A Synthetic Data Generation Framework for Grounded Dialogues

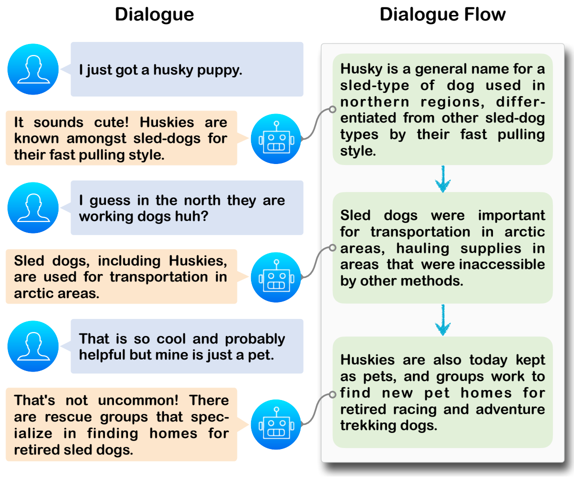
Jianzhu Bao[1,5]*, **Rui Wang**[1,6], **Yasheng Wang**[3], **Aixin Sun**[2],
Yitong Li[3,4], **Fei Mi**[3], **Ruifeng Xu**[1,5,6]†
[1]Harbin Insitute of Technology, Shenzhen, China
[2]Nanyang Technological University, Singapore
[3]Huawei Noah's Ark Lab, [4]Huawei Technologies Co., Ltd.
[5]Peng Cheng Laboratory, Shenzhen, China
[6]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
jianzhubao@gmail.com, ruiwangnlp@outlook.com, axsun@ntu.edu.sg,
{wangyasheng, mifei2, liyitong3}@huawei.com, xuruifeng@hit.edu.cn

Abstract

Training grounded response generation models often requires a large collection of grounded dialogues. However, it is costly to build such dialogues. In this paper, we present a synthetic data generation framework (SynDG) for grounded dialogues. The generation process utilizes large pre-trained language models and freely available knowledge data (*e.g.,* Wikipedia pages, persona profiles, etc.). The key idea of designing SynDG is to consider dialogue flow and coherence in the generation process. Specifically, given knowledge data, we first heuristically determine a dialogue flow, which is a series of knowledge pieces. Then, we employ T5 to incrementally turn the dialogue flow into a dialogue. To ensure coherence of both the dialogue flow and the synthetic dialogue, we design a two-level filtering strategy, at the flow-level and the utterance-level respectively. Experiments on two public benchmarks show that the synthetic grounded dialogue data produced by our framework is able to significantly boost model performance in both full training data and low-resource scenarios.

1 Introduction

Grounded dialogue systems are designed to engage in conversation with humans by incorporating external knowledge to provide relevant and informative responses (Ghazvininejad et al., 2018; Dinan et al., 2019; Gopalakrishnan et al., 2019; Zhou et al., 2018b). In recent years, various advanced techniques have been developed to train grounded dialogue models (Zheng et al., 2020; Cui et al., 2021; Xu et al., 2022; Li et al., 2022a). Despite the notable progress, training these models often requires large amounts of data. However, it is expensive and time-consuming to build a collection of dialogue data that is naturally grounded on documents or knowledge (Li et al., 2020, 2022b).



I just got a husky puppy.

It sounds cute! Huskies are known amongst sled-dogs for their fast pulling style.

I guess in the north they are working dogs huh?

Sled dogs, including Huskies, are used for transportation in arctic areas.

That is so cool and probably helpful but mine is just a pet.

That's not uncommon! There are rescue groups that specialize in finding homes for retired sled dogs.

Husky is a general name for a sled-type of dog used in northern regions, differentiated from other sled-dog types by their fast pulling style.

Sled dogs were important for transportation in arctic areas, hauling supplies in areas that were inaccessible by other methods.

Huskies are also today kept as pets, and groups work to find new pet homes for retired racing and adventure trekking dogs.

Figure 1: An illustrated example from the Wizard of Wikipedia dataset (Dinan et al., 2019). This example shows the dialogue flow in knowledge-grounded dialogues, *i.e.,* a sequence of knowledge pieces. As each agent response is grounded to a specific piece of knowledge, the dialogue flow implies the outline of the conversation.

One solution is to generate grounded dialogue data from unstructured knowledge, by using large pre-trained language models (LMs). Previous work on this topic has explored synthetic dialogue data generation with reinforcement learning (Lin et al., 2022) or user simulation (Wu et al., 2022). However, a key missing component in all these methods is the modeling of *dialogue flow*.

Dialogue flow can be viewed as the outline of a dialogue. The flow reflects the dialogue's content and trajectory, *i.e.,* the topics discussed in each session and the topic shifts between sessions. We consider the dialogue flow of a grounded dialogue as the sequence of the grounded knowledge pieces. Figure 1 shows an example dialogue along with

[*] This work was partially done during the internship at Huawei Noah's Ark Lab.

[†] Corresponding Author

*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers*, pages 10866–10882
July 9-14, 2023 ©2023 Association for Computational Linguistics

its associated dialogue flow. In this example, the grounded knowledge is primarily from a Wikipedia page about "*husky*" dogs. This dialogue follows a smooth knowledge flow, transitioning from "*husky*" to "*sled dogs*" and then to "*huskies as pets*". However, if we replace the second knowledge piece with "*'Esquimaux' or 'Eskimo' was a common term for pre-Columbian Arctic inhabitants of North America.*", which is also from the same Wikipedia page, then the flow becomes less consistent. As the backbone guiding the dialogue generation process, a carefully planned dialogue flow is crucial for the coherence and smoothness of the resulting dialogue.

To this end, we propose a novel framework named SynDG, to synthetically generate coherent grounded dialogues. The generated dialogues are meant to be used as auxiliary training data. In SynDG, we first determine the dialogue flow by task-specific heuristics, from the unstructured knowledge data (*e.g.,* Wikipedia pages, persona profiles, etc.). Then, we employ T5 (Raffel et al., 2020), a large pre-trained LM, to transform the dialogue flow into a synthetic dialogue, with sequential utterance generation, one at a time. To ensure the quality of the synthetic dialogue, we propose a two-level filtering strategy based on T5: flow-level filtering and utterance-level filtering. The flow-level filtering is designed to select dialogue flows with higher consistency, whereas the utterance-level filtering aims to eliminate the synthetic dialogues with poor coherence.

We conduct experiments on two grounded dialogue benchmarks, in both full training data and low-resource scenarios. We use the synthetic grounded dialogue data produced by our framework as additional training data for commonly used grounded dialogue models. Both the automatic and human evaluation results show that our synthetic data leads to significant improvement on model performance. Further analysis also reveals that model performance increases along the increase in the number of synthetic dialogues.

2 Related Work

2.1 Grounded Dialogue

Recent years have witnessed a growing interest in developing dialogue systems that can carry out knowledge-grounded (Zhang et al., 2018; Dinan et al., 2019; Zheng et al., 2020; Tao et al., 2021; Jang et al., 2022) or persona-grounded (Zhang et al., 2018; Cao et al., 2022) conversation.

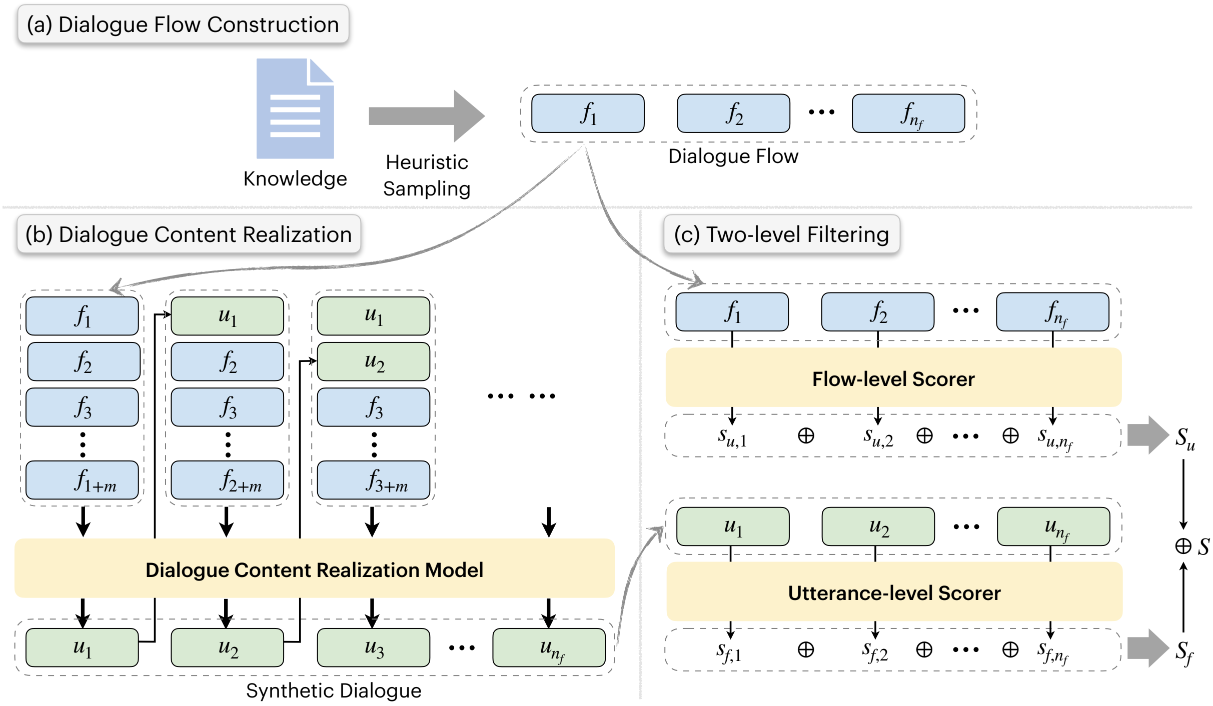
One line of research focuses on knowledge selection (Lian et al., 2019; Kim et al., 2020; Chen et al., 2020; Meng et al., 2020; Li et al., 2022a) or knowledge retrieval (Hedayatnia et al., 2020; Shuster et al., 2021; Li et al., 2022c). The models aim to identify the appropriate knowledge for each dialogue turn. Some other work aims to generate meaningful and informative responses by incorporating the grounded knowledge (Zhou et al., 2018a; Ghazvininejad et al., 2018; Li et al., 2019; Sun et al., 2022). Some recent studies have also explored retrieval-free approaches for end-to-end knowledge-grounded dialogues (Cui et al., 2021; Xu et al., 2022), with the goal of learning knowledge through the parameters of pre-trained LMs.

In particular, obtaining high-quality dialogue data that is naturally grounded on certain knowledge is known to be difficult (Zhao et al., 2020; Li et al., 2020). Zhao et al. (2020) and Liu et al. (2021) explore to train knowledge-grounded dialogue models in a low-resource scenario, where only limited knowledge-grounded dialogues are available. Li et al. (2020) and Tao et al. (2021) investigate the task of knowledge-grounded dialogue generation, in a zero-resource scenario, by using only independent knowledge resources and dialogues without knowledge grounding as training data. However, all aforementioned low-resource approaches rely on large-scale dialogue data and knowledge data for training. In this work, we explore an alternative solution to deal with the low-resource challenge, *i.e.,* synthetic grounded dialogue data generation from unstructured knowledge. The synthetic data can then serve as extra training data for grounded dialogue models.

2.2 Synthetic Dialogue Data Generation

With the superior development of pre-trained models (Devlin et al., 2019; Brown et al., 2020; Bommasani et al., 2021), many researchers have started to exploit the generation of synthetic dialogue data for training better dialogue models (Zheng et al., 2022; Dai et al., 2022; Mehri et al., 2022).

Mohapatra et al. (2021) and Wu et al. (2022) both employ two pre-trained models, as a user bot and an agent bot respectively, to simulate the interaction between two human annotators. Zheng et al. (2022) use template prompts to guide GPT-J, a large-scale pre-trained LM with 6B parameters, to generate emotional support conversations.



(a) Dialogue Flow Construction
Knowledge
Heuristic Sampling
f_1 f_2 ⋯ f_{n_f}
Dialogue Flow
(b) Dialogue Content Realization
f_1 f_2 f_3 f_{1+m}
u_1 f_2 f_3 f_{2+m}
u_1 u_2 f_3 f_{3+m}
⋯ ⋯
Dialogue Content Realization Model
u_1 u_2 u_3 ⋯ u_{n_f}
Synthetic Dialogue
(c) Two-level Filtering
Flow-level Scorer
$s_{u,1}$ ⊕ $s_{u,2}$ ⊕ ⋯ ⊕ s_{u,n_f}
S_u
⊕ S
Utterance-level Scorer
$s_{f,1}$ ⊕ $s_{f,2}$ ⊕ ⋯ ⊕ s_{f,n_f}
S_f

Figure 2: The overview of SynDG. We omit the superscript s for u_i and f_i for simplicity. When generating each synthetic dialogue, SynDG (a) constructs a dialogue flow by heuristic sampling from unstructured knowledge data; (b) incrementally realizes every knowledge piece in the flow as an utterance by a fine-tuned T5 model, producing a synthetic dialogue; (c) scores the synthetic dialogue both at the flow-level and utterance-level with two T5-based scorer. Finally, SynDG filters out the synthetic dialogues of low quality based on their scores to obtain the final set of synthetic dialogues.

Dai et al. (2022) propose to transform a document into a dialogue between the writer and the reader, in which they sequentially treat each sentence in the document as the writer's utterance and generate the reader's questions by pre-trained models. Closely related to our work, Lin et al. (2022) explore using the reinforcement learning-based generative conversational networks (Papangelis et al., 2021) to generate synthetic conversational data for knowledge-grounded dialogue models.

Yet, all the reviewed solutions above are dedicated to directly generating synthetic dialogues. The significance of explicitly simulating or modeling dialogue flows is overlooked, which is closely related to the quality of the synthetic dialogue. In this work, we explicitly construct a dialogue flow before turning it into a synthetic dialogue.

3 Task Formulation

Given a set of training grounded dialogues $\mathcal{D}^t = \{C_i^t, K_i^t, r_i^t\}_{i=1}^{N_t}$, where C_i^t is the dialogue context that is a concatenation of previous utterances, $K_i^t = [k_1^t, k_2^t, \dots, k_{n_k^t}]$ is the knowledge corpus containing several knowledge pieces, r_i^t is the knowledgeable response, the grounded dialogue generation task aims to learn a generation model $P(r^t|C^t, K^t)$ from $\mathcal{D}^t$. In the following, we omit the subscript i for simplicity. Note that only certain knowledge pieces in K^t are associated to the response r^t, while others are redundant. Besides, we use $U^t = [u_1^t, u_2^t, \dots]$ to denote all the utterances in a dialogue.

In this paper, we aim to automatically construct a set of synthetic grounded dialogues $\mathcal{D}^s = \{C_i^s, K_i^s, r_i^s\}_{i=1}^{N_s}$ without any human annotation. Then, the generation model P could be better learned form $\{\mathcal{D}^t \cup \mathcal{D}^s\}$.

4 Methodology

Our framework, named SynDG, is illustrated in Figure 2. SynDG first explicitly constructs a dialogue flow by task-specific heuristics, then realizes it into a synthetic dialogue with pre-trained models. Further, a two-level filtering strategy is proposed to ensure the quality of synthetic dialogues.

4.1 Dialogue Flow Construction

Formally, a dialogue flow is defined as a sequence of knowledge pieces $F = [f_1, f_2, \dots, f_{n_f}]$, where n_f is the length of F, f_i could be a single knowledge piece in K^t, a concatenation of several knowledge pieces, or a special token "[none]" indicating no knowledge. We let each f_i correspond to an utterance, so that n_f also denotes the number of utterances throughout a dialogue. Here, similar to most two-party dialogue benchmarks (Zhang et al., 2018; Dinan et al., 2019; Rashkin et al., 2019), we assume that the two participants take turns speaking, that is, $[f_1, f_3, f_5, \dots]$ are from one speaker, while $[f_2, f_4, f_6, \dots]$ are from another speaker.

From each training dialogue in $\mathcal{D}^t$, we can easily obtain a training dialogue flow $F^t = [f_1^t, f_2^t, \dots, f_{n_f}^t]$, because the knowledge pieces corresponding to each utterance are available.

Following the dialogue flow patterns of the training set, we construct our synthetic dialogue flows, $F^s = [f_1^s, f_2^s, \dots, f_{n_f}^s]$, by task-specific heuristics. In this work, we apply our framework to a persona-grounded dialogue benchmark, PersonaChat (Zhang et al., 2018), and an open-domain knowledge-grounded dialogue benchmark, Wizard of Wikipedia (WoW) (Dinan et al., 2019). For PersonaChat, we first randomly sample the persona sentences of the user and agent from the training set as the knowledge corpus K^s. Each persona sentence is viewed as a knowledge piece. Then, according to heuristic constraints, we sample zero, one, or more persona sentences from the knowledge corpus for each turn, thus forming a dialogue flow. For Wizard of Wikipedia, similar to Li et al. (2022a), we use the chosen topic passage and the retrieved passages in the first turn as the knowledge corpus K^s. Then, for each turn of a synthetic dialogue, at most one knowledge piece is sampled from K^s with heuristic constraints.

The heuristic constraints are defined based on our observation and summary of the dialogue flow patterns from the training set of PersonaChat/WoW. Although this heuristic sampling-based dialogue flow construction method is not universally applicable, it can be migrated to other grounded dialogue datasets with minor modifications. We describe in detail the aforementioned heuristic constraints and the specific process of constructing the dialogue flow in Appendix A. We also provide some suggestions for designing heuristic strategies on other datasets in Appendix A.3.

4.2 Dialogue Content Realization

After obtaining the synthetic dialogue flows, we train a dialogue content realization model to realize every piece of knowledge in a flow as an utterance, step by step. In this way, a synthetic dialogue flow is progressively transformed into a synthetic dialogue.

We fine-tune a pre-trained sequence-to-sequence model, T5, by a dialogue reconstruction task as our dialogue content realization model. The fine-tuning data is constructed from $\mathcal{D}^t$. During fine-tuning, each utterance within a dialogue U^t is considered as the target sequence, while its previous dialogue history and subsequent flows are combined as the source sequence. To be specific, for the i-th utterance u_i^t, the target sequence is itself, and the source sequence is:

$$(u_1^t, u_2^t, \dots, u_{i-1}^t, [\mathrm{t}], f_i^t, [/\mathrm{t}], f_{i+1}^t, \dots, f_{i+m}^t) \tag{1}$$

where u_j^t denotes the utterance in the dialogue history, $[\mathrm{t}]$ and $[/\mathrm{t}]$ indicate the target utterance to be generated should be mainly grounded to f_i^t, m is the number of subsequent knowledge pieces retained from its dialogue flow F^t. By appending the subsequent knowledge pieces, the model can take into account the future of the dialogue (*i.e.,* what will be talked about next). Thus, it can generate the i-th utterance more appropriately, making the final synthetic dialogue more coherent.

In practice, we prepend each u_j^t and f_i^t with a special token "[user]" or "[agent]" to distinguish two speakers. The standard negative log-likelihood loss is used to optimize this model.

After fine-tuning T5, we leverage it to incrementally turn the previously constructed dialogue flow $F^s = [f_1^s, f_2^s, \dots, f_{n_f}^s]$ into a synthetic dialogue $U^s = [u_1^s, u_2^s, \dots, u_{n_f}^s]$. Further, by treating each agent utterance $u_j^{s,a} \in [u_2^s, u_4^s, \dots]$ as a response r^s, we can obtain a set of synthetic grounded dialogues $\mathcal{D}^s = \{C_i^s, K_i^s, r_i^s\}_{i=1}^{N_s}$, where K_i^s is the knowledge corpus used during dialogue flow construction.

4.3 Two-level Filtering

To further improve the quality of the synthetic grounded dialogues, we design a two-level filtering strategy. It scores synthetic dialogues both at the flow-level and at the utterance-level to drop the low-quality dialogues. More precisely, inspired by Ke et al. (2022), we train two models based on T5 by a text infilling task to score the dialogue flow F^s and the synthetic dialogue U^s, respectively.

The training data is also constructed from $\mathcal{D}^t$. At the utterance-level, we first mask each utterance in a dialogue U^t in turn, then fine-tune a T5 model as our utterance-level scorer P_u to predict the masked utterance. Formally, for the i-th utterance u_i^t in U^t, we mask it to obtain the source sequence $U_{m(i)}^t$:

$$(u_1^t, \ldots, u_{i-1}^t, [\text{mask}], u_{i+1^t}, \ldots, u_{n_f}^t) \tag{2}$$

Accordingly, the target sequence to predict is u_i^t. This model is also optimized by the negative log-likelihood loss.

During inference, the utterance-level score of u_i^t can be computed via the log probability:

$$\begin{aligned} s_u(u_i^t) &= \log P_u\left(u_i^t | U_{m(i)}^t\right) \\ &= \sum_{j=1}^{|u_i^t|} P\left(u_{i,j}^t | U_{m(i)}^t, u_{i,<j}^t\right) \end{aligned} \tag{3}$$

In this way, for each synthetic dialogue U^s, we can obtain n_f scores by masking each utterance. We take the average of these scores as the overall utterance-score S_u.

Similarly, we fine-tune another T5 model as our flow-level scorer P_f by replacing U^t with F^t as the training data. We apply this model on our constructed dialogue flow F^s to get the overall flow-level score S_f. Lastly, we sum up S_u and S_f as the final quality score S of a synthetic dialogue.

5 Experimental Setups

5.1 Datasets

We conduct experiments on two publicly available and widely used grounded dialogue benchmarks: Wizard of Wikipedia (WoW) (Dinan et al., 2019) and PersonaChat (Zhang et al., 2018).

The Wizard of Wikipedia benchmark (Dinan et al., 2019) is a collection of multi-turn knowledge-grounded dialogues between two speakers. One speaker (the "wizard") has access to a collection of knowledge and the other (the "apprentice") tries to learn about a specific topic. WoW is collected by crowd-sourcing and is divided into a training set, a validation set, and a test set. The validation/test set is further divided into two subsets: Validation/Test Seen and Validation/Test Unseen. Test Unseen contains dialogues about topics that are not present in the training or validation set, while Test Seen does not guarantee this.

The PersonaChat benchmark (Zhang et al., 2018) consists of dialogues between pairs of crowdworkers. Each crowdworker is assigned certain sentences defining his/her personality, and is asked to engage in a conversation with others according to the assigned personality.

5.2 Baselines

For WoW, we use baselines based on BlenderBot (Roller et al., 2021), since it is commonly used in recent work (Lin et al., 2022; Cui et al., 2021). For PersonaChat, we adopt the GPT2-based (Radford et al., 2019) baselines in order to compare with the recent work, Cao et al. (2022).

5.2.1 WoW Dataset

For WoW, we conduct experiments under two settings, *i.e.,* grounded knowledge available (KA) and grounded knowledge unavailable (KU) settings. The former generates responses given the ground-truth knowledge, while the latter requires knowledge selection first.

BB For both settings, we choose BlenderBot (BB) as our response generation model, and concatenate the dialogue context with the ground-truth/predicted knowledge as input. For the knowledge selection model under the KU setting, we fine-tune RoBERTa (Liu et al., 2019) for binary classification to rank and predict the grounded knowledge piece. The input is the concatenation of the dialogue context and each candidate knowledge piece.

BB-SynDG We use the synthetic dialogue produced by our SynDG framework as extra training data for BB training.

5.2.2 PersonaChat Dataset

GPT2 We fine-tune GPT2 by concatenating the personas and the dialogue history as the input sequence.

GPT2-BT Cao et al. (2022) augment the training dialogue data by back translation, and then fine-tune GPT2 with both the augmented and the original data.

GPT2-D^3 D^3 (Cao et al., 2022) is a data augmentation method designed for PersonaChat, which incorporates multiple techniques and models, such as BERT, GPT2, back translation, etc.

GPT2-SynDG We replace the augmented dialogues in GPT2-D^3 with our synthetic dialogues.

Based on the models boosted by our SynDG framework (BB-SynDG and GPT2-SynDG), we further conduct ablation studies by removing flow-level filtering (***w/o FF***), utterance-level filtering (***w/o UF***), or both (***w/o FF&UF***). Besides, we also report the results using random sampling (**BB-RS** and **GPT2-RS**) instead of heuristic sampling when determining the dialogue flow for comparison.

Further, we demonstrate SynDG's capability in low-resource scenarios by using only 1/16 and 1/32 of the original training data.[1]

5.3 Evaluation Metrics

Automatic Evaluation. For the automatic evaluation, we adopt the widely used BLEU-4 (B-4) (Papineni et al., 2002), ROUGE-L (R-L) (Lin, 2004), and perplexity (PPL). Besides, for WoW, we follow Li et al. (2022c) to use F1 score to measure the unigram overlap between the generated response and the ground-truth response (F1), and the unigram overlap between the generated response and the ground-truth knowledge (KF1). Also, the knowledge selection performance under the KU setting is measured by the accuracy (ACC).

Human Evaluation. For a more comprehensive analysis, we conduct a human evaluation containing two aspects. (1) **Human likeness**: It measures the fluency, coherence, and engagement of the response, *i.e.,* whether it resembles a human response. (2) **Informativeness**: For WoW, it indicates whether a response contains appropriate, correct, and factual knowledge information. For PersonaChat, it measures whether a response is consistent with at least one persona sentence. We respectively sample 100 responses from the test set of WoW (Seen/Unseen) and PersonaChat. We adopt pair-wise comparison to conduct human evaluation, where we compare models before and after using our synthetic dialogue data. For each pair of responses generated from two models, 3 annotators are assigned to give their preferences (win, lose, or tie) in terms of the two aspects.

5.4 Implementation Details

Settings for Generating Synthetic Dialogue. The dialogue content realization model, the flow-level scorer, and the utterance-level scorer are all fine-tuned from T5-Large.[2] The AdamW optimizer (Kingma and Ba, 2015) is employed for parameter optimization with a learning rate of 1e-4. We train our model 3 epochs with a batch size of 8 and select the best checkpoint according to the loss on the validation set. The number of subsequent knowledge pieces m described in Equation 1 is set to 2 for WoW and 1 for PersonaChat.[3] Our models are implemented in PyTorch (Paszke et al., 2019) and trained on a NVIDIA Tesla V100 GPU.[4] For decoding at inference, we use a top-k sampling scheme with k = 70 and a temperature of 0.7.

The training data for the utterance-level scorer can be directly obtained from the training set of WoW/PersonaChat. Nevertheless, training the dialogue content realization model and the flow-level scorer needs the ground-truth dialogue flow data. Here, we can directly derive the dialogue flow from the training set of WoW since each utterance corresponds to one or zero knowledge piece in WoW. Unfortunately, the explicit correspondence between persona sentences and utterances is not given in PersonaChat, so we use the same method as in Cao et al. (2022) to predict the correspondence by a RoBERTa-based model first.

Settings for Baselines. On WoW, the response generation model and the knowledge selection model are respectively fine-tuned from BlenderBot-small and RoBERTa-base. The hyper-parameters for training the response generation model are consistent with Cui et al. (2021). The knowledge selection model is trained by AdamW (lr = 2e-5) with 3 epochs and a batch size of 128, and the negative sampling strategy is used with 4 negative samples during training. We generate 36,860 synthetic dialogues and select 18,430 of them as extra training data, equal to the number of dialogues in the original WoW training set. On PersonaChat, we use the code released by Cao et al. (2022) to implement all baselines with the same hyper-parameters.[5] The number of synthetic dialogues we include is 6,600 (selected from 10k synthetic dialogues by the two-level filtering), yielding 52,800 training samples, which is less than the number of augmented samples in Cao et al. (2022).

6 Results and Analysis

6.1 Automatic Evaluation

WoW Table 1 shows the automatic evaluation results on WoW Test Seen and Test Unseen sets. In

[1] We also show the results using 1/4 and 1/8 training data in Appendix D.

[2] Using T5-Base also brings noticeable improvements, as shown in Appendix E.

[3] We discuss the impact of m in Appendix C.

[4] Code is available at `https://github.com/HITSZ-HLT/SynDG`.

[5] `https://github.com/caoyu-noob/D3`

Set.	Models	WoW Seen						WoW Unseen					
		B-4	R-L	F-1	KF-1	PPL(⇓)	ACC	B-4	R-L	F-1	KF-1	PPL(⇓)	ACC
		Full Data											
KA	BB	14.60	30.98	34.21	50.27	**9.15**	-	13.94	30.70	33.46	49.33	**9.97**	-
	BB-RS	14.77	31.28	34.20	50.49	9.47	-	14.12	31.01	33.82	50.12	10.96	-
	w/ FF&UF	15.24	31.59	34.75	50.48	9.33	-	14.43	31.17	33.92	50.24	10.11	-
	BB-SynDG	**15.84**	**32.18**	**35.30**	**53.37**	9.36	-	14.67	**31.67**	34.36	**53.03**	10.22	-
	w/o FF	15.60	32.13	35.27	52.02	9.36	-	**14.82**	31.61	**34.58**	51.98	10.20	-
	w/o UF	15.48	32.08	35.13	52.16	9.35	-	14.55	31.31	34.12	51.32	10.17	-
	w/o FF&UF	15.17	31.66	34.61	50.39	9.36	-	14.30	31.11	34.01	50.06	10.18	-
KU	BB	5.52	18.32	20.11	19.85	**19.88**	22.32	5.00	18.29	19.45	18.98	**24.67**	22.01
	BB-RS	5.30	18.18	19.72	19.79	21.14	21.63	5.30	18.12	19.45	19.23	25.60	21.99
	w/ FF&UF	5.69	18.65	20.34	19.93	20.25	22.48	5.38	18.37	19.53	19.54	25.32	22.35
	BB-SynDG	**5.89**	18.56	20.20	**20.55**	21.28	**23.64**	**5.42**	18.47	**19.68**	**20.80**	25.67	**23.44**
	w/o FF	5.83	**18.83**	**20.43**	20.22	21.77	23.34	5.36	18.33	19.57	20.22	25.56	23.24
	w/o UF	5.65	18.55	20.17	20.26	21.09	22.79	5.40	**18.50**	19.65	19.88	25.64	22.55
	w/o FF&UF	5.71	18.48	20.11	19.86	21.16	22.62	5.33	18.48	19.41	19.68	25.77	22.42
		Low Resource											
KA	BB 1/16	11.84	27.77	30.27	43.47	10.80	-	11.27	27.41	29.65	42.03	11.52	-
	BB-SynDG 1/16	14.05	30.67	33.12	50.93	11.28	-	13.42	30.32	32.53	50.45	12.11	-
	BB 1/32	10.88	26.65	29.08	41.46	11.10	-	10.43	26.13	28.41	39.95	11.79	-
	BB-SynDG 1/32	12.94	28.97	31.72	49.29	11.56	-	12.54	28.94	31.28	49.06	12.38	-
KU	BB 1/16	3.30	16.12	17.22	14.85	23.84	11.82	3.66	16.42	17.53	15.66	28.90	13.91
	BB-SynDG 1/16	4.54	17.23	18.11	17.66	26.86	17.23	4.92	17.53	18.16	18.68	32.26	18.07
	BB 1/32	3.16	16.11	17.07	13.54	24.92	10.40	3.10	15.91	16.90	13.54	30.80	11.77
	BB-SynDG 1/32	4.42	16.61	17.63	16.79	27.70	16.56	4.21	16.84	17.46	17.58	33.12	17.33

Table 1: Automatic evaluation results on WoW [%]. The best scores are in **bold**.

Models	B-4	R-L	PPL(⇓)
	Full Data		
GPT2	3.70	19.71	17.66
GPT2-BT*	3.94	-	16.96
GPT2-D^3*	4.18	-	15.69
GPT2-RS	3.95	19.73	16.66
w/o FF&UF	4.22	20.18	14.77
GPT2-SynDG	**4.26**	**20.40**	**14.52**
w/o FF	4.13	20.20	14.52
w/o UF	4.21	20.33	14.54
w/o FF&UF	4.01	19.96	14.88
	Low Resource		
GPT2 1/16	1.65	13.61	35.52
GPT2-SynDG 1/16	2.80	16.92	21.01
GPT2 1/32	1.38	12.24	58.50
GPT2-SynDG 1/32	2.53	16.74	23.42

Table 2: Automatic evaluation results on PersonaChat [%]. $*$ denotes the results derived from Cao et al. (2022)

Set.	A vs. B	HL			Informativeness		
		Win	Lose	κ	Win	Lose	κ
		WoW Seen					
KA	BB-SynDG / BB	34.67	26.33	.43	37.00	30.67	.42
KU	BB-SynDG / BB	28.00	23.67	.45	25.00	20.67	.48
		WoW Unseen					
KA	BB-SynDG / BB	38.67	32.00	.46	38.00	28.33	.52
KU	BB-SynDG / BB	28.67	24.67	.41	24.00	19.67	.50
		PersonaChat					
GPT2-SynDG / GPT2		30.67	17.66	.48	29.00	11.00	.47

Table 3: Human evaluation results in terms of the winning/losing rates of SynDG [%]. κ is the Fleiss' Kappa. HL is short for human likeness.

the full training data scenario, BB-SynDG under the KA setting achieves significantly better BLEU-4, ROUGE-L, F-1, and KF-1 scores than BB on both seen and unseen topics, demonstrating the usefulness of our generated synthetic dialogues. Under the KU setting, our BB-SynDG can improve the performance on both the response generation task and the knowledge selection task on top of BB. These observations suggest that the synthetic dialogues generated by our proposed framework not only help the model to generate better responses, but also enhance its ability to ground knowledge. Also, we can observe that both the flow-level filtering and the utterance-level filtering contribute noticeable improvement on BB-SynDG under the KA/KU setting. Concretely, either removing the flow-level filtering (*w/o FF*) or the utterance-level filtering (*w/o UF*) causes some performance degradation, and removing both of them (*w/o FF&UF*) results in further decreases. By comparing (BB-SynDG *w/o FF&UF*) with BB-RS, we find that random sampling to obtain dialogue flow is less effective and even harms the performance (BB-RS on WoW Seen under the KU setting), while our proposed heuristic sampling method works better. Also, adding two-level filtering to BB-RS also achieves a considerable performance improvement, again demonstrating the usefulness of the two-level

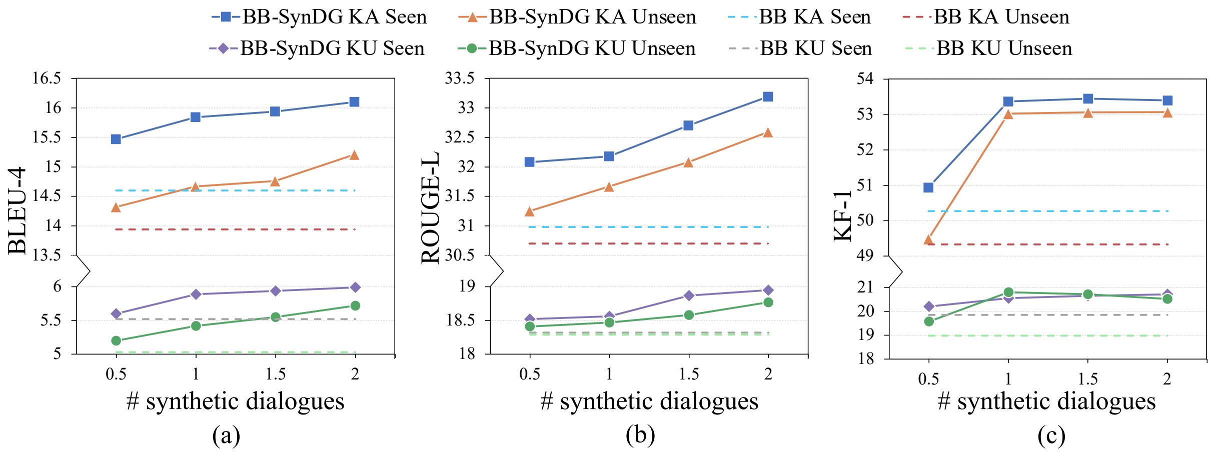


Figure 3: Impact of the number of synthetic dialogues on WoW. The vertical axis denotes the evaluation metrics. The horizontal axis indicates the number of synthetic dialogues is how many times the size of the WoW training set.

filtering strategy.

In the low resource scenario, more significant performance improvements can be observed. Surprisingly, under the KA setting, BB-SynDG with only 1/16 of the training data can already achieve comparable performance to BB with full training data, indicating that SynDG can mitigate the low resource problem in grounded dialogues.

PersonaChat Table 2 shows the automatic evaluation results on PersonaChat. In the full training data scenario, compared to GPT2, adding our synthetic dialogues (GPT2-SynDG) significantly improves the performance. Although GPT2-D^3 includes more augmented training data generated through sophisticated techniques, GPT2-SynDG still outperforms it. In general, GPT2-SynDG performs better than GPT2-BT, GPT2-CVAE, and GPT2-D^3, showing that our framework for generating synthetic dialogues is superior to existing data augmentation techniques on PersonaChat. In addition, the ablation experiments (*w/o FF*, *w/o UF*, *w/o FF&UF*, and GPT2-RS) demonstrate similar results to those on WoW, that is, our proposed heuristic sampling and two-level filtering strategy are essential for generating high-quality and useful synthetic dialogues. Also, incorporating the SynDG framework can improve the model results more significantly in the low-resource scenario.

It is worth noting that introducing synthetic dialogues on WoW makes the PPL score decrease, while it improves on PersonaChat. We hypothesize that the reason may be that the dialogues in WoW involve more complicated knowledge and more diverse utterance than dialogues in PersonaChat. As a result, on WoW, the quality of synthetic dialogues automatically generated by the large LM has a larger gap with the human-written dialogues.

6.2 Human Evaluation

Human evaluation results are shown in Table 3. The results indicate that the introduction of SynDG brings the base model (BB and GPT2) a significant improvement in generating more natural and knowledgeable responses. On WoW, the advantage of SynDG under the KA setting is more evident than under the KU setting, which follows the results of the automatic metrics.

6.3 Impact of the Number of Synthetic Dialogues

With SynDG, we can automatically generate numerous synthetic dialogues. However, how many synthetic dialogues are appropriate to integrate as extra training samples? To answer this question, we show the model performance on WoW with respect to different numbers of synthetic dialogues in Figure 3.

From Figures 3(a) and 3(b), we can observe that the BLEU-4 and ROUGE-L scores tend to increase as the number of synthetic dialogues grows, showing the potential of our proposed SynDG framework. However, through Figure 3(c), we found that the KF-1 score tends to be stable after a rapid increase. We speculate that this may be due to that the scale of the LM we used limits the upper bound of the quality of the synthesized dialogues. We can also find that increasing the amount of synthetic data may not improve performance indefinitely. The improvement becomes less obvious when the amount of synthetic data reaches twice the amount of the original data. The results of F1 and PPL scores are not shown. This is because the trend of F1 score is similar to that of BLUE-4 and ROUGE-L scores, while the variation of PPL score is not significant.

7 Conclusion

In this paper, we propose a framework, SynDG, to automatically construct synthetic training data for the grounded dialogue task. We first construct dialogue flows based on unstructured knowledge, then transform them into synthetic dialogues by large LMs, and finally filter and retain the generated dialogues with high quality. The experimental results demonstrate the effectiveness of our proposed framework in both full training data and low-resource scenarios. Further analysis shows that the model performance tends to increase as the number of synthetic dialogues increases. For future work, we plan to investigate more efficient strategies for determining dialogue flows and take larger LMs to produce synthetic dialogues with higher quality.

Limitations

As discussed in Appendix B, there is still a gap between the synthetic dialogues and the human-written dialogues in terms of quality. The synthetic dialogues sometimes do not express knowledge with sufficient accuracy. Also, some of the synthetic dialogues are less coherent and diverse than the human-written ones. We believe that these issues can be mitigated in two aspects. First, similar to (Zheng et al., 2022), employing larger LMs can help generate utterances with higher quality. Second, introducing knowledge graph and textual reasoning techniques to produce better dialogue flows.

In addition, using large LMs inevitably requires more computational resources. However, it is still a cheaper and promising alternative to hiring expensive labor.

Ethics Statement

The paper focuses on generating synthetic dialogues for training grounded dialogue systems. Our framework is developed based on the commonly used large pre-trained LM, T5 (Raffel et al., 2020). It is trained on large scale web data that is known to contain biased or discriminatory content. However, how to remove bias from large LMs is still a hard research problem so far. The datasets we use are publicly available and contain no personal identifiable information.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (62006062, 62176076), Natural Science Foundation of GuangDong 2023A1515012922, the Shenzhen Foundational Research Funding (JCYJ20220818102415032), the Major Key Project of PCL2021A06, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 2022B1212010005.

References

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Yu Cao, Wei Bi, Meng Fang, Shuming Shi, and Dacheng Tao. 2022. A model-agnostic data manipulation method for persona-based dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7984–8002, Dublin, Ireland. Association for Computational Linguistics.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3426–3437, Online. Association for Computational Linguistics.

Leyang Cui, Yu Wu, Shujie Liu, and Yue Zhang. 2021. Knowledge enhanced fine-tuning for better handling unseen entities in dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2328–2337, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y. Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 4558–4586. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

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

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421, Dublin, Ireland. Association for Computational Linguistics.

Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Dong-Hoon Shin, Seungryong Kim, and Heuiseok Lim. 2022. Call for customized conversation: Customized conversation grounding persona and knowledge. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10803–10812. AAAI Press.

Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. CTRLEval: An unsupervised reference-free metric for evaluating controlled text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2319, Dublin, Ireland. Association for Computational Linguistics.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sha Li, Mahdi Namazifar, Di Jin, Mohit Bansal, Heng Ji, Yang Liu, and Dilek Hakkani-Tur. 2022a. Enhancing knowledge selection for grounded dialogues via document semantic graphs. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2810–2823, Seattle, United States. Association for Computational Linguistics.

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022b. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Asso-*

ciation for Computational Linguistics: Human Language Technologies, pages 206–218, Seattle, United States. Association for Computational Linguistics.

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022c. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 206–218. Association for Computational Linguistics.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 12–21, Florence, Italy. Association for Computational Linguistics.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5081–5087. ijcai.org.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yen Ting Lin, Alexandros Papangelis, Seokhwan Kim, and Dilek Hakkani-Tur. 2022. Knowledge-grounded conversational data augmentation with generative conversational networks. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 26–38, Edinburgh, UK. Association for Computational Linguistics.

Shilei Liu, Xiaofeng Zhao, Bochao Li, Feiliang Ren, Longhui Zhang, and Shujuan Yin. 2021. A three-stage learning framework for low-resource knowledge-grounded dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2262–2272. Association for Computational Linguistics.

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

Shikib Mehri, Yasemin Altun, and Maxine Eskénazi. 2022. LAD: language models as data for zero-shot dialog. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2022, Edinburgh, UK, 07-09 September 2022*, pages 595–604. Association for Computational Linguistics.

Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1151–1160. ACM.

Biswesh Mohapatra, Gaurav Pandey, Danish Contractor, and Sachindra Joshi. 2021. Simulated chats for building dialog systems: Learning to generate conversations from instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1190–1203, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexandros Papangelis, Karthik Gopalakrishnan, Aishwarya Padmakumar, Seokhwan Kim, Gokhan Tur, and Dilek Hakkani-Tur. 2021. Generative conversational networks. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 111–120, Singapore and Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

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

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 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

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

Qingfeng Sun, Can Xu, Huang Hu, Yujing Wang, Jian Miao, Xiubo Geng, Yining Chen, Fei Xu, and Daxin Jiang. 2022. Stylized knowledge-grounded dialogue generation via disentangled template rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3304–3318. Association for Computational Linguistics.

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

Qingyang Wu, Song Feng, Derek Chen, Sachindra Joshi, Luis Lastras, and Zhou Yu. 2022. DG2: Data augmentation through document grounded dialogue generation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 204–216, Edinburgh, UK. Association for Computational Linguistics.

Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2022. Retrieval-free knowledge-grounded dialogue response generation with adapters. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 93–107, Dublin, Ireland. Association for Computational Linguistics.

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

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020. Low-resource knowledge-grounded dialogue generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 115–125, Online. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, and Minlie Huang. 2022. Augesc: Large-scale data augmentation for emotional support conversation with pre-trained language models. *CoRR*, abs/2202.13047.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4623–4629. ijcai.org.

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

A Details of Dialogue Flow Construction

We sample the dialogue flows according to several heuristic constraints, which are based on our observation and summary of the patterns of the dialogue flows from the training set of PersonaChat/WoW. The patterns we identified in PersonaChat/WoW are obvious and do not require much manual effort. So we think our task-specific heuristics are cheaper compared to manual data annotation. Also, this heuristic sampling-based dialogue flow construction method can be migrated to other grounded dialogue datasets with minor task-specific modifications. As a simple example, given an article, when determining the knowledge pieces corresponding to the first utterance, we can sample the first sentence of the article with a high probability. Afterwards, the remaining sentences of the article can be sampled evenly to obtain the subsequent knowledge pieces. The reason is that the beginning of an article usually expresses the main idea of the whole article and is more suitable for starting a conversation.

A.1 PersonaChat

First, all the persona sentences in the training set are collected as a candidate persona pool. Second, we randomly sample 10 persona sentences from the candidate persona pool, then divide them equally into two groups as the user and agent's persona profiles, *i.e.,* K_u and K_a. We can also regard K_u and K_a as the knowledge corpus K^s described in Section 3. Finally, based on K_u and K_a, we sample $n_f = 16$ persona sentences to form the dialogue flow $F^s = [f^s_1, f^s_2, ...f^s_{n_f}]$ according to the following constraints:

- $[f_1, f_3, f_5, ...]$ are sampled from K_u, while $[f_2, f_4, f_6, ...]$ are sampled from K_a.
- When sampling f_i, there is a 0.5 probability that it is uniformly sampled from K_u/K_a. Otherwise, we set f_i to "[none]", indicating that the utterance in this turn does not need to be grounded to certain persona sentences.
- If f_i is not "[none]", there is a 0.1 probability that it contains two uniformly sampled persona sentences. Otherwise, it contains only one.
- Each persona sentence in K_u and K_a can only be sampled at most twice.

A.2 WoW

We sample dialogue flows based on the training instances in WoW. Each training instance in WoW can sample out multiple different dialogue flows.

Specifically, for each training instance, we use its chosen topic passage K_t and the retrieved passages in its first turn K_r as the knowledge corpus K^s. Usually, a dialogue in WoW mainly focuses on its chosen topic, and occasionally mentions other related topics of the retrieved passages. Thus, the knowledge pieces in its dialogue flow are mainly from K_t, and only a few of them are from K_r.

Based on the above observation, we sample $n_f = 10$ knowledge pieces from K_t and K_r to obtain the dialogue flow $F^s = [f^s_1, f^s_2, ...f^s_{n_f}]$ according to the following constraints:

- $[f_1, f_3, f_5, ...]$ are all "[none]", while only $[f_2, f_4, f_6, ...]$ are sampled from K_t/K_r. This is because only the wizard's utterances are grounded to knowledge.
- If f_i is not "[none]", there is a 0.9 probability that it should be sampled form K_t. Otherwise, it will be uniformly sampled form K_r.
- When sampling f_i from K_t, there is a 0.9 probability that it is the first knowledge piece. Otherwise, it will be uniformly sampled from the rest of the knowledge pieces. The reason behind is that the first knowledge piece in K_t is usually the central topic sentence of a passage and is therefore more likely to be discussed in a dialogue.
- Each knowledge piece in K_t and K_r can only be sampled at most once.

A.3 Other Datasets

Different datasets may exhibit different dialogue flow patterns. Therefore, it is reasonable to summarize them manually. For other datasets, we suggest the following steps to design the dialogue flow:

- Calculate the distribution d_1 of the number of knowledge pieces corresponding to each utterance.
- Calculate the distribution d_2 of sources of knowledge pieces. For example, in WoW, the majority of knowledge pieces come from "chosen-topic-passages".

Set.	Models	WoW Seen						WoW Unseen					
		B-4	R-L	F-1	KF-1	PPL($\Downarrow$)	ACC	B-4	R-L	F-1	KF-1	PPL($\Downarrow$)	ACC
KA	$m = 0$	15.25	31.83	34.96	51.38	9.32	-	14.41	31.30	34.13	51.05	10.13	-
	$m = 2$	**15.84**	32.18	**35.30**	**53.37**	9.36	-	**14.67**	31.67	**34.36**	53.03	10.22	-
	$m = 4$	15.67	**32.36**	35.13	52.62	9.32	-	14.38	**31.74**	34.32	**53.05**	**10.12**	-
	$m = 6$	15.27	32.02	34.59	52.57	**9.30**	-	14.38	31.39	33.61	52.53	10.15	-
KU	$m = 0$	5.43	18.31	19.96	20.11	20.97	22.61	5.16	18.14	19.30	19.80	25.55	22.12
	$m = 2$	5.89	18.56	20.20	**20.55**	**21.28**	**23.64**	**5.42**	18.47	**19.68**	**20.80**	**25.67**	**23.44**
	$m = 4$	**5.92**	**18.92**	**20.35**	20.43	20.86	23.49	5.39	18.58	19.60	20.32	25.46	23.42
	$m = 6$	5.49	18.82	20.09	19.90	**21.28**	23.47	5.21	**18.67**	19.50	20.54	25.39	23.37

Table 4: Impact of m on WoW [%].

Models	B-4	R-L	PPL($\Downarrow$)
$m = 0$	4.00	19.84	15.30
$m = 1$	4.26	**20.40**	14.52
$m = 2$	**4.27**	20.36	14.54
$m = 3$	4.18	20.31	**14.37**

Table 5: Impact of m on PersonaChat [%].

- Design a heuristic sampling strategy based on the results above. Specifically, when determining the knowledge pieces corresponding to each utterance, we need to first sample how many knowledge pieces are needed based on d_1, and then sample the specific knowledge pieces based on d_2.

B Case Study

In Table 6, we show a synthetic dialogue generated by our framework for WoW. The dialogue flow suggests smooth topic shifts within this synthetic dialogue, from the look of narcissus to their genus and then to their history. Regarding the utterance generated by the LM, they are grounded to the corresponding knowledge and are not simply copied. Overall, this synthetic dialogue is fluent, smooth and coherent. Nevertheless, there is still a gap between it and the human-generated dialogues. For example, in the last turn of the synthetic dialogue, the "adjacent areas of southwest Europe" is missed, which indicates that the knowledge is not expressed accurately enough.

C Impact of m

As described in Section 4.2, m is the number of subsequent knowledge pieces retained from F^s when generating each utterance. The value of m determines how much of the future dialogue information the dialogue content realization model can see. In Table 4 and Table 5, we explore the impact of m on the performance of the downstream model. Note that, we choose $m \in \{0, 2, 4, 6\}$ for WoW, since only the utterances of "wizard" have grounded knowledge piece.

From the results, we can see that if the dialogue content realization model does not see any future information ($m = 0$), then the downstream model has poor performance. When $m > 0$, the impact of m on the downstream model is not significant. Considering the computational efficiency, we set $m = 2$ on WoW and $m = 1$ on PersonaChat.

D More Results in the Low-resource Scenarios

The results of using 1/4 and 1/8 of the original training data are shown in Tables 7 and 9. Our SynDG consistently achieves performance improvements.

E Results with T5-Base

The results of using T5-Base instead of T5-Large are shown in Tables 8 and 10. Using T5-Base can also achieve noticeable performance gains.

Synthetic Dialogue	Dialogue Flow
[Apprentice]: I love the flowers narcissus, do you have any?	[none]
[Wizard]: I have some of them in my garden. They are white or yellow and have contrasting coloured tepal.	Narcissus (plant) | The flowers are generally white or yellow (also orange or pink in garden varieties), with either uniform or contrasting coloured tepals and corona.
[Apprentice]: Cool, I used to have some in my garden but now I don't have the space to grow my own flowers.	[none]
[Wizard]: It is quite a small family, but they are considered springtime perennials.	Narcissus (plant) | Narcissus is a genus of predominantly spring perennial plants in the Amaryllidaceae (amaryllis) family.
[Apprentice]: Sounds pretty, they're very common in flower gardens. I wonder who discovered them though.	[none]
[Wizard]: They were well known in the ancient civilisation, but were formally described by Linnaeus in 1753.	Narcissus (plant) | "Narcissus" were well known in ancient civilisation, both medicinally and botanically, but formally described by Linnaeus in his "Species Plantarum" (1753).
[Apprentice]: Wow that's a long time ago. Do you know the origin of the name Narcissus?	[none]
[Wizard]: Well, the exact origin of the name "Narcissus" is unknown, but it is believed to be related to the Greek word for intoxicated.	Narcissus (plant) | The exact origin of the name "Narcissus" is unknown, but it is often linked to a Greek word for intoxicated (narcotic) and the myth of the youth of that name who fell in love with his own reflection.
[Apprentice]: That's interesting, I wonder if they're native to Europe still.	[none]
[Wizard]: I do not know, but the species arose in the Late Oligocene and Miocene epochs in the Iberian peninsula.	Narcissus (plant) | Narcissus (plant) | The genus arose some time in the Late Oligocene to Early Miocene epochs, in the Iberian peninsula and adjacent areas of southwest Europe.

Table 6: Case study

Set.	Models	WoW Seen						WoW Unseen					
		B-4	R-L	F-1	KF-1	PPL(⇓)	ACC	B-4	R-L	F-1	KF-1	PPL(⇓)	ACC
KA	BB 1/4	13.08	29.65	32.46	47.34	10.03	-	12.52	29.15	31.53	46.36	10.84	-
	BB-SynDG 1/4	15.10	32.07	34.57	52.27	10.27	-	14.13	31.43	33.52	51.43	11.03	-
	BB 1/8	12.32	28.61	31.24	45.51	10.46	-	11.81	28.40	30.78	43.97	11.19	-
	BB-SynDG 1/8	13.50	30.36	32.73	47.02	10.56	-	12.66	30.03	32.28	46.52	11.33	-
KU	BB 1/4	4.30	17.15	18.57	17.66	21.73	19.09	4.75	17.51	18.56	18.37	25.99	20.54
	BB-SynDG 1/4	5.12	18.31	19.21	19.67	23.88	20.93	5.17	18.43	19.08	19.50	28.70	21.49
	BB 1/8	3.93	16.89	18.13	16.31	22.47	18.01	4.30	17.49	18.52	16.85	27.45	18.96
	BB-SynDG 1/8	4.61	17.53	18.53	17.50	24.82	19.19	4.51	17.67	18.29	17.44	29.37	20.81

Table 7: More results in the low-resource scenarios on WoW [%].

Set.	Models	WoW Seen						WoW Unseen					
		B-4	R-L	F-1	KF-1	PPL(⇓)	ACC	B-4	R-L	F-1	KF-1	PPL(⇓)	ACC
KA	BB	14.60	30.98	34.21	50.27	**9.15**	-	13.94	30.70	33.46	49.33	**9.97**	-
	BB-SynDG-Large	**15.84**	**32.18**	**35.30**	**53.37**	9.36	-	**14.67**	**31.67**	**34.36**	**53.03**	10.22	-
	BB-SynDG-Base	15.34	31.96	34.74	50.27	9.99	-	14.09	31.23	34.01	50.07	10.95	-
KU	BB	5.52	18.32	20.11	19.85	**19.88**	22.32	5.00	18.29	19.45	18.98	**24.67**	22.01
	BB-SynDG-Large	**5.89**	**18.56**	**20.20**	**20.55**	21.28	**23.64**	**5.42**	**18.47**	**19.68**	**20.80**	25.67	**23.44**
	BB-SynDG-Base	5.64	18.45	20.18	19.75	21.91	22.94	5.32	18.44	19.49	20.17	25.89	23.02

Table 8: Results with T5-Base on WoW [%].

Models	B-4	R-L	PPL(⇓)
GPT2 1/4	2.70	17.51	20.13
GPT2-SynDG 1/4	3.16	18.28	17.62
GPT2 1/8	2.23	15.56	24.61
GPT2-SynDG 1/8	3.06	18.09	19.26

Table 9: More results in the low-resource scenarios on PersonaChat [%]

Models	B-4	R-L	PPL(⇓)
GPT2	2.70	17.51	20.13
GPT2-SynDG-Large	**4.26**	**20.40**	**14.52**
GPT2-SynDG-Base	3.96	19.33	16.55

Table 10: Results with T5-Base on PersonaChat [%]

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☑ A4. Have you used AI writing assistants when working on this paper?
Grammarly is used to check the grammar.

B ☑ Did you use or create scientific artifacts?

5.1, 5.4

☑ B1. Did you cite the creators of artifacts you used?
5.1, 5.4

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We follow the the license or terms of the used artifacts.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
5.1, 5.4

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The data is safe and are commonly used by many previous efforts.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
5.1

☒ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
These are consistent with previous work.

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5.4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5.4

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We use a single run because of the high computational overhead of the model. Also, we observe stable performance.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5.4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
5.3

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Our annotation is simple and does not use visualization tools. The principles of annotation are given in Section 5.3.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Our annotations are few and simple. Three authors of this paper performed the annotation.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
The annotators are the authors of this paper. We all agree to the use of these data.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our annotation does not include any ethic issues.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We have only 3 annotators, all of whom are the authors of this paper.