Knowledge-Grounded Dialogue Generation with a Unified Knowledge Representation

Abstract

Knowledge-grounded dialogue systems are challenging to build due to the lack of training data and heterogeneous knowledge sources. Existing systems perform poorly on unseen topics due to limited topics covered in the training data. In addition, it is challenging to generalize to the domains that require different types of knowledge sources. To address the above challenges, we present PLUG, a language model that homogenizes different knowledge sources to a unified knowledge representation for knowledge-grounded dialogue generation tasks. We first retrieve relevant information from heterogeneous knowledge sources (e.g., wiki, dictionary, or knowledge graph); Then the retrieved knowledge is transformed into text and concatenated with dialogue history to feed into the language model for generating responses. PLUG is pre-trained on a large-scale knowledge-grounded dialogue corpus. The empirical evaluation on two benchmarks shows that PLUG generalizes well across different knowledge-grounded dialogue tasks. It achieves comparable performance with state-of-the-art methods in the fully-supervised setting and significantly outperforms other approaches in zero-shot and few-shot settings.

1 Introduction

Recent work has shown that conversational models can be trained in an end-to-end fashion (Gao et al., 2019; Roller et al., 2020; Zhang et al., 2019; Adiwardana et al., 2020). Though such models can generate coherent and natural responses consistent with conversation history, there is still a clear gap between conversational AI agents and humans. The primary reason is that existing dialogue systems lack knowledge of the subject and thus cannot deep dive into specific topics with humans. In order to better incorporate knowledge into dialogue, knowledge-grounded dialogue systems have become increasingly popular.

Knowledge-grounded dialogue generation aims to generate informative and meaningful responses based on both conversation context and external knowledge sources. Thus far, researchers have collected knowledge-grounded dialogues for various tasks using crowdsourcing platforms, for instance, open-domain dialogues (Dinan et al., 2019; Zhou et al., 2018) and conversational recommendation dialogues (Li et al., 2018; Moon et al., 2019; Hayati et al., 2020). Workers are asked to base their replies on knowledge from structured knowledge bases (Moon et al., 2019; Tuan et al., 2019) or unstructured documents (Dinan et al., 2019; Zhou et al., 2018; Feng et al., 2020). Taking advantage of recent advances in large-scale language models (Raffel et al., 2019; Lewis et al., 2020a; Guu et al., 2020), researchers have also built knowledge-grounded dialogue systems by fine-tuning such language models in an end-to-end fashion (Shuster et al., 2021; Zhao et al., 2020b; Li et al., 2021).

However, there are two critical challenges in these existing methods. First, it is expensive and time-intensive to collect knowledge-grounded dialogues. As shown in Table 1, most of the datasets only cover a small portion of the knowledge base. Thus, systems which only fine-tune with small

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Knowledge</th>
<th>% Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open-domain</td>
<td>Wizard of Wikipedia</td>
<td>articles</td>
</tr>
<tr>
<td></td>
<td>CMU_DoG</td>
<td>articles</td>
</tr>
<tr>
<td>Recommendation</td>
<td>RE_DIAL</td>
<td>tables</td>
</tr>
<tr>
<td></td>
<td>OPEN_DIAL_KG</td>
<td>graph</td>
</tr>
</tbody>
</table>

Table 1: Knowledge representation and topic coverage statistics of existing knowledge-grounded dialogue datasets. % Topics means the portion of topics or facts in the knowledge database covered by the dataset.

\(^1\)Pre-trained Language model with a Unified knowledge representation for knowledge-grounded dialogues.
training sets generalize poorly on unseen topics in the same knowledge base. Additionally, the formats of knowledge sources vary in different tasks, making the approaches unable to transfer to other domains with different knowledge sources. For example, ReDIAL (Li et al., 2018) adopts a movie database as the knowledge source to recommend movies. Techniques on this task exploit the graph structure. It is not easy to adapt such techniques to document-grounded conversation tasks like Wizard of Wikipedia (Dinan et al., 2019).

In this work, we present PLUG, a model that can unify different knowledge formats for knowledge-grounded dialogue generation. First, we convert different knowledge formats (e.g., knowledge graph, knowledge base, and passages) to unstructured text, each using a different retriever. Then we use a pre-trained language model to process them into a unified representation to incorporate the knowledge into dialogue generation. We pre-train PLUG on different knowledge-ground dialogue corpora, including a large-scale open-domain conversation dataset from Reddit. This allows PLUG to learn knowledge in various formats from different tasks, and thus transfer to any knowledge-grounded dialogue task with few-shot learning techniques.

We evaluate the effectiveness of PLUG by applying it to an open-domain knowledge-grounded dialogue benchmark, Wizard of Wikipedia (Dinan et al., 2019), and a knowledge-grounded conversational recommendation benchmark, ReDIAL (Li et al., 2018). PLUG achieves results comparable to the state-of-the-art method under a fully-supervised setting. It outperforms other methods on both tasks under zero-shot and few-shot settings, demonstrating that PLUG can be grounded on world knowledge in different knowledge sources and generalize to different downstream tasks.

Our contributions are three-fold: (1) We propose a novel knowledge-based pre-trained language model, PLUG, that can be applied to any knowledge-grounded dialogue tasks; (2) Our model achieves slightly better results than state-of-the-art models in fully-supervised settings and shows promising improvements over the current state-of-the-art in zero-shot and few-shot settings; (3) We present extensive experiments to explore the bottlenecks of the task and the future direction of knowledge-grounded dialogues.

2 Approach

We describe our approach in this section. Figure 1 gives a diagram of our proposed method. We first introduce the background of knowledge-grounded dialogues and the backbone language model in Section 2.1. Then, we formalize the task and introduce the details of PLUG in Section 2.2. Finally, we explain the training process of our PLUG, which includes the pre-training dataset selection and the data preprocessing processes in Section 2.3.

2.1 Background: Knowledge-Grounded Pre-training

Traditional knowledge-grounded dialogue includes three steps: information extraction, knowledge prediction, and response generation. Previous work focuses on developing separate modules (Zhou et al., 2020b). Inspired by the recent success of applying a large-scale pre-trained language model on task-oriented dialogue systems (Peng et al., 2020; Hosseini-Asl et al., 2020), we explore the possibility of using a unified knowledge representation in a large-scale language model. In order to properly manage the task in a sequence-to-sequence setup, we choose T5 (Raffel et al., 2020) as our backbone.

T5 is a sequence-to-sequence pre-trained Transformer (Vaswani et al., 2017) model for transfer learning. It is trained by converting various language tasks into text-to-text tasks. After fine-tuning on a dialogue dataset, T5 can generate fluent and coherent responses. Nevertheless, responses are often too generic because they are not grounded on specific knowledge. PLUG is built on the T5 model but grounded on real-world knowledge during training, making it inherit T5’s capability of producing good responses but include more knowledge.

2.2 PLUG

We formulate a knowledge-grounded dialogue as:

\[ D = \{C, R, S\} \]  

where \( C = \{C_i\}_{i=1}^n \) is a dialogue context, and \( R = \{R_i\}_{i=1}^n \) is the response in a dialogue that has \( n \) turns. \( S \) is the external knowledge source for task \( t \). For each dialogue turn, we can formulate a knowledge-grounded dialogue generation task on a single domain \( d \) as \( p(R_t|C_t, S) \).

As shown in Figure 1, each task has its own knowledge source (e.g., documents, databases, and knowledge graphs). In order to make all knowledge-grounded dialogue generation tasks
able to fit in the text-to-text encoder-decoder framework, we follow T5 to feed each dialogue turn into the language model simply by concatenating the context \( C_i = \{c_1, c_2, ..., c_m\} \), and essential knowledge triples \( K_i = \{k_1, k_2, ..., k_n\} \) as a token sequence. The essential knowledge is extracted from the knowledge source \( S \) and represented as text of triples. We train the model to predict the responses token sequence \( R = \{r_1, r_2, ..., r_k\} \). The probability of the responses is formulated as:

\[
p(R_i | C_i) = \prod_{t=1}^{k} p(r_t | C_i, K_i, r_1, ..., r_{t-1})
\]

We will explain how we select and process pre-training datasets in the following sections.

### 2.3 Model training process

We pre-trained the PLUG model using two datasets, Reddit Conversation (Galley et al., 2018) and OpenDialKG (Moon et al., 2019). We will first present the three-step data cleaning process of Reddit Conversation in Section 2.3.1, then we will introduce OpenDialKG in Section 2.3.2.

#### 2.3.1 Reddit Conversation

Reddit Conversation Galley et al. (2018) is a large-scale open-domain conversation dataset. It extracts the conversation threads grounded on a document from the Reddit data. We only keep the conversations grounded on Wikipedia passages for pre-training to recognize better the knowledge used in the dialogue. Since vanilla document-based dialogue in the dataset does not have a knowledge label for each dialogue turn, we apply a hierarchical information extraction method to obtain the essential knowledge in each turn. Our information extraction method includes three steps: knowledge retrieval, statistical ranking, and semantic ranking.

**Knowledge Retriever.** We use a knowledge retriever to retrieve all relevant knowledge in a single turn’s response. We first extract the title of the grounding Wikipedia passage in the dialogue. Then, we retrieve knowledge triples from a large-scale knowledge graph, DBpedia (Lehmann et al., 2015). Specifically, we query the DBpedia via a public SPARQL endpoint and then collect triples whose subject or object is in the Wikipedia passage in the dialogue. For example, we keep triples \(<\text{Barack Obama}, \text{alma mater}, \text{Columbia University}>\) and \(<\text{Michelle Obama}, \text{spouse}, \text{Barack Obama}>\) for the dialogue about Barack Obama. To carry sufficient knowledge to refine in the next step, we retrieve 500 triples for every passage.

**Statistical Ranking.** After retrieving adequate knowledge, we rank the corresponding triples to refine the knowledge. Specifically, we get the TF-IDF (term frequency-inverse document frequency) value for all the retrieved triples. To find the triples related to the context, we concatenate the dialogue history and the response as the query. Then we compute the cosine similarity between the query and every triple. Because every triple has the Wikipedia passage name as the subject or object, a higher cosine similarity score means the query has more similar text with the distinguished text in the triple.

The Reddit data dumps: https://files.pushshift.io/reddit/

---

3https://dbpedia.org/sparql
We rank the query-document similarity score and only keep the top-50 triples in this step.

**Semantic Ranking.** The TF-IDF-based cosine similarity score only counts words overlapping between triples and the query. It will introduce triples whose overlapping words are not meaningful in the context and response. Additionally, the Reddit Conversation dataset is obtained from Reddit conversation threads. It involves many responses that are not grounded on any knowledge. In order to find the triples that have the best semantic similarity with the response and filter out ungrounded responses, in this step, we estimate the semantic similarity score with Sentence-Bert (Reimers and Gurevych, 2019). We rerank the 50 triples from the second step based on the score. Additionally, we abandon the dialogue turns whose best semantic similarity is lower than a threshold because the response cannot find proper knowledge, while we want to pre-train the model with knowledge-grounded turns.

### 2.3.2 OpenDialKG

To generalize our model to various tasks, we also employ OpenDialKG to enrich our pre-training dataset. OpenDialKG consists of two types of tasks, recommendations and chit-chat, across four domains. Unlike the Reddit Conversation dataset, which needs to find the knowledge grounding in every turn, the original OpenDialKG has a Knowledge graph path label for each dialogue, and a triple label for each dialogue turn. The response is grounded on the labeled triple during data collection. Thus, we use the triple in the dataset as the essential knowledge in our pre-training examples.

### 3 Experiments

We demonstrate our approach on two different downstream tasks: open-domain knowledge-grounded dialogue and conversational recommendation. Besides the fully-supervised learning setting, we also explore the performance of our approach in few-shot and zero-shot settings. We describe our implementation details in Section A in Appendix.

#### 3.1 Datasets and Knowledge Sources

We test our approach on Wizard of Wikipedia (WoW; (Dinan et al., 2019)) and REDIAL (Li et al., 2018). Basic dataset statistics are listed in Table 2.

**Wizard of Wikipedia.** This dataset (Dinan et al., 2019) is collected on Amazon Mechanical Turk. Each conversation happens between a “wizard” who has access to knowledge about a specific topic, and an “apprentice” who is interested in the topic. The wizard’s response is grounded on a Wikipedia article in each turn. The data is split as a training set, a validation set, and a test set. The test set has two subsets: Test Seen and Test Unseen. Test Seen contains conversations whose topics are seen in the training set, while topics in Test Unseen are not seen in the training or validation set. To extract the essential knowledge in each dialogue turn, we first keep the top five passages retrieved by the TF-IDF retriever in Shuster et al. (2021). Then we use an Open Information Extraction (OpenIE) annotator⁴ to extract the top three triples from the passages as our essential knowledge. The pre-processing is conducted with the code published on ParlAI.⁵

**REDIAL.** REDIAL (Li et al., 2018) is also collected on Amazon Mechanical Turk. Two crowd-workers, a “movie seeker” and “movie recommender,” are randomly paired. The recommender has access to a movie database and can recommend movies based on movie information, such as actors and movie genres. There are 6,924 different movies mentioned in 51,699 movie slots in the dataset. We follow Li et al. (2018) to split the dataset into training, validation, and test sets. Since recommenders use movie-related knowledge when they recommend movies to seekers, we use it as the essential knowledge for a given turn in this dataset. We experiment with three knowledge sources: (1) We query the movie names mentioned in the dialogue context and retrieve similar movies from the knowledge graph DBpedia, mentioned in Section 2.3, and then input the similar movies in a triple format as the essential knowledge. (2) We query the movie names mentioned in the context and retrieve movie comments from MovieLens,⁶ then use the keywords in the comments as the essential knowledge. (3) We use the output of the recommender

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Valid</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>WoW</td>
<td>18,430</td>
<td>Seen</td>
<td>981</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unseen</td>
<td>967</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>968</td>
</tr>
<tr>
<td>REDIAL</td>
<td>8,004</td>
<td>1,001</td>
<td>1,001</td>
</tr>
</tbody>
</table>

Table 2: Number of conversations in Wizard of Wikipedia (WoW) and REDIAL.

---

⁵https://github.com/facebookresearch/ParlAI
⁶https://grouplens.org/datasets/movielens/
We first evaluate PLUG with all training examples in the training sets to compare its performance with other state-of-the-art systems. Additionally, we experiment with using golden knowledge in the input to explore the upper bound of our method.

Table 3 shows the Wizard of Wikipedia Test Seen and Test Unseen results. We can see that PLUG with retrieved knowledge achieves better BLEU-4, ROUGE-L, and F1 scores than the RAG method and the model without adding knowledge in the input, on both seen and unseen topics. It suggests that our essential knowledge format helps the model generate responses to ground knowledge better. We also observe that PLUG outperforms the model without pre-training on all metrics, which means our pre-training can boost this task.

We list ReDIAL’s results in Table 4. We compare our approach to the state-of-the-art systems and T5-Large models without pre-training. Additionally, we include a comparison to models with different knowledge sources as described in Section 3.1. It shows that our best model (PLUG+KGSF) achieves the new state-of-the-art results on the recommendation item recall metric and distinct metrics. This result is understandable given that our approach is built upon pre-trained language models. Similarly, we also observe noticeable performance gains for the pre-training on this task. However, compared to systems with currently available knowledge sources, it is immediately apparent that the system with golden knowledge outperforms the current state-of-the-art on all metrics by a large margin. This huge gap implies that current knowledge retrievers are the main bottleneck for the conversational recommendation task. We will discuss more details in Section 3.7.

Overall, we observe noticeable improvement brought by pre-training on both tasks, but it is less significant than expected. It implies that the knowledge grounding pattern in the response is limited; a complete training set is more than enough for the T5-Large model to learn the generation task. We will discuss more details in zero-shot and few-shot settings in the following subsections.

### 3.5 Zero-Shot and Few-Shot Results

We focus on zero-shot and few-shot settings because it is more realistic to evaluate dialogue systems. Specifically, we randomly sample 10/50/500 dialogues with different topics from the training sets and observe performance on the complete test sets. We also evaluate under a zero-shot setting. We experiment with knowledge retrieved by exist-
Table 3: Fully-supervised results on Wizard of Wikipedia Test Seen and Test Unseen Sets.

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU4</th>
<th>ROUGE-L</th>
<th>F1</th>
<th>KF1</th>
<th>BLEU4</th>
<th>ROUGE-L</th>
<th>F1</th>
<th>KF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAG-T5-Large (Shuster et al., 2021)</td>
<td>3.8</td>
<td>22.1</td>
<td>21.9</td>
<td>25.9</td>
<td>2.8</td>
<td>20.4</td>
<td>21.4</td>
<td>13.9</td>
</tr>
<tr>
<td>T5-Large-w/o Knowledge</td>
<td>4.1</td>
<td>18.0</td>
<td>18.3</td>
<td>19.2</td>
<td>2.1</td>
<td>15.4</td>
<td>21.4</td>
<td>13.9</td>
</tr>
<tr>
<td>T5-Large-Retrieved Knowledge</td>
<td>5.8</td>
<td>21.8</td>
<td>25.8</td>
<td>22.6</td>
<td>3.4</td>
<td>19.2</td>
<td>22.7</td>
<td>17.6</td>
</tr>
<tr>
<td>T5-Large-Golden Knowledge</td>
<td>11.3</td>
<td>30.8</td>
<td>35.6</td>
<td>46.8</td>
<td>8.7</td>
<td>28.4</td>
<td>33.0</td>
<td>43.6</td>
</tr>
<tr>
<td>PLUG-Retrieved Knowledge</td>
<td>6.0</td>
<td>22.3</td>
<td>26.5</td>
<td>22.4</td>
<td>3.5</td>
<td>19.5</td>
<td>23.3</td>
<td>18.6</td>
</tr>
<tr>
<td>PLUG-Golden Knowledge</td>
<td>11.5</td>
<td>31.1</td>
<td>36.0</td>
<td>47.8</td>
<td>8.8</td>
<td>29.0</td>
<td>33.4</td>
<td>46.0</td>
</tr>
</tbody>
</table>

Table 4: Fully-supervised results on REDIAL.

<table>
<thead>
<tr>
<th>Model</th>
<th>B4</th>
<th>RL</th>
<th>DIST2</th>
<th>DIST4</th>
<th>Rec</th>
</tr>
</thead>
<tbody>
<tr>
<td>KBRD</td>
<td>1.8</td>
<td>16.5</td>
<td>0.48</td>
<td>0.67</td>
<td>0.7</td>
</tr>
<tr>
<td>KGDF</td>
<td>2.3</td>
<td>13.1</td>
<td>0.49</td>
<td>1.28</td>
<td>0.9</td>
</tr>
<tr>
<td><strong>T5-Large</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o KG</td>
<td>3.7</td>
<td>18.3</td>
<td>0.72</td>
<td>1.10</td>
<td>3.4</td>
</tr>
<tr>
<td>Golden</td>
<td>10.4</td>
<td>32.7</td>
<td>1.17</td>
<td>1.60</td>
<td>83.5</td>
</tr>
<tr>
<td>KGDF</td>
<td>3.7</td>
<td>17.4</td>
<td>1.13</td>
<td>2.02</td>
<td>4.7</td>
</tr>
<tr>
<td><strong>PLUG</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o KG</td>
<td>3.9</td>
<td>19.6</td>
<td>0.78</td>
<td>1.31</td>
<td>3.7</td>
</tr>
<tr>
<td>Golden</td>
<td>10.6</td>
<td>33.5</td>
<td>1.26</td>
<td>1.81</td>
<td>84.3</td>
</tr>
<tr>
<td>DBpedia</td>
<td>3.3</td>
<td>18.3</td>
<td>0.45</td>
<td>0.66</td>
<td>0.8</td>
</tr>
<tr>
<td>MovieLens</td>
<td>3.4</td>
<td>17.8</td>
<td>0.91</td>
<td>1.34</td>
<td>2.4</td>
</tr>
<tr>
<td>KGDF</td>
<td>3.8</td>
<td>18.0</td>
<td>1.51</td>
<td>2.84</td>
<td>5.3</td>
</tr>
</tbody>
</table>

Figure 2: Zero-shot and few-shot results on Wizard of Wikipedia Test Seen and Test Unseen sets.

Figure 3: Fully-supervised results on REDIAL.

3.6 Human Evaluation

We conduct a human evaluation on Wizard of Wikipedia to assess the overall quality of the responses of our model compared to T5 and RAG. Specifically, we randomly select 100 responses for each model with the same context from Test Seen and Test Unseen. For the few-shot setting, we use the models trained with 50 dialogues. We hire workers on Amazon Mechanical Turk to rate models. We use the published FiD RAG DPR model at https://parl.ai/projects/hallucination/
els’ responses on a 0 - 2 scale with three metrics: Fluency, Coherence, and Knowledge. The order of the systems shown to workers is shuffled to avoid confounding practice effects. Three different workers evaluate each dialogue turn. Table 5 reports average metrics scores. We observe that responses from our fully-supervised model are more fluent and coherent than those from RAG, which benefits from our simple but effective essential knowledge representation. We can also see significant improvement on all metrics for PLUG under a zero-shot setting compared to the T5 model. Performance improvement under the few-shot setting is less than in the zero-shot setting, but PLUG still outperforms T5 on all metrics, which aligns with the result in automatic evaluation. Interestingly, we observe that responses from the model trained with 50 dialogues have already been very fluent and coherent, which is even higher than those from the fully-supervised model. However, responses from the fully-supervised model contain the most appropriate knowledge, which suggests that the model has learned how to generate high-quality responses in a few-shot setting, but it continues to learn how to ground on knowledge with more training samples.

<table>
<thead>
<tr>
<th>Model</th>
<th>Fluency</th>
<th>Coherence</th>
<th>Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAG</td>
<td>1.06</td>
<td>1.08</td>
<td>1.19</td>
</tr>
<tr>
<td>T5-Large</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Zero-shot</td>
<td>0.87</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>- Few-shot</td>
<td>1.26</td>
<td>1.35</td>
<td>1.31</td>
</tr>
<tr>
<td>PLUG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Zero-shot</td>
<td>1.20</td>
<td>1.34</td>
<td>1.25</td>
</tr>
<tr>
<td>- Few-shot</td>
<td>1.29</td>
<td>1.42</td>
<td>1.39</td>
</tr>
<tr>
<td>- Fully-supervised</td>
<td>1.24</td>
<td>1.37</td>
<td>1.46</td>
</tr>
</tbody>
</table>

Table 5: Human evaluation results of different models on Wizard of Wikipedia.

3.7 Discussion and Analysis

To investigate the enormous performance gap between models with golden knowledge and retrieved knowledge in Table 4, we compare the performance of models with different knowledge sources on the ReDIAL dataset. Specifically, we mix the golden movies information and the retrieved movie information retrieved in the training/validation/test set to simulate knowledge sources with different recall performances. We experiment with 0/20/40/60/80/100 percent of the golden knowledge. 0 means all training samples includes retrieved knowledge (a flawed knowledge source), 100 means all training samples include golden knowledge (a perfect knowledge source). To have a more realistic setting, we compare the performance of PLUG and T5 under the few-shot setting (trained on 50 dialogues), as shown in Figure 4.

We find that the performance gain for both models is linear with respect to the performance of the knowledge source, whereas PLUG has a better boost on the BLEU-4 score and recommendation recall score. The curve with a higher slope shows the potential benefit from our pre-training method when better knowledge sources are available in the future. Furthermore, the gap on DIST-4 between PLUG and T5 is almost the same as golden knowledge increases, but the DIST-4 of T5 surprisingly drops when no golden knowledge is available. It means that T5 requires a better knowledge source in the training set to generate diverse responses under a few-shot setting, while PLUG has learned that ability in the pre-training process and generates diverse responses out-of-the-box. We also note that the performance boost with a better knowledge source is much more than the generation technology in previous work. This massive gap may shed light on the research direction of knowledge-grounded dialogue tasks for future efforts.

4 Related Work

Knowledge-grounded dialogue is becoming an increasingly important topic, with datasets proposed to model its occurrence on different tasks. Dialogues in these datasets are based on various formats of knowledge, such as documents in open-domain conversations (Ghazvininejad et al., 2018; Dinan et al., 2019; Gopalakrishnan et al., 2019), movie database in movie recommendation con-
versations (Li et al., 2018; Hayati et al., 2020), or knowledge graph in recommendation conversations (Moon et al., 2019; Liu et al., 2021b).

One of the principal challenges in knowledge-grounded conversations is incorporating knowledge into dialogue systems. Recent work investigates different techniques of learning a better knowledge representation to fuse knowledge in the response generation process. Ghazvininejad et al. (2018) separately encoded the dialogue history and documents to infuse the response with external world facts. Chen et al. (2019); Wang et al. (2021); Zhou et al. (2020a) joined a knowledge graph representation in a response generation module. Zhu et al. (2017) combined the knowledge from the database with the user intent and fed it into the decoder. Unlike these studies, we use a single encoder for both dialogue context and knowledge.

In order to improve the systems’ performance on unseen topics and train knowledge-grounded dialogue in a low-resource setting, researchers investigate pre-training methods for the knowledge-grounded tasks. Zhao et al. (2020a) pre-trained the dialogue generation model with ungrounded dialogues and the knowledge encoder with the Wikipedia dump separately. Li et al. (2020) proposed a pre-trained latent variable model to learn the way that the knowledge is expressed in the response. Liu et al. (2021a) built a document encoder and a dialogue context encoder, then pre-trained them separately in multiple stages. The knowledge encoder in these studies is pre-trained separately and only accepts the same knowledge format, while we pre-train our model with essential knowledge in the text format, thus fitting different knowledge sources in the downstream tasks.

Inspired by the success of pre-trained language models for a variety of natural language processing tasks (Devlin et al., 2019; Radford et al., 2019; Yang et al., 2019; Ma et al., 2021), another line of work investigates learning knowledge through language models’ parameters (Petroni et al., 2019; Rosset et al., 2020; Roberts et al., 2020). In our pre-training process, we aim to learn extra knowledge and, more importantly, learn how to generate response grounding on the essential knowledge.

Two recent studies are most closely related to our work. Chen et al. (2020) proposed a pre-trained model for data to text tasks. They unified the knowledge format in the pre-training data and downstream tasks, however only depend on the graph structure and do not work on knowledge-grounded dialogues. Shuster et al. (2021) applied the document retrieval augmentation method (Lewis et al., 2020b) on open-domain knowledge-grounded dialogues. However, they do not do pre-training and rely on Wikipedia documents in the decoder, limiting their model to document-based dialogues. We use unified essential knowledge instead of documents in our pre-training, making our model more generalizable. Our approach can be seen as generalizing both lines of work, and showing for the first time that a pre-trained model is effective for various knowledge-grounded tasks with different knowledge formats.

5 Conclusion and Future Work

We present a knowledge-grounded pre-trained language model PLUG that can be applied to various knowledge-grounded dialogue tasks. It subsumes different knowledge sources into a simple but effective unified essential knowledge representation. Evaluation results on two benchmarks indicate that our model performs better in zero-shot and few-shot settings and can generalize to different knowledge-grounded tasks.

As future work, we would like to augment our pre-training datasets with more knowledge sources, and apply our method to other knowledge-grounded tasks such as question answering. Another interesting direction would be to develop better information retrievers since experiments show that the retriever is the main bottleneck in the knowledge-grounded dialogues.
References


A Implementation Details

We process the Reddit monthly submissions and comments dump from 2011 to 2017, consisting of over 894k knowledge-grounded dialogue turns. As detailed in Section 2.3.1, we set the threshold as 0.35 in the semantic ranking. After filtering with our hierarchical information extraction method, over 321k dialogue turns remain. All dialogue turns in the Open DialKG dataset are used in the pre-training. Each dialogue turn is processed to form a sequence of tokens consisting of three segments: dialogue context, essential knowledge, and response. We keep the top-three triples/keywords as our essential knowledge in pre-training and downstream tasks. PLUG is implemented with Huggingface Pytorch Transformers\textsuperscript{8} (Wolf et al., 2020) and initialized with the 800M-parameter T5 model. We use Adam (Kingma and Ba, 2014) with weight decay for pre-training. Training examples are truncated to ensure a maximal length of 512. Models are pre-trained on 8 Nvidia V100 GPUs until we observe no progress on validation data or up to 20 epochs. The best configuration of hyper-parameters is selected through cross-validated grid-search.

B Ethical Considerations

It is essential to consider potential ethical issues in knowledge-grounded dialogue systems. In our work, PLUG is pre-trained on a large-scale dataset Reddit Conversation, which is crawled from the internet. We follow Galley et al. (2018) to filter out dialogues that have profanity content. However, it is still possible to include inappropriate content in the pre-training dataset. In processing the Reddit Conversation dataset during pre-training, we have carefully designed rules to remove knowledge that has profanity words. Additionally, the T5 model may have seen inappropriate content in its pre-training tasks, and it may generate wrong responses even if we input appropriate knowledge. Considerable additional work is needed to detect profanity content when we generate with a pre-trained language model. In addition to these ethical considerations, we have sought to better conduct our human evaluation by transparently communicating with crowd-workers about data use and study intent and compensating workers at a reasonable hourly wage.

\textsuperscript{8}https://github.com/huggingface/transformers is licensed under the Apache License 2.0
C  Human Evaluation Interface

Figure 5 shows the interface of an example in our human evaluation.
A user and an agent are chatting with each other. Now it is the agent’s turn to reply. Please read the dialog context and rate responses from six agents.

**Dialog context:**

User: What is your favorite thing to post on Instagram?
Agent: Usually I just photos or videos. Do you know who owns Instagram?
User: No I don’t, please tell me.
Agent: It's actually owned by Facebook
User: That cool, what kind of picture do you like to post.
Agent: Generally landscape photos. Instagram allows you to use different filters which makes it pretty cool.
User: I used the black and white filter for my dog pictures.

**Response 1**

Agent: That's a cool filter. Do you have a favorite photo of your dog?

**Questions**

1. Is the response fluent?

   - 0 - Not fluent
   - 1 - Neutral
   - 2 - Very fluent

---

Figure 5: Screenshot of human evaluation interface.