REQUEST [cuisine: Mexican, American]
U: I feel like Cambodian tonight.	INFORM [cuisine: Cambodian]
A: <u>(✓) I hear the food is amazing.</u> There is a Cambodian restaurant in Milpitas called Anh Hong. <u>(✓) It's a great place to eat.</u>	OFFER [restaurant_name: Anh Hong, city: Milpitas]
U: That's exactly what I'm looking for.	SELECT
A: Do you want me to make a reservation for you? <u>(X) I can also recommend some other restaurants in the area.</u>	OFFER_INTENT [intent: ReserveRestaurant]
U: Yes, I'd like to try this place out.	AFFIRM_INTENT
A: Okay. At what time would you like to eat there?	REQUEST [slot: time]
U: 6:30 pm. Please make the reservation for four people.	INFORM [time: 6:30 pm, party_size: 4]
A: <u>(X) Okay .</u> Sure. A reservation at Anh Hong in Milpitas for 4 at 6:30 pm today. Is that correct?	CONFIRM [restaurant_name: Anh Hong, city: Milpitas, time: 6:30 pm, party_size: 4]
U: Perfect. What are the prices at Anh Hong like?	REQUEST [slot: price_range]
A: They have moderate pricing. Your reservation has been made. <u>(✓) Thank you.</u> <u>(X) Do you want dessert?</u>	INFORM [price_range: moderate]
U: Thanks. You've been a big help.	THANK_YOU
A: <u>(✓) You are welcome. I hope you enjoy your meal.</u> Have a nice day. <u>(X) I hope you enjoy your meal. Do you have any plans for the rest of the day?</u>	GOODBYE
U: Can you find me some songs? Any album is fine.	INFORM_INTENT [intent: LookupMusic, album: dont-care]
A: I have found 10 songs you may like. How does A Little Bit Stronger by Sara Evans in the album Stronger sound?	INFORM_COUNT [count: 10], OFFER [artist: Sara Evans, song_name: A Little Bit Stronger]
U: That's great.	SELECT
A: <u>(X) I'll have to check it out.</u> <u>(✓) I'm glad you like it.</u> What else can I help with? <u>(✓) I love helping people find new music.</u>	REQ_MORE
U: I want to play the song.	INFORM_INTENT [intent: PlayMedia]
A: Should I play A Little Bit Stronger on the TV? <u>(✓) I think it would be a great song to play on the television. What do you think?</u> <u>(✓) That's my favorite song by Sarah Evans.</u>	CONFIRM [playback_device: TV, song_name: A Little Bit Stronger]
U: Yes, that's correct.	AFFIRM
A: Started playing the song. <u>(X) I can tell you about the lyrics.</u> <u>(X) What else do you want to know about it?</u>	NOTIFY_SUCCESS
U: Thanks, that's all I need.	THANK_YOU
A: <u>(✓) You're welcome.</u> Have a good day. <u>(X) What do you like?</u> <u>(X) I hope you find some good music to listen to. I like all kinds of music. What do you like?</u>	GOODBYE

Table 7: Example dialogues from ACCENTOR-SGD (U: user; A: assistant). All *chit-chat* candidates for augmentation, generated with the state-of-the-art pre-trained language models, are annotated by the crowd workers with good (✓) and bad (X) labels. Note that while most of the bad chit-hat candidates are *fluent*, they are often contextually inappropriate or inconsistent with the rest of the dialogue. The annotation guideline is highlighted in Section 2.3.

## A.5 Generation Samples

U: I'm looking for something interesting to do, please find an activity for me on the 5th of this month.	ST: What type of events do you prefer, maybe [events_category] or [events_category] or anything else? Also, what city should I search in? S+: Sure, I can assist you. So, where are you looking for an activity, and do you have a type of activity you prefer? You could say music, sports, or something like that. AR: What type of events do you prefer, maybe [events_category] or [events_category] or anything else? Also, what city should I search in? RW: What type of events do you prefer, maybe [events_category] or [events_category] or anything else? Also, what city should I search in?
U: I'd like to search for an activity around Seattle, WA.	ST: Is there any particular activity you are interested in? S+: What type of activity would you like to take part in? There are a lot of activities to be found in the city. AR: Seattle is a great place to visit. Is there any particular activity you are interested in? RW: Seattle is a great place to visit. Is there any particular activity you are interested in?
U: I'd like to do something Music related.	ST: I found [events_count] events that you might like. There's a [events_event_name] activity taking place at [events_event_location]. It's happening [events_date] starting at [events_time]. S+: I discovered [events_count] activities you may enjoy. There is an [events_event_name] activity at [events_event_location] on [events_date], starting at [events_time]. AR: I found [events_count] events that you might like. There's a [events_event_name] activity taking place at [events_event_location]. It's happening [events_date] starting at [events_time]. RW: It's a great place to see a concert. I found [events_count] events that you might like. There's a [events_event_name] activity taking place at [events_event_location]. It's happening [events_date] starting at [events_time].
U: That seems fun, please book one ticket for me to that event.	ST: Confirm the following details for me: Booking [events_number_of_seats] ticket for the [events_event_name] show on [events_date] in [events_city_of_event]. S+: Let me see if I got everything correctly: You'd like [events_number_of_seats] ticket for the [events_event_name] show in [events_city_of_event] on [events_date]. AR: Confirm the following details for me: Booking [events_number_of_seats] ticket for the [events_event_name] show on [events_date] in [events_city_of_event]. RW: I'm sure you'll enjoy it. Confirm the following details for me: Booking [events_number_of_seats] ticket for the [events_event_name] show on [events_date] in [events_city_of_event].
U: Yes, that is correct. Please tell me the venue's address.	ST: I have reserved your ticket. The venue is located at [events_address_of_location]. S+: Reservation is confirmed. The venue is located at [events_address_of_location]. AR: I have reserved your ticket. The venue is located at [events_address_of_location]. RW: I have reserved your ticket. The venue is located at [events_address_of_location].
U: Okay, please find me a ride to the event. A shared ride is fine.	ST: Please confirm that you would like to travel to [ridesharing_destination] in a cab for [ridesharing_number_of_riders] person and that you would like a shared ride. S+: How many people will be riding in the cab? AR: Please confirm that you would like to travel to [ridesharing_destination] in a cab for [ridesharing_number_of_riders] person and that you would like a shared ride. RW: Please confirm that you would like to travel to [ridesharing_destination] in a cab for [ridesharing_number_of_riders] person and that you would like a shared ride.
U: Yes that is correct. How long will it take to get there.	ST: Your ride is booked and on the way. The ride is expected to take [ridesharing_approximate_ride_duration] minutes. S+: Your ride is confirmed and is on its way. The ride is expected to take [ridesharing_approximate_ride_duration] minutes. AR: Your ride is booked and on the way. The ride is expected to take [ridesharing_approximate_ride_duration] minutes. RW: Your ride is booked and on the way. The ride is expected to take [ridesharing_approximate_ride_duration] minutes.
U: What's the cost of the ride?	ST: The cost is [ridesharing_ride_fare]. S+: The price is [ridesharing_ride_fare]. AR: The cost is [ridesharing_ride_fare]. RW: The cost is [ridesharing_ride_fare].
U: Thanks, I don't need anything else.	ST: Enjoy your day. S+: Enjoy the rest of your day. If you have any questions, please don't hesitate to contact me. AR: You are welcome. Enjoy your day. RW: You are welcome. Enjoy your day.

Table 8: Sampled responses generated for a complete task-oriented dialogue (U: user, ST: SimpleTOD, S+: SimpleTOD+, AR: Arranger, RW: Rewriter).

As shown in Table 8, we observe that compared with SimpleTOD+, both Arranger and Rewriter tend to add chit-chat to the beginning of task-oriented responses. This is perhaps because the underlying off-the-shelf chit-chat model takes only  $u_1, s_1, \dots, u_i$  as input, making it more likely to generate a suitable chit-chat to start, rather than end the  $i$ -th system turn. The responses generated by Arranger and Rewriter are similar because Rewriter generates responses by copying contents from the responses output by the underlying off-the-shelf models without modification for most of the time (87.0% of dialogues on the test set).