Retrieval-free Knowledge Injection through Multi-Document Traversal for Dialogue Models

Rui Wang[1,6*], **Jianzhu Bao**[1,5], **Fei Mi**[2†], **Yi Chen**[1,6], **Hongru Wang**[4], **Yasheng Wang**[2], **Yitong Li**[2,3], **Lifeng Shang**[2], **Kam-Fai Wong**[4], **Ruifeng Xu**[1,5,6†]
[1]Harbin Institute of Technology, Shenzhen, China
[2]Huawei Noah's Ark Lab, [3]Huawei Technologies Co., Ltd
[4]The Chinese University of Hong Kong, [5]Peng Cheng Laboratory, Shenzhen, China
[6]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
ruiwangnlp@outlook.com, mifei2@huawei.com, xuruifeng@hit.edu.cn

Abstract

Dialogue models are often enriched with extensive external knowledge to provide informative responses through a retrieval-augmented pipeline. Nevertheless, retrieval-augmented approaches rely on finely annotated retrieval training data and knowledge-grounded response generation data, making it costly to transfer. To tackle this challenge, this paper proposed a retrieval-free approach, KiDG, by automatically turning knowledge documents into simulated multi-turn dialogues through a Multi-Document Traversal algorithm. The simulated knowledge-intensive dialogues constructed by KiDG in one domain can be easily used to train and enhance pre-trained dialogue models' knowledge w.r.t. this domain without costly annotation. We conduct extensive experiments comparing retrieval-augmented models and a variety of retrieval-free models. We found that dialogue models enhanced with data simulated with KiDG largely outperform state-of-the-art retrieval-free methods, and it achieves comparable performance compared to retrieval-augmented methods while being better, and cheaper at domain transfer. We have released the code and data at https://github.com/DevoAllen/KiDG.

1 Introduction

Knowledge plays a crucial role in dialogue systems, which is helpful in improving the informativeness, logicality, and reliability of generated responses. To encourage Pretrained Dialogue Models (PDMs) to produce knowledge-grounded responses, existing research mainly follows two lines: retrieval-based (Dinan et al., 2019; Zhao et al., 2020) and retrieval-free(Xu et al., 2022).

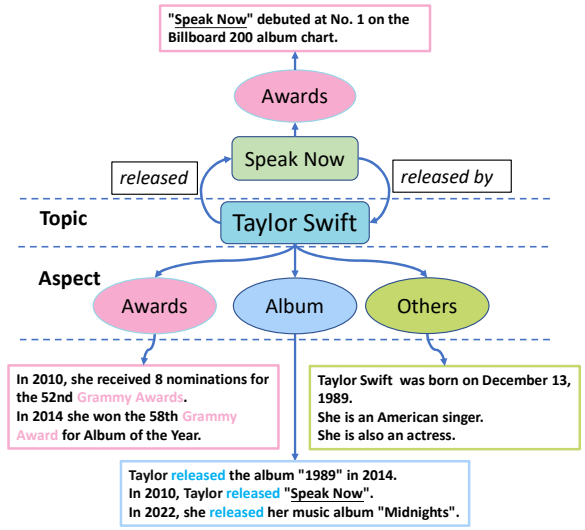

Figure 1: The structure of the document "*Taylor Swift*" and her album "*Speak Now*". A document usually concentrates on one **topic**. Sentences within the document describe different **aspects** of the document's topic.

The retrieval-based paradigm explicitly provides DMs with ready-to-use knowledge by a knowledge retriever (Karpukhin et al., 2020) and trains DMs to apply it during response generation on knowledge-grounded conversations (Izacard and Grave, 2021). Yet, it relies on finely annotated retrieval training data as well as knowledge-grounded conversations, making it costly to apply and transfer. The retrieval-free paradigm aims to directly incorporate knowledge into DMs through training on simulated knowledge-intensive conversational data. Therefore, it's more convenient and cheaper to be applied at scale.

Recently, Dai et al. (2022) proposed dialogue inpainting to automatically transform a single document into a multi-turn dialogue. However, dialogue data produced by the existing inpainting approach only considers sentences from the same

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

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6608–6619
July 9-14, 2023

document, and they are always in the original order as in the source document. This limitation harms the model trained with such data to be inflexible to different knowledge flows during conversation. Once the conversation context changes, PDMs might fail to recall the correct knowledge, which has been demonstrated in our experiments (§3.3). Xu et al. (2022) incorporate topic information into PDMs using separate topic adapters in a retrieval-free manner, yet only limited topics are considered at a coarse granularity.

In this paper, we focus on exploiting the connections among sentences from multiple knowledge documents to construct knowledge-intensive and topic-diversified dialogues at a fine granularity. As shown in Fig.1, we notice that a knowledge document typically concentrates on a particular topic, and the sentences within the same document usually talk about different aspects of the topic. Moreover, there are also relations between the topics of different but related documents. We attempt to infuse these fine-grained relations into simulated dialogues by imitating human conversation behaviors: diving into an aspect of the document's topic, then jumping into a related aspect or a new topic at the right time to attract the listener's interest based on background knowledge. However, there are several challenges to constructing such dialogues: (1) how to distinguish the fine-grained aspects of a particular topic within a single document; (2) how to gather multiple topic-related documents together in an efficient way when we are faced with abundant documents; and (3) how to simulate in-depth and topic-diversified dialogues according to the rich aspect and topic relationships both inside and among knowledge documents.

To overcome the above challenges, we propose **K**nowledge-**i**ntensive **D**ialogues Generation with Aspect based Topic **G**raph (**KiDG**), as shown in Figure 2. Firstly, KiDG automatically builds an **A**spect **G**raph (**AG**) to capture the aspect relevance among sentences within a single document. Then, it connects AGs of topic-related documents to construct a larger **A**spect-based **T**opic **G**raph (**ATG**), which both retains the aspect relevance inside each document and further models the topic relationship among multiple associated documents. Finally, KiDG employs a **M**ulti-**D**ocument **T**ravelsal (**MDT**) algorithm to walk through ATG and sample a series of aspect/topic-related sentences, which are organized in a logically coherent order and turned into a simulated dialogue in a human-like manner.

With the proposed KiDG, we automatically construct a high-quality knowledge-intensive dialogue dataset, KiDial. Experimental results show that PDM further trained on KiDial achieves state-of-the-art performance compared with retrieval-free baselines, and shows competitive ability with retrieval-based models. In addition, we scale up the knowledge corpus to produce three versions of KiDial (i.e., small, base, and large) and compare the performance of PDMs pre-trained on them. We find that a larger size of KiDial could further enhance PDMs to be both more proactive and more knowledgeable during conversation while maintaining low hallucination.

2 Method

Given a knowledge corpus $\mathcal{D} = \{d_i\}_{i=1}^{M}$, where $d_i = \{e_i, \mathcal{S}_i\}$ is a knowledge document composed of a number of sentences $\mathcal{S}_i = \{s_i^j\}_{j=1}^{N}$, and a title e_i, which usually indicates the topic of the document and is utilized as the entry for web search. We propose **K**nowledge-**i**ntensive **D**ialogs Generation with Aspect based Topic **G**raph(**KiDG**) to automatically construct a large-scale simulated dialogue corpus. As illustrated in Fig.2, the procedure of KiDG can be divided into 3 stages: (1) Construct an **A**spect **G**raph (**AG**) to capture the aspect relations inside a single document. (2) Construct an **A**spect-based **T**opic **G**raph (**ATG**) to associate the AGs of topic-related documents. (3) Use a **M**ulti-**D**ocument **T**raversal (**MDT**) algorithm to traverse ATG and simulate knowledge-intensive and logically-coherent dialogues.

2.1 Aspect Graph Construction

For each document d_i, we construct an **A**spect **G**raph (AG) $\mathcal{A}_i$ as a weighted fully connected bidirectional graph, whose nodes are the sentences $\mathcal{S}_i$ within d_i, and edge weights $w_i^{u,v}$ are BertScore (Zhang et al., 2019a) similarity between two sentences $s_i^u, s_i^v \in \mathcal{S}_i$ to measure the fine-grained aspect relevance.

$$w_i^{u,v} = BertScore(s_i^u, s_i^v) \quad (1)$$

BertScore computes text similarity based on token-level cosine similarity using the pre-trained contextual embeddings from BERT (Devlin et al., 2018). The intuition is that sentences describing the same or similar aspects of a topic tend to have a higher semantic similarity and lexical overlap, as shown

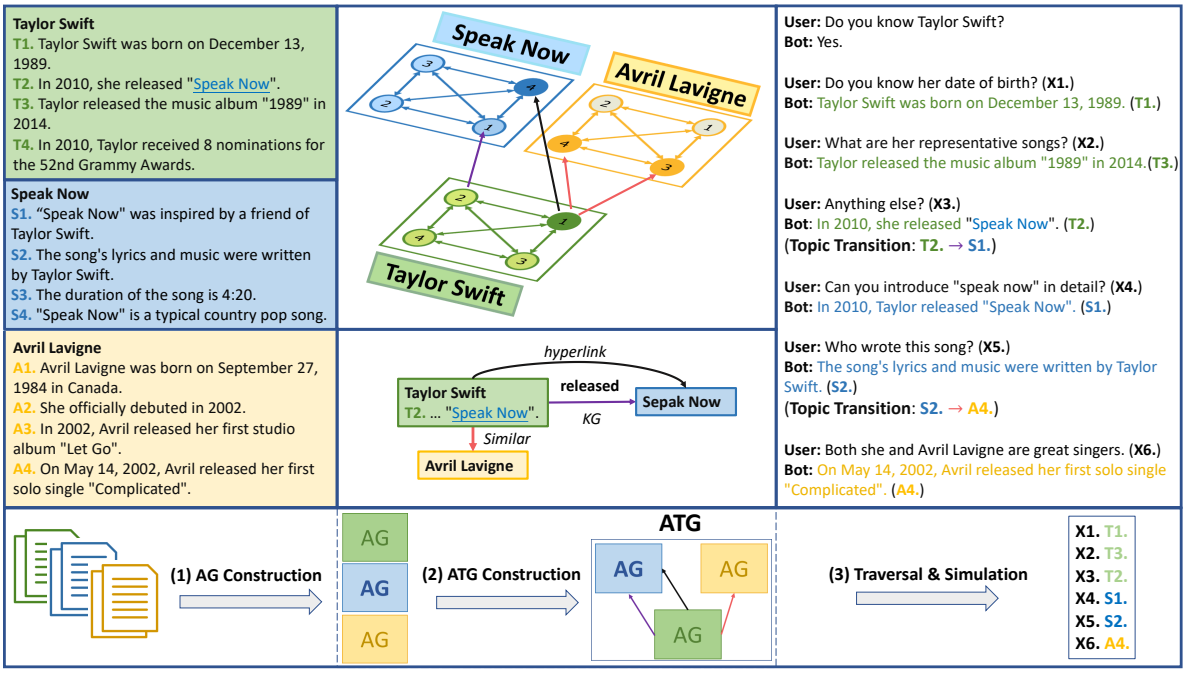

Figure 2: Overview of KiDG. To construct in-depth and topic-diversified dialogues, KiDG (1) builds an AG for each document to distinguish aspects; (2) connects topic-related documents to construct ATG; (3) utilizes the traversal algorithm MDT to walk through ATG to sample a sequence of knowledge sentences and transform them to a simulated dialogue. Note that, the Representative Nodes are marked in the darkest color, **black**, **purple** and **red** arrows represent the topic relations derived from Web Hyperlink, Knowledge Graph and Word vector respectively.

in Figure 1. Thus BertScore is relatively suitable for this scenario.

2.2 Aspect-based Topic Graph Construction

We further associate $\mathcal{A}_i$ of each document d_i with other AGs and build a large **A**spect-based **T**opic **G**raph (ATG) $\mathcal{G}$ to capture the coarse-grained topic relevance between documents. In order to reduce the size of ATG, we only connect $\mathcal{A}_i$ with closely related AGs based on three sources of clues: Web Hyperlink, Knowledge Graph, and Word Vector.

• **Web Hyperlink** is usually a digital reference providing direct access from one distinctively marked term in a document to another in a different document. This kind of clue naturally exists in online knowledge documents and the marked term is usually the title (topic) of the linked document. Therefore, the documents connected by a hyperlink are naturally topic-related.

• **Knowledge Graph** is a knowledge base that stores interlinked descriptions of entities as triples. For example, (*Taylor Swift*, *released*, *Speak Now*) describes that "*Taylor Swift* has a *released* relation with *Speak Now*". We observe that most titles (topics) of knowledge documents are also entities in the Knowledge Graph. Thus, we use it to complement the missing relations in hyperlinks.

• **Word Vector** is a continuous dense vector for word representation (Mikolov et al., 2013). Semantically correlated words/phrases usually have a higher cosine similarity. For instance, *Taylor Swift* and *Aviral Lavigne* are close in the vector space since they are both singers. Terms associated by this approach are likely to be described in similar contexts (comparing the documents of *Taylor Swift* and *Aviral Lavigne* in Figure 2) but are usually not directly linked by the previous two methods.

We connect $\mathcal{A}_i$ with the AGs whose document titles are directly linked by Web Hyperlink or Knowledge Graph or share the top-5 similarity with the title of $\mathcal{A}_i$ in the Word Vector space. Note that, the resulting ATG $\mathcal{G}_i$ is not a fully connected graph. We only add directed edges from the Representative Nodes of $\mathcal{A}_i$ to those of the associated AGs to reduce computing overhead. The Representative Nodes are marked in the darkest color in Figure 2. The weights for the newly added cross-topic edges are computed the same as Eq. 1.

• **Representive Nodes** are the sentence nodes in $\mathcal{A}_i$ whose contents are most likely to be mentioned when talking about a particular topic. We denote the representative nodes as $\mathcal{V}_i \in \mathcal{S}_i$ and set them

to be the nodes with the top-k highest sum of edge weights in $\mathcal{A}_i$:

$$\mathcal{V}_i = \underset{v \in \mathcal{S}_i}{ArgTopk}\,(k, \sum_{u \in \mathcal{S}_i, u \neq v,} w_i^{u,v}), \quad (2)$$

$$k = max(1, \lfloor \frac{|\mathcal{S}_i|}{5} \rfloor). \quad (3)$$

$|\mathcal{V}_i| = k$ is the number of representative nodes we select in $\mathcal{A}_i$. Since the edge weight $w_i^{u,v}$ measures the aspect relevance between two sentences (§2.1), the obtained representative nodes are analogous to the topic centers which we often use to start a conversation around a topic and extend to other relevant contents along with the aspect relevance.

2.3 Dialogue Simulation based on Multi-Document Traversal

2.3.1 Multi-Document Traversal Algorithm

We propose the Multi-Document Traversal (MDT) algorithm, which is a refined Weighted Random Walking to obtain a sequence of aspect/topic-related sentences by walking through ATG following two principles: (1) sentences that are similar in the aspect or topic level should be placed close to each other in this sequence; (2) if sentences from one aspect (or topic) are nearly exhausted, the probability of sampling the next sentence from another one should be increased. To meet the above principles, MDT is composed of the Aspect Graph Walking (AGW) algorithm to traverse the AG of a document, and the Adaptive Main Document Transition (AMDT) to fulfill the topic transfer.

MDT executes T time steps and returns a sentence sequence $\mathcal{Y}$. At each step, it mainly focuses on one document d_i, called the Main Document. First, MDT utilizes AGW to traverse d_i in a relevance-first way and append a sentence into $\mathcal{Y}$, to concentrate on a specific aspect of d_i. Although AGW could dive into an aspect of d_i and traverse sentences of another aspect, it could not fulfill topic transitions, i.e., visiting another document. Then AMDT will consider all of the Representative Nodes included in $\mathcal{Y}$ and determine whether jump into one of the documents they connected.

Aspect Graph Walking The Aspect Graph Walking (AGW) algorithm is a Weighted Random Walking algorithm. In step-1, the AGW samples a start sentence y_1 from $\mathcal{V}_i$ with uniformed probability, then $\mathcal{Y} = [y_1]$. In step-t, the AGW samples the next sentence y_t from the transition distribution $\{w_i^{y_{t-1}, y_t}, y_t \in \mathcal{S}_i\}$ of y_{t-1}. The more relevant the sentence is to y_{t-1}, the higher the chance it will be selected, in order to dive into an aspect. After T steps, we obtain the sequence $\mathcal{Y} = [y_1, y_2, ..., y_T]$. For repeated sentences in $\mathcal{Y}$, we only keep the one that appears first.

Adpative Main Document Transition In a complete execution of MDT, the AGW first chooses a sentence in Main Document d_i, then the Adaptive Main Document Transition (AMDT) determines whether to change the Main Document.

Towards this purpose, AMDT maintains a set of Transition Acts (TA) and assigns each action a transition weight to balance between in-depth discussion and topic transition. Denote the Transition Weights as TW, the next transition behavior is sampled from TA based on $\texttt{softmax}(\text{TW})$. TA records the reachable nodes outside d_i, and TW stores the probability to choose corresponding nodes.

In step-1, TA=$[No]$ and TW=$[1.0]$, "No" means "do not change the Main Document". Obviously, AMDT would not change the Main Document at the beginning to dive into the topic of d_i. In step-t, AMDT first samples a node r_j following uniform distribution from all of the visited Representative Nodes in $\mathcal{Y}$. The set of outside nodes connected to r_j is O_j. Denote the list of edge weights between r_j and the nodes in O_j as W_j, we have

$$\begin{aligned} &\text{TA} = \text{concat}(No, O_j); \\ &w_{No} = \max(W_j) \\ &\text{TW} = \text{concat}(w_{No}, W_j) \end{aligned} \quad (4)$$

We set the weight for the "No" action to be the largest among all the actions. It is worth mentioning that the representative nodes in Q are not always from the current Main Document d_i since the Main Document maybe has already changed several times before d_i. Considering the next Main Document based on $\mathcal{Y}$ rather than d_i could boost the topic's diversity of dialogues. If speakers are discussing d_i, when the information from d_i is exhausted, they tend to change the conversation topic. Hence, to simulate human conversation behaviors, the vanilla softmax is not applicable for the AMDT. So we introduce $\theta(t)$, an *adaptive temperature* which increases with the number of visited sentences from d_i, and derive the final transition probability $Q(t)$.

$$\begin{aligned} &\theta(t) = \tau \cdot |V \cap d_i| \\ &Q(t) = \text{softmax}(TW/\theta(t)) \end{aligned} \quad (5)$$

The higher the $\theta(t)$ is, the greater the probability of switching Main Document is. In practice, we set $\theta(t)$ to 2. In addition, if y_i and y_{i+1} do not belong to the same document, a Topic Transition Prompt will be added between them to provide topic transition hints. e.g., "Except for *A*, do you know *B*?" and "Yes,...".

2.3.2 Dialogue Simulation

We leverage the dialogue inpainting model (Dai et al., 2022), which takes the sequence $\mathcal{Y}$ as the utterances from one speaker and repairs another speaker's utterances in an autoregressive manner. To provide basic topic relations, we design starting prompt p, e.g., "Have you ever heard of `A`?" and append it before $\mathcal{Y}$. Now the $\mathcal{Y}$ will be: $\mathcal{Y} = \{p, y_1, y_2, ..., y_T\}$. We first feed $\{p, \texttt{[m]}, y_1\}$ to the inpainting model to get x_1, then feed $\{p, x_1, y_1, \texttt{[m]}, y_2\}$ to generate x_2. We keep doing this until the conversation is complete. Note that the Topic Transition Prompt already contains two speakers' words, hence there will be no repaired utterances between them.

3 Experiment

In this section, we show that simulated dialogues from KiDG could boost PDMs' performance in knowledge-grounded response generation tasks. From our exhaustive experiments, we found that the enhanced PDMs obtain state-of-the-art performance under retrieval-free settings and even achieve comparable performance compared with retrieval-based approaches. In addition, as the simulation dialogue scale increases, the PDMs tend to generate more proactive and precise responses.

3.1 KiDial Construction and Pre-training PDMs

We apply KiDG to document corpora to generate a large dialogue dataset KiDial with the open-sourced knowledge graph [1] and work vectors [2], containing a Small version based on a knowledge corpus of KdConv(Zhou et al., 2020), Base, and Large versions originating from a well-known Chinese encyclopedia website[3]. The dialogue inpainting model is initialized from BART-Large and trained in a large QA dataset and 0.9M conversations translated from WikiDialog(Dai et al., 2022). We input $\{p, \texttt{[m]}, y_1\}$ and force the model to generate $\{x_1\}$, rather than feeding $\{p, \texttt{[m]}, y_1, y_2, ..., y_T\}$ as Dai et al. (2022) did. In this way, we could eliminate the gap between training and inference of the inpainting model.

Datasets Scale	Small	Base	Large
# dialogues	36K	751K	3.75M
# documents	12K	214K	1.3M
# utterances	1.38M	15.2M	75.9M
# topic-turns	35K	730K	3.66M
# topic-turns per dialogue	1.05	1.03	1.02
# contexts per knowledge	2.17	2.27	3.29
rate of eligibility	93%	90%	94%

Table 1: Statistics of KiDial. The high rate of eligibility means that most of the dialogues are logically coherent, and Cohen's Kappa score is 0.76.

Then we sample 50 dialogues from each of the 3 versions of KiDial respectively and invited 2 human annotators to judge whether the knowledge sentences in dialogues are highly related to context and whether the topic transitions are proper, denoted as the *rate of eligibility* in Table 1. The dataset statistics are shown in Tab. 1.

We feed the conversations from KiDial to enhance the PDMs. We use the BART-Large from Shao et al. (2021) and CDialGPT trained on LCCC-Large(Wang et al., 2020). The pre-training setting for CDialGPT is the same as Wang et al. (2020). To enhance BART, we treat the response generation as a text-infilling task(Lewis et al., 2020) and add role labels `[S1]` and `[S2]` to help BART distinguish the utterances from different speakers. All of the training is finished in 8 Nvidia V100 GPUs.

3.2 Evaluation on Knowledge Grounded Dialog Datasets

3.2.1 Experiment Setup

Datasets We empirically measure the impact of KiDial on knowledge-grounded dialogue systems. Hence we construct the KiDial-Small based on the knowledge corpus of KdConv(Zhou et al., 2020) in §3.1, in which case, the dialogues of KdConv are the perfect source to evaluate the PDMs trained on KiDial-Small.

Baselines We select baselines under 2 experiment settings, including both the retrieval-based and the retrieval-free settings.

• **Retrieval-based** These methods are combined with a retriever and the PDM. We utilize **BM25** and **DPR**(Karpukhin et al., 2020) as retrievers. The **DPR** is combined by two Chinese BERT [4].

[1] https://github.com/ownthink/robot
[2] https://ai.tencent.com/ailab/nlp/en/embedding.html
[3] https://baike.baidu.com/
[4] https://huggingface.co/bert-base-chinese

The Top-3 retrieved knowledge is concatenated to dialogue history as input for **CDialGPT** and **BART**. Meanwhile, A special token [KNW] is added to differentiate the knowledge and history. Then the retriever and PDMs are jointly fine-tuned on KdConv's training set to learn how to generate grounded responses.

• **Retrieval-free** In this setting, the PDMs need to fulfill the response generation without knowledge retrieval. **KnowExpert**(Xu et al., 2022), where knowledge is infused to n adapters with documents from n topics. Since the documents of KdConv are already split into 3 topics, we set 3 topic adapters here. Then KnowExpert is finetuned on KdConv's training set for adaption.

We construct other baselines by enhancing the PDMs with the following knowledge resources and then finetuning them on KdConv's training set. Note that BART+X means the BART is trained on dialogue dataset X, and the same applies to CDialGPT.

1) **KB**. Knowledge documents are split into pseudo-dialogues to train PDMs. 2) **MD** is the set of sentence sequences the KiDG constructed. 3) **SDial**(Dai et al., 2022). Sentence sequences in **KB** are transformed into dialogues by the dialogue inpainting model. 4) **KiDial-S**. The PDMs are trained on KiDial-Small. 5) **Shuff**. With two sentences next to each other as a group, we shuffle the conversations from **SDial**.

• **Ground-truth Knowledge** In addition, we offer ground-truth knowledge to the retrieval-based model to provide ceiling performance on KdConv. We also report the performance of **HRED**(Zhou et al., 2020), which has a memory module to incorporate knowledge into responses.

3.2.2 Metrics

Automatical Metrics We utilize the Perplexity(PPL) of the ground-truth response, the average of BLEU (Papineni et al., 2002), the Uni-gram F1 and Distinct-2 (Li et al., 2016) to evaluate the generation results automatically.

Human Evaluation. We randomly select 100 dialogue history and response pairs for evaluation. Following Zhou et al. (2020), we evaluate the generation results from 2 different perspectives. 1) **Fluency** (Flu.) To test whether the generated utterances are grammatically correct. 2) **Coherency**(Coh.) The response must be coherent to grounded knowledge at the utterance level and relevant to dialogue context at the dialogue level.

We performed a pairwise comparison of the responses generated by PDMs trained on KiDial-S with other baseline models. Three annotators evaluated dialogues based on the above two metrics to determine which one is better. In this comparison, the model that outperforms its counterpart receives 2 points, while the underperforming model gets 0 points. In the case of a tie, each model is awarded 1 point. The average score was used to measure overall performance and Kappa was reported in Table 4.

Model	Bleu-Avg	F1	Dist-2	PPL↓
Ground-truth Knowledge				
HRED	18.87	-	11.03	**11.15**
CDial	24.31	38.23	12.55	6.41
BART	**30.79**	**44.35**	13.61	6.07
Retrieval-based				
CDial+BM25@3	13.72	24.87	12.35	8.59
CDial+DPR@3	**19.28**	**32.16**	**14.33**	**7.22**
BART+BM25@3	22.35	37.70	**20.56**	**7.45**
BART+DPR@3	**28.07**	**42.97**	12.52	7.86
Retrieval-free				
KnowExpert	16.37	30.72	17.10	9.47
CDial	13.02	27.06	12.88	8.11
CDial+Shuff	13.81	28.09	13.51	7.52
CDial+KB	15.02	29.65	15.14	6.89
CDial+MD	16.69	31.04	16.84	**6.46**
CDial+SDial	15.79	30.26	16.01	8.43
CDial+KiDial-S	**17.52**	**31.60**	**18.06**	7.45
BART	18.74	30.29	14.80	7.37
BART+Shuff	19.70	34.08	17.95	7.13
BART+KB	20.34	37.67	12.41	6.54
BART+MD	21.70	35.08	**18.53**	7.50
BART+SDial	23.68	39.39	14.35	**6.29**
BART+KiDial-S	**27.72**	**41.45**	13.27	6.77

Table 2: Automatic evaluation results on KdConv. The best results under different settings are in **bold**.

3.2.3 Automatic Evaluation

The evaluation results on KdConv are shown in Table 2. After training on KiDial-Small, BART and CDialGPT achieve state-of-the-art performance on most of the metrics compared with retrieval-free paradigms. Moreover, they even outperform retrieval-based methods with a weak retriever, i.e., BM25, and get comparable performance compared to those with DPR. These results show the **KiDial** could significantly improve the PDMs in knowledge memorization and understanding. **Shuff** brings less improvement than others, which means the proper organization of knowledge is essential. Hence PDMs trained on **MD** outperform those trained on **KB**, for the fine-grained topic relations in **MD** help the model understand knowledge better. Moreover, **CDial+MD** outperforms **KnowEx-**

Model	Bleu-Avg	F1	Dist-2
BART			
w/ Shuff	6.76	16.91	4.59
w/ KB	7.45	16.50	12.65
w/ MD	7.61	16.68	13.33
w/ SDial	11.61	21.26	11.32
w/ KiDial-S	**12.68**	**21.80**	**15.86**
CDial			
KnowExpert	4.53	10.78	8.47
w/ Shuff	4.14	10.35	10.16
w/ KB	4.36	10.29	4.53
w/ MD	6.50	14.05	18.43
w/ SDial	8.09	17.47	9.18
w/ KiDial-S	**8.61**	**19.04**	**17.36**

Table 3: Zero-shot performance.

BART-L	Flu.	Coh.	κ
KiDial-S **v.s.** SDial	1.11	1.36	0.49
KiDial-S **v.s.** DPR@3	1.09	1.24	0.51

Table 4: Human evaluate performance of BART on KdConv's test set. A score larger than 1 means that our model is better than its counterparts. κ means the Fleiss' Kappa.

pert, which proves that the coarse topic relations in KnowExpert are insufficient.

Abalation Study Without KiDG, PDMs trained on **KB** and **SDial** are worse than those trained on **MD** and **KiDial-S**, because (1) there are no explicit relations between topics in **KB** and **SDial**. PDMs need to understand those relations by themselves; (2) the context diversity for a knowledge text in **SDial** is limited. However, as shown in Table 1, the flourishing contexts in **KiDial** could help the model learn knowledge from various perspectives. Without dialogue inpainting, **KB** and **MD** are worse than **SDial** and **KiDial-S**, for the inpainted dialogues could provide more context information. It is worth mentioning that the PDMs obtain larger performances boost from **MD** to **KiDial-S** than from **KB** to **SDial**. We believe that it's because the sentence sequences obtained by our method are infused with fine-grained topic information than the original sentence order of documents, which helps the PDMs better absorb knowledge.

3.2.4 Human Evaluation

In addition, Table 4 shows the human evaluation results. The results reveal that dialogues from BART+KiDial tend to be more coherent with dialogue history and ground-truth knowledge while maintaining high fluency.

3.3 Analysis

Zero-shot Performance A more knowledgeable PDM will perform better in the zero-shot scenario. However, the retrieval-based methods are not applicable in the zero-shot scenario, for they need grounded dialogues to learn how to incorporate knowledge into responses. As shown in Tab.3, the PDMs pre-trained in KiDial-Small outperform other baselines, which proves that KiDial-Small is a better source for PDMs to learn from. Note that models trained on SDial and KiDial-S are significantly better than other baselines. We attribute this to the inpainted dialogues providing more context information to serve as a hint to elicit the knowledge.

Generalization to Different Contexts In the previous section, we have proven that although KiDial and other knowledge sources originated from the same document corpus, PDMs trained on KiDial still perform better than other baselines. In this section, we show that KiDial could equip PDMs with the generalization ability to diversified contexts.

If a model is better at handling diversified conversation contexts, it will have a more stable performance when injecting the same knowledge into responses in different contexts. Hence we assess the model according to the variance of uni-gram F1 score on the test samples grounded by the same knowledge. We first identify the knowledge in KdConv which is grounded in more than one dialogue response and these responses form a **unit**. Then all of the units are grouped according to the number of different contexts in them. Then we calculate the variance of the F1 score for responses in every unit in the specific group, then report the average of these variances. We illustrate the results in Fig. 3. We can conclude that when the context of knowledge becomes more complex, the stability of model performance will have greater fluctuations. But BART+KiDial-S is still lower than others, which proves that KiDial makes BART understands knowledge better and can handle more complex and diverse context environments.

Finally, we explore how the model's capabilities change as the size of KiDial increases. For PDMs trained in larger KiDial, there are no grounding dialogues to evaluate the models' abilities. Hence we employ self-talk to evaluate how much knowledge the model could generate, and how much is correct. Besides PDMs enhanced with different versions of KiDial, we also introduce BART+DPR@3 as

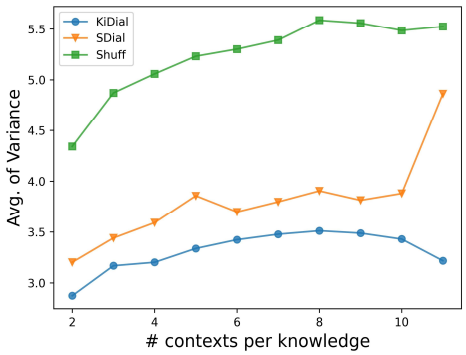

Figure 3: Impact of context diversity of knowledge and performance stability. The horizontal and vertical axis denotes the number of contexts the knowledge corresponds to and the average variance of **F1** respectively.

a comparison. The purpose of this is to validate whether it has learned the general ability to incorporate knowledge into the response during the fine-tuning on KdConv, rather than just understanding the knowledge in the dataset.

Choice of Starting Topic We provide starting topics for models to perform self-talk. We randomly select 10 topics, half of them from documents of KdConv and the other half from encyclopedia documents of KiDial (i.e., **Pedia** as shown in Table 5). Then we rewrite the topics to complete sentences as the starting utterances. For every topic, the self-talk conducts 10 rounds, and 5 history and response pairs are sampled for evaluation. Thus, there are 50 samples evaluated for every model.

Evaluation Perspectives We evaluate the generation results based on: 1) Informativeness, how much information the response contains; 2) Groundness(Thoppilan et al., 2022), how much claims in response could be associated with authoritative knowledge; 3) Proactiveness, whether the model would like to perform active information exchange or topic transition. We compare different methods by asking 3 human annotators to give absolute scores(0 for bad,1 for good, and 2 for excellent)for each response based on three metrics. We reported the average score in Table 5.

Analysis The evaluation results are shown in Table 5. BART+DPR@3 achieves the best Groundness score on topics of KdConv due to the gain brought by DPR. However, BART+DPR@3 suffers a great performance drop on topics from Pedia, which means that it has not learned a general ability to infuse knowledge into responses. Since annotating the knowledge-grounded dialogues in other

	KdConv			Pedia		
BART	**Infor.**	**Grou.**	**Proa.**	**Infor.**	**Grou.**	**Proa.**
DPR@3	1.32	**1.80**	1.44	1.03	0.77	1.32
KiDial-S	1.08	1.64	1.20	1.24	0.93	1.28
KiDial-B	**1.52**	1.24	**1.60**	**1.37**	1.47	1.40
KiDial-L	1.20	1.52	1.48	1.28	**1.57**	**1.67**

Table 5: Human evaluation results on self-talk conversations from modes. The best and the second-best results are in **bold** and underlined.

domains is a tedious process, they are difficult to transfer to other domains.

In contrast, PDMs trained on KiDial-Large or KiDial-Base do not show a large performance difference on topics from two different sources. Moreover, the larger KiDial improves the PDMs in generating more informative, precise, and proactive responses. Surprisingly, on topics of KdConv, the BART enhanced by KiDial-Base and KiDial-Large appear to be more proactive than BART+DPR@. We attribute this to that the dialogues in KiDial are composed of utterances from a knowledgeable speaker and a supportive listener. Thus, the model could learn to play both roles. When the dialogue history becomes boring, the model will introduce more information.

4 Related Work

Knowledge-grounded dialogues are helpful for enhancing pre-trained dialogue models (PDMs) (Zhang et al., 2019b; Zhou et al., 2021; Bao et al., 2020; Thoppilan et al., 2022; Mi et al., 2022) to be more knowledgeable. Existing research can be classified into two directions: retrieval-based (Dinan et al., 2019; Zhao et al., 2020; Li et al., 2020) and retrieval-free (Xu et al., 2022) paradigms.

The retrieval-based paradigm is composed of a knowledge retriever and a generator. The retriever uses sparse or dense representations (Karpukhin et al., 2020) to obtain relevant knowledge for response generation. As the input length of PDMs is limited, the retriever is responsible to fulfill fine-grained knowledge retrieval rather than a batch of relevant documents and more information is not always better, since there exists much noise. Hence the retrieval-based methods need knowledge and dialogue utterances aligned dataset to learn how to fetch knowledge and how to incorporate it into responses. However, the data annotation process is tedious and labor-intensive.

To alleviate this problem, Xu et al. (2022) proposed the retrieval-free paradigm. They first

train topic experts with documents from several topics. Then the model is fine-tuned with knowledge-grounded dialogues for adaption. However, they utilize topic relations at a coarse granularity, i.e., document-level only. Recently, Dai et al. (2022) devised the dialogue inpainting to produce knowledge-grounded dialogues: transforming documents into two-person conversations with T5 (Raffel et al., 2019). But the generated dialogues only contain sentences in the same document and always in the original order as the source document, which hurts the generalization abilities of PDMs.

5 Conclusion

In this paper, we propose KiDG, a retrieval-free approach to incorporate knowledge into PDMs by automatically turning knowledge documents into simulated dialogues. KiDG exploits both the fine-grained aspect relations in a single document and the coarse-grained topic relations between documents though Multi-Document Traversal. Our experiments show that the KiDial generated by KiDG can improve the PDMs to achieve state-of-the-art performance under retrieval-free settings and achieve performance comparable to retrieval-based methods. Our further analysis proves that a larger KiDial can enhance the PDMs to generate more proactive and informative responses.

Limitations

The simulated dialogues constructed by KiDG are a powerful source of training data for retrieval-free knowledge-grounded dialogue systems. However, there is a clear style difference between the generated utterance and the original document sentences: one is the oral expression and the other is a more formal style.

But as shown in Table 5, the PDMs trained on KiDial appear to be more proactive and knowledgeable during conversations. The generated utterances serve as a type of prompt to help the model understand the knowledge. In the meanwhile, our KiDG embeds the knowledge into different contexts, alleviating the one-to-many problem in some degree.

Although generating dialogues needs to cost GPU resources, it is still a cheaper and quicker way to acquire large-scale knowledge-intensive dialogues.

Ethics Statement

This paper proposes a method to exploit fine-grained aspect/topic-relations between documents and construct topic-diversified dialogues to enhance retrieval-free dialogue systems. The documents we used in this paper and the generated dialogues have been carefully filtered to make sure there is no offensive and toxic 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

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. 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 *Proceedings of the 39th International Conference on Machine Learning*, pages 4558–4586.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

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*, volume 33, pages 8475–8485. Curran Associates, Inc.

Fei Mi, Yitong Li, Yulong Zeng, Jingyan Zhou, Yasheng Wang, Chuanfei Xu, Lifeng Shang, Xin Jiang, Shiqi Zhao, and Qun Liu. 2022. Pangu-bot: Efficient generative dialogue pre-training from pre-trained language model.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *ArXiv*, abs/2109.05729.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. A. Krivokon, Willard James Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Hartz Søraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Díaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravindran Rajakumar, Alena Butryna, Matthew Lamm, V. O. Kuzmina, Joseph Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Huai hsin Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *ArXiv*, abs/2201.08239.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *Natural Language Processing and Chinese Computing*.

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. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B. Dolan. 2019b. Dialogpt : Large-scale generative pre-training for conversational response generation. In *Annual Meeting of the Association for Computational Linguistics*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390. Association for Computational Linguistics.

Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, Bosi Wen, Xiaoyan Zhu, Minlie Huang, and Jie Tang. 2021. Eva: An open-domain chinese dialogue system with large-scale generative pre-training. *ArXiv*, abs/2108.01547.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7098–7108. Association for Computational Linguistics.

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 sec. 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3.1

☑ B1. Did you cite the creators of artifacts you used?
3.1

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We follow 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)?
3.1

☒ 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 carefully cleaned.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3.1

☒ 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.
All of this is consistent with previous work.

C ☑ Did you run computational experiments?

3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3.1

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?
3.1,3.2

☒ 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 the model is computational and 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.)?
3.1,3.2

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
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 3. Some data samples will be in supplement materials.

☒ 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.