KPT: Keyword-Guided Pre-training for Grounded Dialog Generation

Qi Zhu[1*], **Fei Mi**[2†], **Zheng Zhang**[1], **Yasheng Wang**[2], **Yitong Li**[2],
Xin Jiang[2], **Qun Liu**[2], **Xiaoyan Zhu**[1], **Minlie Huang**[1†]
[1]CoAI Group, DCST, IAI, BNRIST, Tsinghua University
[2]Huawei Noah's Ark Lab
zhu-q18@mails.tsinghua.edu.cn, {z-zhang,zxy-dcs,aihuang}@tsinghua.edu.cn
{mifei2,wangyasheng,liyitong3,Jiang.Xin,qun.liu}@huawei.com

Abstract

Incorporating external knowledge into the response generation process is essential to building more helpful and reliable dialog agents. However, collecting knowledge-grounded conversations is often costly, calling for a better pre-trained model for grounded dialog generation that generalizes well w.r.t. different types of knowledge. In this work, we propose **KPT** (**K**eyword-guided **P**re-**T**raining), a novel self-supervised pre-training method for grounded dialog generation without relying on extra knowledge annotation. Specifically, we use a pre-trained language model to extract the most uncertain tokens in the dialog as *keywords*. With these keywords, we construct two kinds of knowledge and pre-train a knowledge-grounded response generation model, aiming at handling two different scenarios: (1) the knowledge should be faithfully grounded; (2) it can be selectively used. For the former, the grounding knowledge consists of keywords extracted from the response. For the latter, the grounding knowledge is additionally augmented with keywords extracted from other utterances in the same dialog. Since the knowledge is extracted from the dialog itself, KPT can be easily performed on a large volume and variety of dialogue data. We considered three data sources (open-domain, task-oriented, conversational QA) with a total of 2.5M dialogues. We conduct extensive experiments on various few-shot knowledge-grounded generation tasks, including grounding on dialog acts, knowledge graphs, persona descriptions, and Wikipedia passages. Our comprehensive experiments and analyses demonstrate that KPT consistently outperforms state-of-the-art methods on these tasks with diverse grounding knowledge.

1 Introduction

Grounding the response generation on external knowledge is a necessary skill of an engaging and reliable dialog agent (Shuster et al. 2021; Thoppilan et al. 2022; Shuster et al. 2022). Such a dialog agent could handle various scenarios by grounding on relevant knowledge such as common sense knowledge graphs (Zhou et al. 2018), persona descriptions (Zhang et al. 2018), and Wikipedia passages (Dinan et al. 2019). However, the collection of human-annotated knowledge-grounded dialogs is expensive, which highlights

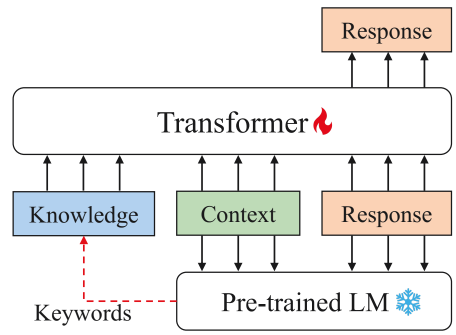

Figure 1: An illustration of *keyword-guided pre-training*. We extract the most uncertain tokens in a dialog (measured by the bottom frozen pre-trained LM) as keywords and create pseudo-knowledge from these keywords. The pre-training objective (upper) is to generate the response according to the context and knowledge.

the model's few-shot learning ability to ground on new kinds of knowledge. To this end, we propose **K**eyword-guided **P**re-**T**raining (KPT), a self-supervised pre-training method that only needs dialogs without any extra knowledge annotation but benefits multiple downstream dialog generation tasks grounded on different knowledge sources.

Most of the previous dialog pre-training methods aim to capture the coherence of dialog, such as predicting the next utterance (Zhang et al. 2020), masked utterance (Mehri et al. 2019), and correct utterance order (Gu, Yoo, and Ha 2021). However, for grounded dialog generation, learning to incorporate external knowledge for a given dialog context is more critical. Recently proposed GODEL (Peng et al. 2022) utilizes several annotated grounded dialog corpora for pre-training by concatenating the knowledge and context as input. In contrast, our proposed KPT can be performed on any dialog corpus without requiring grounding annotation, which could benefit more downstream tasks by including more diverse dialogs in pre-training.

Pre-training a general knowledge-grounded dialog generation model faces two main challenges: the diverse knowledge types of downstream tasks and the lack of large annotated knowledge-grounded dialog datasets. To address these challenges, we regard the keywords as the basic unit

[*]Work done during an internship at Huawei Noah's Ark Lab.
[†]Corresponding authors.

of knowledge and mine the keywords from the dialog itself as Feng et al. (2021). As shown in Figure 1, we extract the most uncertain tokens measured by a pre-trained language model (LM) as keywords and construct pseudo-knowledge input using these keywords. We identify two scenarios of grounded dialog generation and design two pre-training tasks accordingly which only differ in knowledge construction: (1) models need to faithfully ground on the given knowledge, where we use keywords from the response as the knowledge; (2) models need to selectively use part of (or even none of) the knowledge based on the context, where we use keywords sampled from other turns of the same dialog along with (or not) keywords from the response as the knowledge. We pre-train the models jointly on these two tasks on 2.5M dialogs covering open-domain dialogs, task-oriented dialogs, and conversational question answering. To verify the effectiveness of KPT, we conduct few-shot fine-tuning experiments on four grounded dialog generation tasks where the knowledge sources include dialog acts, knowledge graphs, persona descriptions, and Wikipedia passages. We show that KPT consistently outperforms GODEL and standard response generation pre-training. We further analyze the effects of each pre-training task, keyword ratio, and the pre-trained LM used for keyword extraction.

In summary, our contributions are:

1. We propose KPT, a novel self-supervised dialog pre-training method for grounded dialog generation, which does not need any extra knowledge annotation.
2. We identify two scenarios of grounded dialog generation and design corresponding pre-training objectives to handle them in a unified framework.
3. Experiments on diverse grounded dialog generation tasks with different knowledge sources show that our proposed KPT consistently outperforms state-of-the-art baselines.

2 Related Work

2.1 Pre-training

Pre-training is a widely used technique to transfer knowledge from a source domain to a target domain and is especially effective when the target domain data is limited. In recent years, big models pre-trained on large-scale text corpora have been applied to various downstream tasks and achieved remarkable success (Devlin et al. 2019; Radford et al. 2019; Raffel et al. 2020; Lewis et al. 2020). To further improve the performance on dialog modeling, researchers perform pre-training on dialog corpora to reduce domain gap (Zhang et al. 2020; Freitas et al. 2020; Roller et al. 2021; Wu et al. 2020; Mi et al. 2022) and devise novel pre-training objectives to capture dialog-specific features (Mehri et al. 2019; Gu, Yoo, and Ha 2021; Xu and Zhao 2021). However, most proposed self-supervised pre-training objectives focus on learning the internal characteristics of dialog data instead of how to incorporate external knowledge into the dialog.

To learn a knowledge-grounded dialog generation model, existing works either use human-annotated knowledge-grounded conversation corpora for training (Thoppilan et al. 2022; Peng et al. 2021, 2022; He et al. 2022) or augment raw dialogs with retrieved relevant knowledge (Ghazvininejad et al. 2018; Shuster et al. 2021; Li, He, and Mi 2022; Wang et al. 2022; Zhang et al. 2022). However, they all cover limited domains and knowledge types, questioning the transferability of pre-trained models to new domains and different types of knowledge. Recently proposed GODEL (Peng et al. 2022) is the most related work to ours, which introduces an additional grounded pre-training phrase between general pre-training and fine-tuning, utilizing several grounded dialog corpora including question answering (grounded on passage), chit-chat (grounded on world knowledge), and task-oriented dialog (grounded on domain knowledge). In contrast, our proposed pre-training method is more general without requiring such knowledge annotation and is demonstrated to be more effective in diverse grounded dialog generation tasks with different types of knowledge.

2.2 Grounded Dialog Generation

Grounding response generation on different types of knowledge is a vital ability for an informative, engaging, and reliable dialog agent. Grounding knowledge could be common sense (Zhou et al. 2018), persona (Zhang et al. 2018), world knowledge (Dinan et al. 2019), domain knowledge (Eric et al. 2017), etc., in various forms ranging from structured knowledge such as dialog act (Wen et al. 2015), knowledge graph (Zhou et al. 2018), and table (Eric et al. 2017), to unstructured knowledge such as keyword (Zhong et al. 2021), sentence (Song et al. 2018), and document (Dinan et al. 2019). Since collecting natural knowledge-grounded conversation is often costly, learning a grounded dialog generation model with a few samples is a challenging but important topic. However, the heterogeneity of knowledge leads to separated communities and specialized models (Zhou et al. 2018; Wen et al. 2015; Zhao et al. 2019), preventing the study of general grounded dialog generation models. Recently, Xie et al. (2022) propose to unify a series of structured knowledge grounded text generation tasks by serializing the input and output and applying the same text-to-text transformer. Inspired by them, we unify the downstream tasks in a similar way. But we focus on improving diverse grounded response generation tasks via pre-training and our proposed KPT does not need extra knowledge annotation.

3 Method

3.1 Overview

Given the dialog context $C_t = [U_1, U_2, ..., U_{t-1}]$, the goal of a knowledge-grounded dialog agent is to generate the next utterance U_t grounded on knowledge K_t. The knowledge K_t could be persona descriptions (Zhang et al. 2018), dialog acts (Wen et al. 2015), knowledge graph (Zhou et al. 2018), etc. Different grounding tasks may have different ways to retrieve relevant knowledge K_t according to the context C_t, such as rules, domain APIs, and specialized knowledge retrievers. Therefore, we separate the knowledge retrieval from knowledge-grounded response generation and focus on using pre-training to improve the latter (i.e. $P(U_t|C_t, K_t)$) where K_t is already available.

Dataset	Dialogs	Samples	Avg. Tok.	non-SW%
DailyDialog	11K	76K	11.2	46.4
SGD	16K	309K	10.4	52.1
Taskmaster-1	11K	211K	8.8	52.8
Taskmaster-2	14K	218K	9.5	53.7
Taskmaster-3	19K	360K	10.6	53.7
MetaLWOZ	34K	323K	7.5	49.6
Reddit	1.4M	5.2M	15.1	51.9
WikiDialog	1.0M	4.7M	18.7	61.4
Total	2.5M	11.4M		

Table 1: Statistics of pre-training datasets. "Avg. Tok." is the average tokens per turn. "non-SW%" is the average ratio of non-stop words in a turn. For WikiDialog, we only use system turns that come from wiki passages.

To utilize large-scale dialog corpora without knowledge annotation for pre-training, we regard the keyword as the basic unit of knowledge and mine keywords from dialogs directly without any external knowledge sources or annotation efforts. Inspired by Feng et al. (2021), we extract keywords in a dialog according to their LM loss judged by a pre-trained LM (§3.3). In §3.4, we identify two kinds of knowledge usages on downstream tasks: 1) express all the knowledge in the response; 2) only select context-relevant knowledge, if any, to compose the response. Thus, we devise two keyword-guided pre-training objectives that mimic the two kinds of knowledge usages respectively: generate the response according to the context and 1) keywords in the response or 2) randomly sampled keywords in the dialog, as illustrated in Table 2. Following Xie et al. (2022) and Peng et al. (2022), we concatenate serialized knowledge with the context as the input and use T5 (Raffel et al. 2020) to generate the response for both pre-training and fine-tuning.

3.2 Pre-training Datasets

As shown in Table 1, our pre-training datasets include DailyDialog (Li et al. 2017), Schema-Guided Dialog (Rastogi et al. 2020), Taskmaster-1/2/3 (Byrne et al. 2019, 2021), MetaLWOZ (Li et al. 2020), DSTC8-Reddit (Lee et al. 2019), and WikiDialog (Dai et al. 2022), covering chit-chats, goal-oriented dialogs, and information seeking dialogs. We filter out the Reddit dialogs that contain URLs and dialogs that contain too long utterances (>256 tokens) in Reddit and WikiDialog. WikiDialog is an artificial dataset containing 11M information-seeking dialogs where user utterances are generated by a model. Therefore, we only use system turns for pre-training to ensure data quality. Since WikiDialog is much larger than other pre-training datasets, we randomly choose 1M dialogs for pre-training to avoid an imbalanced training set. Reddit and WikiDialog have comparable samples and constitute most of the pre-training data. Nearly 50% tokens are non-stop words across different datasets.

3.3 Keyword Extraction

As there is no golden standard for keyword definition, we propose to extract words that convey information that could hardly be inferred from the context directly. To this end, we select the most unpredictable words according to the context as the keywords. Specifically, we concatenate the utterances $U_t = w_1^t w_2^t ... w_{|U_t|}^t$ in a dialog as one sequence $U_1 U_2 ... U_T$ and sort the words according to their conditional generation probability given by a pre-trained language model:

$$P_{LM}(w_i^t | C_t, w_{<i}^t) \tag{1}$$

If a word is split into multiple tokens, its generation probability is the geometric average of these tokens' generation probabilities. Then, we select the top α words (α is set to 30% empirically) that have the lowest generation probability and are not stop words as keywords. We ignore stop words since they often carry little information. Finally, we merge adjacent keywords into ones and group the keywords by their belonging sentences to preserve the information of which keywords should be expressed together. We denote keywords in a utterance U that consists of N sentences as $KW = \{\{k_1^1, k_2^1, ...\}, ..., \{k_1^N, k_2^N, ...\}\}$, where k_*^i are keywords in the i-th sentence.

Algorithm 1: Prepare keyword-guided pre-training data

Input: Dialog corpus $\mathcal{D}_{cov}$, Pre-trained LM ϕ
Parameter: Keyword ratio α
Output: Pre-training dataset $\mathcal{D}_{pre}$

```
Initialize D_pre as an empty list
for dialog D = U_1 U_2 ... U_T in D_cov do
    Extract keywords as stated in §3.3 with φ and α
    for context-response pair (C_t, U_t) in D do
        K_g ← KW_t
        Add (C_t, K_g, U_t) to D_pre
        Initialize K_n as an empty list
        Sample random integer j in range [1, 5]
        Sample min(j, T − 1) turns from all {U_i}_{i≠t} and
        add their keywords KW_i to K_n
        if r ~ uniform(0, 1) < 0.8 then
            Add KW_t to K_n
        end if
        Add (C_t, K_n, U_t) to D_pre
    end for
end for
return D_pre
```

3.4 Keyword-guided Pre-training

For each context-response pair (C_t, U_t), we construct two kinds of knowledge inputs: golden keywords K_g from the response U_t and noisy keywords K_n from the whole dialog $U_1 U_2 ... U_T$, as shown in Algorithm 1. The former simulates the knowledge that should be faithfully grounded on (e.g., dialog acts, relational triples) while the latter simulates the knowledge that should be selectively used (e.g., document, database, persona). As illustrated in Table 2, we use different prompts to distinguish these two kinds of knowledge. Then we mix (C_t, K_g, U_t) and (C_t, K_n, U_t) in a ratio of 1:1 for keyword-guided response generation pre-training.

Response Generation with Golden Keywords This task is designed for the situation where the given knowledge must

Knowledge Source	Knowledge	Input	Output
Golden keywords (pre-train)	[["CoD", "hit"], ["trying", "become battlefield"]]	generate a response: **grounded knowledge:** | **CoD : hit** | **trying : become battlefield** | context: ... user: Yes, weapons now fire projectiles instead of hit scans. system:	But why, **CoD** has always been **hit** scan. Is CoD **trying** to **become battlefield**?
Dialog acts (Multi-WOZ)	{"inform-hotel": {"choice": "9"}, "recommend-hotel": {"name": "Autumn House"}}	generate a response: **grounded knowledge:** | **inform-hotel : choice : 9** | **recommend-hotel : name : Autumn House** | context: user: A guesthouse, please. Are there any in the cheap price range? system:	Oh, yes, there are **9**. I do recommend the **Autumn House**.
Knowledge graph (OpenDialKG)	[["Stranger in a Strange Land", "has_genre", "Science Fiction"]]	generate a response: **grounded knowledge:** | **Stranger in a Strange Land : has_genre : Science Fiction** | context: user: I like Stranger in a Strange Land. Can you recommend something for me? system:	Sure, do you like this book because it is the **genre** of **science fiction**?
Noisy keywords (pre-train)	[["Bullet travel"], ["CoD", "hit"], ["trying", "become battlefield"],...]	generate a response: **all knowledge:** | **Bullet travel** | **CoD : hit** | **trying : become battlefield** | **...** | context: ... user: Yes, weapons now fire projectiles instead of hit scans. system:	But why, **CoD** has always been **hit** scan. Is CoD **trying** to **become battlefield**?
Persona (PersonaChat)	["i have one cat.", "i am a kindergarten teacher.", ...]	generate a response: **all knowledge:** | **i have one cat.** | **i am a kindergarten teacher.** | **...** | context: ... user: I have a dog. his name is max. system:	That is cute I **have a cat**.
Wikipedia (WoW)	["Radiohead are an English rock band from Abingdon, Oxfordshire, formed in 1985.", ...]	generate a response: **all knowledge:** | **Radiohead are an English rock band from Abingdon, Oxfordshire, formed in 1985.** | **...** | context: user: I just heard a song by Radiohead. It was great! system:	I love Radiohead! It's hard to believe that they've **been together since 1985**!

Table 2: Examples for pre-training and downstream tasks, categorized into two groups according to how to use the knowledge. In each row, we show the grounding knowledge, model input (serialized knowledge and context), and the golden response. We use "*grounded knowledge*" to prompt the knowledge that should be faithfully grounded (the first three rows) and "*all knowledge*" to prompt the knowledge that should be selectively used (the last three rows).

appear in the response. We regard the keywords KW_t extracted from the response U_t as golden keywords K_g. To serialize K_g, we use ":" to join keywords in a sentence as a group and then use "|" to join keyword groups, both in random order. As shown in Table 2, the input of the model is the concatenation of K_g and C_t while the output is U_t. The loss function is:

$$\mathcal{L}_{golden}(\theta) = -\sum_{i=1}^{|U_t|} \log p_\theta(w_i^t | C_t, w_{<i}^t, K_g) \qquad (2)$$

Response Generation with Noisy Keywords In some cases (Zhang et al. 2018; Dinan et al. 2019), we only have a set of relevant knowledge instead of the exact grounding knowledge, where the model needs to decide whether to use knowledge and which knowledge to use. To simulate these cases, we construct noisy keywords K_n which consists of keywords from multiple turns in the same dialog to serve as "negative" knowledge, as detailed in Algorithm 1. To model the scenario where knowledge is not needed, KW_t is only included in K_n in 80% of the time. After mixing keyword groups from different turns, the serialization of K_n is the same as K_g. The loss function is:

$$\mathcal{L}_{noisy}(\theta) = -\sum_{i=1}^{|U_t|} \log p_\theta(w_i^t | C_t, w_{<i}^t, K_n) \qquad (3)$$

4 Experimental Setup

4.1 Evaluation Datasets and Metrics

We fine-tune models with KPT on four grounded dialog generation datasets covering various kinds of knowledge:

- **MultiWOZ 2.1 (MWOZ)** (Budzianowski et al. 2018; Eric et al. 2020) is a widely used multi-domain task-oriented dialog dataset. We conduct natural language generation (NLG) experiments on this dataset, where the response should strictly follow the given dialog acts. We treat each dialog act that consists of intent, slot, and value as a keyword group. Since the dialog acts already contain sufficient information, we use the latest three turns instead of the whole history as context to save computation.
- **OpenDialKG (ODKG)** (Moon et al. 2019) is an open-domain dialog dataset grounded on a knowledge graph (KG). Each turn is manually annotated with KG paths that link previously mentioned entities to the entities in this turn. For serialization, we treat each relational triple that consists of head entity, relation, and tail entity as a keyword group. Since the original data partition is not publicly available, we randomly split the data into training (70%), validation (15%), and test set (15%).
- **PersonaChat (PC)** (Zhang et al. 2018) is an open-domain dialog dataset where each speaker chats accord-

ing to a pre-defined persona described by five sentences. We directly use "|" to join all persona sentences as the knowledge input.

- **Wizard of Wikipedia (WoW)** (Dinan et al. 2019) is an open-domain dialog dataset where system utterances are grounded on passages retrieved from Wikipedia. Each passage has a few sentences. Similar to PersonaChat, we use "|" to join all sentences as the knowledge input. We merge the original seen and unseen test sets into one for evaluation.

Grounding knowledge of MultiWOZ and OpenDialKG is similar to the golden keywords in pre-training, while grounding knowledge of PersonaChat and WoW is similar to the noisy keywords. Therefore, we use the corresponding prompt for each task and show the example in Table 2.

We evaluate generated responses in two aspects: similarity to reference responses and knowledge utility. For the former, we use perplexity (PPL), corpus-level BLEU-4 (Papineni et al. 2002; Post 2018), unigram F1 for non-stop words (Miller et al. 2017), and Rouge-L (Lin 2004). When it comes to knowledge utility, it measures how much the model utilizes the provided grounding knowledge, and the metrics are dataset-specific. For MultiWOZ, previous works (Wen et al. 2015) use Slot Error Rate (SER) to evaluate the ratio of missing and redundant slots in generated responses. We use $1-$SER to be consistent with knowledge utility metrics of other datasets where bigger numbers are better. For OpenDialKG, we evaluate micro entity F1 between generated responses and references, where a turn's candidate entities are the head and tail entities of grounding relational triples. For PersonaChat and WoW, we calculate unigram F1 for non-stop words between generated responses and the concatenation of knowledge sentences.

4.2 Baselines and Training Details

Since our goal is to verify the effect of keyword-guided pre-training, we consider following baselines that are based on the same model architecture but pre-trained differently:

- **T5** (Raffel et al. 2020): Our backbone pre-trained model.
- **T5-RG**: We further pre-train T5 on our pre-training data to generate a response according to the context. It can be viewed as a T5 version of DialoGPT (Zhang et al. 2020).
- **GODEL** (Peng et al. 2022): Initialized by T5, GODEL is first trained with response generation objective on 147M Reddit dialogs extracted by Zhang et al. (2020) and then trained with knowledge-grounded response generation objective on several datasets with knowledge annotation, including DSTC7 (Galley et al. 2019), MS MARCO (Nguyen et al. 2016), UnifiedQA (Khashabi et al. 2020), and Schema-Guided Dialog (Rastogi et al. 2020).

In our main experiment, we use GPT-2 Large (Radford et al. 2019) for keyword extraction and perform keyword-guided pre-training (KPT) on T5 solely or after response generation pre-training (RG), denoted as **T5-KPT** and **T5-RG-KPT** respectively. We use ConvLab-3 (Zhu et al. 2022) for dataset loading and model training.

We consider two sizes of model: 60M T5-small and 220M T5-base. Since GODEL does not have a T5-small version, we only use the released T5-base version for comparison. For both RG and KPT, we pre-train the models for 1 epoch. The fine-tuning settings are the same for all models and downstream tasks. We use the same knowledge serialization for all models. During pre-training, we set the keyword ratio α to 0.3, and it is empirically chosen as illustrated in §5.3. To perform K-shot fine-tuning, we randomly sample K dialogs from the original training and validation set respectively. We alter K in $\{50, 100, 200\}$ and sample the data 5 times for each K in order to estimate the performance variance. We fine-tune the models until the validation loss does not decrease for 5 consecutive epochs. Models with the lowest validation losses during training are selected as the final models. We use Adafactor optimizer with a constant learning rate 1e-3 for both pre-training and fine-tuning. We set the batch size per GPU to 64 and use 8/2 Tesla V100 32G GPUs for pre-training/fine-tuning. During inference, We adopt greedy decoding for a fair comparison of different models.

5 Experiments and Analysis

We organize our experiments as follows. We compare keyword-guided pre-training with other pre-training methods in §5.1 and analyze the effect of each pre-training objective in Sec. §5.2. We explore different keyword ratios and pre-trained LMs for keyword extraction in Sec. §5.3 and §5.4. Finally, we conduct human evaluation in §5.5. Due to the space limit, we place completed experimental results in the appendix for reference.

5.1 Main Experiment

Table 3 and 4 show the few-shot fine-tuning results of models based on T5-base and T5-small respectively. Comparing T5 and T5-RG, We can see that the response generation pre-training does not always bring benefits for grounded dialog generation. T5-RG is worse than T5 on MultiWOZ and OpenDialKG, where the response should be faithfully grounded on the given knowledge. We contend that RG pre-training teaches the model to generate the response based on the context and therefore tends to ignore the (heterogeneous) grounding knowledge. On the contrary, T5-KPT and T5-RG-KPT outperform T5 and T5-RG in most cases, especially w.r.t. the knowledge utility score, demonstrating the effectiveness of our proposed KPT on grounded dialog generation tasks. T5-KPT is comparable to T5-RG-KPT, indicating that RG pre-training before KPT is not necessary for grounded dialog generation. Nevertheless, we use T5-RG-KPT for subsequent analysis since RG pre-training may handle some scenarios where knowledge is not needed. GODEL outperforms T5 in most settings but it is still inferior to KPT. Although GODEL has been trained on several knowledge-grounded dialog datasets, the specific types of knowledge of these datasets may limit the transfer effect to different downstream knowledge grounding tasks.

5.2 Pre-training Objectives Ablation

To investigate the effect of each pre-training objective, we pre-train the grounded response generation model using ei-

Dataset	Model	BLEU-4			Unigram F1			Rouge-L			Knowledge Utility		
		50	100	200	50	100	200	50	100	200	50	100	200
MultiWOZ	T5	16.7	20.8	26.7	35.0	45.1	50.3	33.2	40.8	45.7	65.9	82.0	88.2
	GODEL	17.8	21.5	26.1	38.3	44.1	48.9	36.0	40.3	44.7	69.5	82.4	88.3
	T5-RG	17.4	20.9	25.5	36.2	41.7	47.8	35.1	39.2	44.3	64.2	75.0	81.9
	T5-KPT	23.1	**27.1**	28.7	**46.2**	**51.0**	53.4	41.8	**45.6**	47.9	84.0	**91.8**	93.7
	T5-RG-KPT	**23.4**	26.8	**29.7**	**46.2**	50.9	**54.3**	**42.0**	45.4	**48.5**	**84.2**	91.7	**95.0**
OpenDialKG	T5	7.9	7.3	7.5	25.0	23.1	23.1	23.0	23.1	23.8	54.6	57.2	52.9
	GODEL	9.4	9.9	11.0	27.8	33.1	31.8	27.4	30.7	29.7	59.9	68.4	68.4
	T5-RG	7.8	10.1	11.0	23.3	28.0	29.9	24.5	28.2	29.6	52.6	60.8	65.1
	T5-KPT	**9.7**	**12.5**	**13.2**	**30.4**	**34.8**	**36.5**	**28.9**	**31.9**	**33.8**	**63.6**	**72.9**	**74.0**
	T5-RG-KPT	8.3	12.1	12.4	27.5	34.7	35.2	27.2	31.8	32.9	57.7	71.3	72.0
PersonaChat	T5	1.5	2.2	2.5	5.9	5.2	7.4	15.2	16.4	17.3	11.9	11.8	17.5
	GODEL	2.5	2.8	3.6	7.9	7.9	8.9	17.2	17.6	18.6	16.0	16.3	18.7
	T5-RG	2.8	3.1	3.7	7.3	8.3	9.0	17.7	18.9	19.1	14.2	13.9	18.1
	T5-KPT	**2.9**	3.3	3.5	**8.8**	8.9	9.3	**18.4**	18.7	19.0	**20.1**	**18.4**	19.5
	T5-RG-KPT	**2.9**	**3.7**	**3.8**	8.6	**9.1**	**9.6**	**18.4**	**19.3**	**19.3**	19.1	17.1	**20.8**
WoW	T5	4.0	2.3	3.1	12.0	8.9	9.4	16.1	14.0	14.9	**17.1**	7.7	9.7
	GODEL	2.8	4.1	4.5	9.4	11.6	12.7	15.3	17.1	17.5	6.8	10.6	**13.6**
	T5-RG	2.5	3.8	4.5	9.4	11.3	12.3	15.5	17.0	17.8	5.1	8.1	10.0
	T5-KPT	**5.3**	**4.8**	5.1	**13.7**	**13.1**	**13.5**	**18.2**	**17.9**	18.4	14.0	**12.4**	12.6
	T5-RG-KPT	5.2	4.7	**5.2**	13.1	12.8	**13.5**	17.6	17.6	**18.5**	13.2	11.7	12.5

Table 3: Performance of models based on T5-base. Results of 50/100/200-shot are shown in different columns. We report the means of 5 runs and move the standard deviations to the appendix to save space. The best results are in bold.

ther golden keywords K_g or noisy keywords K_n after RG pre-training, and they are referred to as T5-RG-K_g and T5-RG-K_n respectively. Results of different model ablations of T5-small model size are shown in Table 4. We could see that these two kinds of keywords contribute differently. On MultiWOZ, the superior performance of T5-RG-KPT mainly comes from K_g, while K_n contributes more on PersonaChat. This observation is also intuitive as MultiWOZ needs faithfully grounding on the knowledge the most, while PersonaChat needs selectively grounding the most. On WoW, these two types of keywords have similar contributions to T5-RG-KPT on top of T5-RG. On OpenDialKG, their influences are not clear, since the relative performances differ across shots.

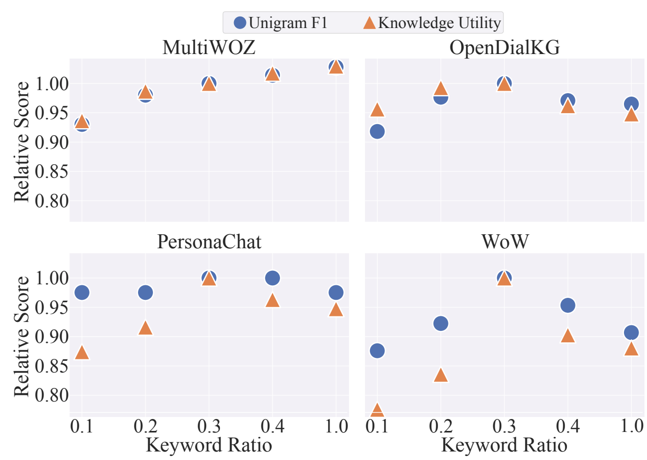

Figure 2: Performance of models based on T5-small with different keyword ratios (α). We average 50, 100, and 200 shots' performance and report the relative scores compared to the setting of α=0.3.

5.3 Keyword Ratio

In this section, we investigate the influence of keyword ratio in KPT. As we mentioned before, keywords are chosen from non-stop words (roughly 50% of all words) to serve as pseudo grounding knowledge. Therefore, we vary the keyword ratio α in $\{0.1, 0.2, 0.3, 0.4, 1.0\}$, where 1.0 means we use all non-stop words. When α is 0, KPT degenerates to RG. Larger α leaks more information, which encourages the model to ground on the knowledge more. However, too large α will make the model ignore the context. As shown in Figure 2, as α increases, the performance on all tasks except MultiWOZ first increases until α=0.3 then decreases, demonstrating the necessity of choosing appropriate α. Performance on MultiWOZ continually increases with larger α, which can be attributed to the fact that this task is heavily grounded on knowledge and rarely rely on context.

5.4 Keyword Extraction Model

We also explore different keyword extraction methods in KPT. Besides GPT-2 Large (774M), we also try DialoGPT Large (762M) as the keyword extraction model. We com-

Dataset	Model	Unigram F1	Knowledge Utility
MWOZ	T5	38.0/44.9/49.7	77.1/83.3/88.9
	T5-RG	36.5/41.7/47.4	67.2/76.6/82.5
	T5-KPT	47.1/50.7/53.6	89.5/91.8/94.6
	T5-RG-KPT	46.8/50.6/53.0	88.0/92.0/94.0
	T5-RG-K_g	**48.0/50.7/52.5**	**91.7/92.8/93.8**
	T5-RG-K_n	41.7/46.7/51.0	78.8/83.9/88.5
ODKG	T5	20.4/25.6/27.9	46.7/58.0/61.5
	T5-RG	21.3/26.1/29.6	47.1/57.8/63.5
	T5-KPT	32.4/35.1/34.9	67.6/72.2/71.2
	T5-RG-KPT	32.7/35.2/34.6	69.4/72.9/70.9
	T5-RG-K_g	32.7/**33.9/35.0**	68.4/**70.5/71.1**
	T5-RG-K_n	**33.8**/33.7/33.8	**70.1**/70.0/69.1
PC	T5	4.0/4.5/5.6	15.1/17.1/17.4
	T5-RG	7.2/7.3/7.5	13.9/14.6/16.5
	T5-KPT	7.9/7.8/8.0	22.2/19.8/18.2
	T5-RG-KPT	8.1/7.9/8.2	18.1/20.0/19.1
	T5-RG-K_g	7.7/7.4/7.5	**21.5**/14.4/17.9
	T5-RG-K_n	**7.9/7.9/8.2**	15.7/**18.1/19.0**
WoW	T5	10.1/9.0/8.6	13.7/11.0/10.5
	T5-RG	9.3/10.0/10.6	7.1/8.4/9.6
	T5-KPT	13.2/11.7/12.2	15.1/10.0/12.1
	T5-RG-KPT	13.3/12.2/13.1	15.2/11.8/13.1
	T5-RG-K_g	12.3/11.0/12.1	**14.7**/10.1/**11.4**
	T5-RG-K_n	**12.5/11.6/12.3**	12.9/**10.3/11.4**

Table 4: Performance of models based on T5-small. Results of 50/100/200-shot are separated by a slash. The best results among T5-RG-K_g and T5-RG-K_n are in bold.

Dataset	KE Model	Unigram F1	Knowledge Utility
MWOZ	GPT-2	46.8/50.6/53.0	88.0/**92.0**/94.0
	DialoGPT	**48.1**/51.0/53.0	**89.8**/91.7/94.3
	Random	47.1/**51.1/54.0**	87.3/91.8/**94.7**
ODKG	GPT-2	**32.7/35.2/34.6**	**69.4/72.9/70.9**
	DialoGPT	29.6/32.8/33.4	64.6/69.4/69.6
	Random	32.6/34.6/33.8	68.4/72.2/69.4
PC	GPT-2	**8.1/7.9/8.2**	**18.1/20.0/19.1**
	DialoGPT	7.8/7.6/8.0	14.1/18.3/17.0
	Random	7.7/7.7/8.0	13.5/19.4/18.0
WoW	GPT-2	**13.3/12.2/13.1**	**15.2**/11.8/**13.1**
	DialoGPT	11.8/12.1/12.7	12.2/**11.9**/13.0
	Random	12.1/11.6/11.9	12.2/10.6/11.0

Table 5: Performance of T5-RG-KPT models based on T5-small but with different keyword extraction models. Results of 50/100/200-shot are separated by a slash, and best results are in bold.

Win Rate	Our vs T5-RG	Our vs GODEL
MWOZ-Coherence	35.3 vs 34.7	42.7 vs 31.7
MWOZ-Knowledge	50.3 vs 22.7**	46.0 vs 24.0**
ODKG-Coherence	49.3 vs 40.3*	48.7 vs 41.3*
ODKG-Knowledge	52.3 vs 39.0	44.7 vs 40.7
PC-Coherence	43.7 vs 40.0	49.0 vs 37.0
PC-Knowledge	47.3 vs 40.3	41.3 vs 39.7**
WoW-Coherence	48.3 vs 44.0	47.3 vs 44.7
WoW-Knowledge	54.0 vs 39.3	48.0 vs 44.7

Table 6: Pairwise human evaluation. * means Fleiss' kappa $\kappa<0$ (poor agreement) and ** means $\kappa>0.2$ (fair agreement), while others are within [0, 0.2] (slight agreement).

pare the results of these two keyword extraction models and a vanilla random extraction approach in Table 5 for T5-RG-KPT of T5-small model size. Compared with using GPT-2, random extraction leads to worse performance on PersonaChat and WoW, while using DialoGPT performs poorly on OpenDialKG and PersonaChat. This result shows the influence of bias in keyword extraction on the final performance of the grounded response generation task. We assume that GPT-2 may create a more accurate estimation of Eq. (1) on diverse keyword-guided pre-training corpora than DialoGPT since it is pre-trained on more diverse text corpora. Nevertheless, we also found that the performance gaps between different keyword extraction methods are not significant, and using random extraction can already achieve satisfying performance.

5.5 Human Evaluation

We further employed workers on Amazon Mechanical Turk to evaluate the responses generated by T5-RG-KPT and two baselines T5-RG and GODEL. All models are of T5-base model size and fine-tuned on 50 dialogs. For a sample, we asked 3 workers to compare 2 responses (ours and the baseline model's) and select the better one w.r.t. context coherence and knowledge utility, resulting in 4 (tasks) * 2 (baselines) * 2 (aspects) * 100 (samples) * 3 (workers) = 4800 scores. 141 workers attended the human evaluation in total.

The result is shown in Table 6. Since dialog evaluation is rather subjective and subtle, the low agreement is acceptable. We can see that T5-RG-KPT is better than T5-RG and GODEL, consistent with the automatic results. Compared with T5-RG, our model mostly improve knowledge utility, while compared with GODEL, the improvement in context coherence is more obvious.

6 Conclusion

In this paper, we propose *keyword-guided pre-training* (KPT), a novel dialog pre-training method for grounded dialog generation. KPT does not need the dialogs to be annotated with grounding knowledge and outperforms standard response generation pre-training and supervised knowledge-grounded pre-training on several downstream tasks with diverse knowledge sources. Further analysis shows that our two pre-training objectives complement each other, enabling KPT to handle scenarios that make use of knowledge at different levels. Altogether, we contend that KPT is an effective, scalable, and versatile pre-training technique towards build better grounded dialog systems.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2021ZD0113304), the National Science Foundation for Distinguished Young Scholars (with No. 62125604), the NSFC projects (Key project with No. 61936010 and regular project with No. 61876096), and sponsored by Tsinghua-Toyota Joint Research Fund.

References

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *EMNLP*, 5016–5026. Brussels, Belgium: Association for Computational Linguistics.

Byrne, B.; Krishnamoorthi, K.; Ganesh, S.; and Kale, M. 2021. TicketTalk: Toward human-level performance with end-to-end, transaction-based dialog systems. In *ACL-IJCNLP*, 671–680. Online: Association for Computational Linguistics.

Byrne, B.; Krishnamoorthi, K.; Sankar, C.; Neelakantan, A.; Goodrich, B.; Duckworth, D.; Yavuz, S.; Dubey, A.; Kim, K.-Y.; and Cedilnik, A. 2019. Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. In *EMNLP-IJCNLP*, 4516–4525. Hong Kong, China: Association for Computational Linguistics.

Dai, Z.; Chaganty, A. T.; Zhao, V.; Amini, A.; Green, M.; Rashid, Q.; and Guu, K. 2022. Dialog Inpainting: Turning Documents to Dialogs. In *ICML*. PMLR.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *ICLR*.

Eric, M.; Goel, R.; Paul, S.; Sethi, A.; Agarwal, S.; Gao, S.; Kumar, A.; Goyal, A.; Ku, P.; and Hakkani-Tur, D. 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*, 422–428. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Eric, M.; Krishnan, L.; Charette, F.; and Manning, C. D. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *SIGdial*, 37–49. Saarbrücken, Germany: Association for Computational Linguistics.

Feng, X.; Feng, X.; Qin, L.; Qin, B.; and Liu, T. 2021. Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization. In *ACL-IJCNLP*, 1479–1491. Online: Association for Computational Linguistics.

Freitas, D. D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; and Le, Q. V. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977*.

Galley, M.; Brockett, C.; Gao, X.; Gao, J.; and Dolan, B. 2019. Grounded response generation task at dstc7. In *DSTC-7@AAAI*.

Ghazvininejad, M.; Brockett, C.; Chang, M.-W.; Dolan, B.; Gao, J.; Yih, W.-t.; and Galley, M. 2018. A knowledge-grounded neural conversation model. In *AAAI*.

Gu, X.; Yoo, K. M.; and Ha, J.-W. 2021. DialogBERT: Discourse-Aware Response Generation via Learning to Recover and Rank Utterances. In *AAAI*.

He, W.; Dai, Y.; Yang, M.; Sun, J.; Huang, F.; Si, L.; and Li, Y. 2022. Unified Dialog Model Pre-training for Task-Oriented Dialog Understanding and Generation. In *SIGIR*, 187–200.

Khashabi, D.; Min, S.; Khot, T.; Sabharwal, A.; Tafjord, O.; Clark, P.; and Hajishirzi, H. 2020. UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In *Findings of EMNLP*, 1896–1907. Online: Association for Computational Linguistics.

Lee, S.; Schulz, H.; Atkinson, A.; Gao, J.; Suleman, K.; El Asri, L.; Adada, M.; Huang, M.; Sharma, S.; Tay, W.; and Li, X. 2019. Multi-Domain Task-Completion Dialog Challenge. In *DSTC-8@AAAI*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*, 7871–7880. Online: Association for Computational Linguistics.

Li, J.; He, B.; and Mi, F. 2022. Exploring Effective Information Utilization in Multi-Turn Topic-Driven Conversations. *arXiv preprint arXiv:2209.00250*.

Li, J.; Peng, B.; Lee, S.; Gao, J.; Takanobu, R.; Zhu, Q.; Huang, M.; Schulz, H.; Atkinson, A.; and Adada, M. 2020. Results of the Multi-Domain Task-Completion Dialog Challenge. In *DSTC-8@AAAI*.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP 2017*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Mehri, S.; Razumovskaia, E.; Zhao, T.; and Eskenazi, M. 2019. Pretraining Methods for Dialog Context Representation Learning. In *ACL*, 3836–3845. Florence, Italy: Association for Computational Linguistics.

Mi, F.; Li, Y.; Zeng, Y.; Zhou, J.; Wang, Y.; Xu, C.; Shang, L.; Jiang, X.; Zhao, S.; and Liu, Q. 2022. PANGUBOT: Efficient Generative Dialogue Pre-training from Pre-trained Language Model. *arXiv preprint arXiv:2203.17090*.

Miller, A.; Feng, W.; Batra, D.; Bordes, A.; Fisch, A.; Lu, J.; Parikh, D.; and Weston, J. 2017. ParlAI: A Dialog Research

Software Platform. In *EMNLP: System Demonstrations*, 79–84. Copenhagen, Denmark: Association for Computational Linguistics.

Moon, S.; Shah, P.; Kumar, A.; and Subba, R. 2019. OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs. In *ACL*, 845–854. Florence, Italy: Association for Computational Linguistics.

Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; and Deng, L. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPs*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Peng, B.; Galley, M.; He, P.; Brockett, C.; Liden, L.; Nouri, E.; Yu, Z.; Dolan, B.; and Gao, J. 2022. GODEL: Large-Scale Pre-training for Goal-Directed Dialog. *arXiv preprint arXiv:2206.11309*.

Peng, B.; Li, C.; Li, J.; Shayandeh, S.; Liden, L.; and Gao, J. 2021. Soloist: Building Task Bots at Scale with Transfer Learning and Machine Teaching. *TACL*, 9: 807–824.

Post, M. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191. Brussels, Belgium: Association for Computational Linguistics.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Rastogi, A.; Zang, X.; Sunkara, S.; Gupta, R.; and Khaitan, P. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *AAAI*.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E. M.; Boureau, Y.-L.; and Weston, J. 2021. Recipes for Building an Open-Domain Chatbot. In *ACL*, 300–325. Online: Association for Computational Linguistics.

Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; and Weston, J. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of EMNLP*, 3784–3803. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Shuster, K.; Xu, J.; Komeili, M.; Ju, D.; Smith, E. M.; Roller, S.; Ung, M.; Chen, M.; Arora, K.; Lane, J.; et al. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Song, Y.; te Li, C.; Nie, J.-Y.; Zhang, M.; Zhao, D.; and Yan, R. 2018. An Ensemble of Retrieval-Based and Generation-Based Human-Computer Conversation Systems. In *IJCAI*.

Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Wang, Y.; Li, Y.; Wang, Y.; Mi, F.; Zhou, P.; Wang, X.; Liu, J.; Jiang, X.; and Liu, Q. 2022. Pan More Gold from the Sand: Refining Open-domain Dialogue Training with Noisy Self-Retrieval Generation. In *COLING*, 636–647. International Committee on Computational Linguistics.

Wen, T.-H.; Gašić, M.; Mrkšić, N.; Su, P.-H.; Vandyke, D.; and Young, S. 2015. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In *EMNLP*, 1711–1721. Lisbon, Portugal: Association for Computational Linguistics.

Wu, C.-S.; Hoi, S. C.; Socher, R.; and Xiong, C. 2020. TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue. In *EMNLP*, 917–929. Online: Association for Computational Linguistics.

Xie, T.; Wu, C. H.; Shi, P.; Zhong, R.; Scholak, T.; Yasunaga, M.; Wu, C.-S.; Zhong, M.; Yin, P.; Wang, S. I.; Zhong, V.; Wang, B.; Li, C.; Boyle, C.; Ni, A.; Yao, Z.; Radev, D.; Xiong, C.; Kong, L.; Zhang, R.; Smith, N. A.; Zettlemoyer, L.; and Yu, T. 2022. UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models. *arXiv preprint arXiv:2201.05966*.

Xu, Y.; and Zhao, H. 2021. Dialogue-oriented Pre-training. In *Findings of ACL-IJCNLP*, 2663–2673. Online: Association for Computational Linguistics.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *ACL*, 2204–2213. Melbourne, Australia: Association for Computational Linguistics.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *ACL: System Demonstrations*, 270–278. Online: Association for Computational Linguistics.

Zhang, Y.; Sun, S.; Gao, X.; Fang, Y.; Brockett, C.; Galley, M.; Gao, J.; and Dolan, B. 2022. RetGen: A Joint framework for Retrieval and Grounded Text Generation Modeling. In *AAAI*.

Zhao, X.; Wu, W.; Tao, C.; Xu, C.; Zhao, D.; and Yan, R. 2019. Low-Resource Knowledge-Grounded Dialogue Generation. In *ICLR*.

Zhong, P.; Liu, Y.; Wang, H.; and Miao, C. 2021. Keyword-guided neural conversational model. In *AAAI*.

Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*.

Zhu, Q.; Geishauser, C.; chin Lin, H.; van Niekerk, C.; Peng, B.; Zhang, Z.; Heck, M.; Lubis, N.; Wan, D.; Zhu, X.; Gao, J.; Gašić, M.; and Huang, M. 2022. ConvLab-3: A Flexible Dialogue System Toolkit Based on a Unified Data Format. *arXiv preprint arXiv:2211.17148*.