Knowledge-Grounded Dialogue Generation with Pre-trained Language Models

Xueliang Zhao[1,2], **Wei Wu**[3], **Can Xu**[4], **Chongyang Tao**[4], **Dongyan Zhao**[1,2], **Rui Yan**[1,2,5*]
[1]Wangxuan Institute of Computer Technology, Peking University, Beijing, China
[2]Center for Data Science, AAIS, Peking University, Beijing, China
[3]Meituan, Beijing, China [4]Microsoft Corporation, Beijing, China
[5]Beijing Academy of Artificial Intelligence (BAAI), Beijing, China
{xl.zhao,zhaody,ruiyan}@pku.edu.cn
{wuwei19850318,chongyangtao}@gmail.com

Abstract

We study knowledge-grounded dialogue generation with pre-trained language models. To leverage the redundant external knowledge under capacity constraint, we propose equipping response generation defined by a pre-trained language model with a knowledge selection module, and an unsupervised approach to jointly optimizing knowledge selection and response generation with unlabeled dialogues. Empirical results on two benchmarks indicate that our model can significantly outperform state-of-the-art methods in both automatic evaluation and human judgment.

1 Introduction

With advances in neural machine learning (Sutskever et al., 2014; Gehring et al., 2017; Vaswani et al., 2017) and availability of the huge amount of human conversations on social media (Adiwardana et al., 2020), building an open domain dialogue system with data-driven approaches has attracted increasing attention from the community of artificial intelligence and natural language processing. In this work, we are interested in generative approaches. Generative models for open domain dialogues are notorious for replying with generic and bland responses, resulting in meaningless and boring conversations (Li et al., 2015). Such deficiency is particularly severe when human participants attempt to dive into specific topics in conversation (Dinan et al., 2019). As a result, there is still a big gap between conversation with existing systems and conversation with humans.

Very recently, there emerge two lines of research that seem promising to bridge the gap. One is to apply large-scale pre-trained language models, such as GPT-2 (Radford et al., 2019), to the task of open domain dialogue generation. Prototypes such as DialoGPT (Zhang et al., 2019c) have exhibited compelling performance on generating responses that make sense under conversation contexts and at the same time carry specific content for keeping the conversation going. While the giant language models can memorize enough patterns in language during pre-training, they only capture "average" semantics of the data (Zhang et al., 2019c). As a result, responses could still be bland or inappropriate when specific knowledge is required, as illustrated by the example in Table 1. The other line is to ground dialogue generation by extra knowledge such as unstructured documents (Zhao et al., 2020). By the means, the documents (e.g., wiki articles) serve as content sources, and make a dialogue system knowledgeable regarding to a variety of concepts in discussion. However, collecting enough dialogues that are naturally grounded on documents for model training is not trivial. Although some benchmarks built upon crowd-sourcing have been released by recent papers (Zhou et al., 2018b; Dinan et al., 2019; Gopalakrishnan et al., 2019), the small training size

Context	
A	I just discovered star trek and I really like watching star trek .
B	Gene Roddenberry created it based upon science fiction and it is American media.
	...
A	If I remember Captain Kirk was not the original captain .
B	The Star Trek Canon of the series an animated had 5 spin offs.
A	I watched a little of the next generation but could not get into it like i did with the original show .
Response	
Human	These adventures went on but were short lived and six feature films.
DialoGPT	I think it's worth it.

Table 1: An example from the test set (Test Seen) of Wizard of Wikipedia (Dinan et al., 2019) .

*Corresponding author: Rui Yan (ruiyan@pku.edu.cn).

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 3377–3390,
November 16–20, 2020.

makes the generation models generalize badly on unseen topics (Dinan et al., 2019) and the cost of building such data also prevents from transferring the techniques proved on the benchmarks to new domains and new languages.

Encouraged by the results on pre-training for dialogue generation and knowledge-grounded dialogue generation, and motivated by the problems in both sides, we consider bringing the two together in this work. Specifically, we propose knowledge-grounded dialogue generation with pre-trained language models in order to endow a generative model with both rich knowledge and good generalization ability[1]. The challenge is that pre-trained language models often set constraints on the maximum number of tokens they can handle (e.g., the maximum number for GPT-2 (Radford et al., 2019) is 1024), and thus hinders exploitation of the knowledge text which could be rather long and redundant (e.g., in Wizard of Wikipedia (Dinan et al., 2019), on average each conversation context is associated with 61.2 sentences retrieved from wiki articles, and the average number of tokens in the extra knowledge is 1625.6). Indeed, the conflict between model capacity and the ability required for processing long knowledge input represents an essential obstacle for applying pre-trained language models to knowledge-grounded dialogue generation, since on the one hand we always have to set up an upper bound to the capacity of pre-trained models in order to handle massive text corpus, and on the other hand we need to keep sufficient candidates with rich enough content in the procedure of response generation in order to guarantee the recall of relevant knowledge.

To overcome the challenge, we consider equipping the pre-trained response generation model with a knowledge selection module whereby the redundant knowledge input is slimmed with relevant information (regarding to conversation contexts) kept to meet the capacity constraint. While some recent papers on knowledge-grounded dialogues have paid attention to the problem of knowledge selection (Lian et al., 2019; Kim et al., 2020; Ren et al., 2019), the knowledge selection module is either deeply coupled with the specially configured models (Lian et al., 2019; Ren et al., 2019) and thus is incompatible with the pre-trained language models, or it is learned with human annotations (Dinan et al., 2019; Kim et al., 2018) which are difficult to obtain in practice (e.g., the dataset in (Zhou et al., 2018b) does not contain annotations for knowledge selection). Therefore, we propose an unsupervised approach where learning of knowledge selection and fine-tuning of response generation are jointly conducted with unlabeled dialogues. Specifically, we build the knowledge selection module on the basis of BERT, and formalize knowledge selection as a sequence prediction process, by which the model can take advantage of the pre-training techniques and dynamically determine the relevant knowledge for a given context. The learning algorithm starts from training with pseudo ground-truth that is constructed by making full use of responses as an alternation of human annotations, and then alternatively updates the knowledge selection model and the response generation model through a reinforcement learning approach and a curriculum learning approach respectively. Thus, knowledge selection is further optimized with the feedback from response generation, and the knowledge used for fine-tuning the response generation model gradually moves from the pseudo ground-truth to the prediction of the knowledge selection module.

We test the proposed method on two benchmarks of knowledge-grounded dialogue generation: Wizard of Wikipedia (Dinan et al., 2019) and CMU Document Grounded Conversations (Zhou et al., 2018b). Evaluation results indicate that our model can significantly outperform state-of-the-art methods as well as a few pre-trained models used in heuristic ways, and thus achieves new state-of-the-art on the benchmarks. Moreover, as a byproduct, the knowledge selection module also outperforms the state-of-the-art model in terms of accuracy of knowledge selection on Wizard of Wikipedia, implying that other models could also benefit from the component.

Our contributions in this paper are three-fold: (1) proposal of a knowledge selection module for applying pre-trained language models to the task of knowledge-grounded dialogue generation; (2) proposal of an unsupervised approach in which learning of knowledge selection and fine-tuning of the pre-trained model are conducted in a joint manner; and (3) empirical verification of the effectiveness of the proposed method on benchmarks of knowledge-grounded dialogue generation.

[1]In this paper, we assume that knowledge is retrieved from documents.

2 Related Work

Early work on end-to-end open domain dialogue generation is inspired by the research of machine translation (Ritter et al., 2011; Shang et al., 2015; Vinyals and Le, 2015). Later, the vanilla encoder-decoder architecture is widely extended to improve diversity of responses (Li et al., 2015; Xing et al., 2017a; Zhao et al., 2017; Tao et al., 2018); to model the structure of conversation contexts (Serban et al., 2016, 2017; Xing et al., 2017b; Zhang et al., 2019a); to control attributes of responses (Xu et al., 2019; Zhou et al., 2017; Zhang et al., 2018a; Wang et al., 2018; See et al., 2019); and to bias responses to some specific personas (Li et al., 2016; Zhang et al., 2018b). Recently, grounding dialogue generation by extra knowledge is emerging as an important step towards human-like conversational AI, where the knowledge could be obtained from knowledge graphs (Zhou et al., 2018a; Moon et al., 2019; Tuan et al., 2019), retrieved from unstructured documents (Dinan et al., 2019; Lian et al., 2019; Zhao et al., 2020; Kim et al., 2020), or extracted from visual background (Mostafazadeh et al., 2017; Shuster et al., 2018; Huber et al., 2018). In this work, we study document-grounded dialogue generation. Rather than learning from scratch like most existing work, we take advantage of the pre-trained language models and achieve new state-of-the-art on the benchmarks of the task.

Big, deep neural language models pre-trained on huge unlabeled text corpus have led to strong improvements on numerous natural language understanding and natural language generation benchmarks (Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019; Radford et al., 2019; Song et al., 2019; Dong et al., 2019; Lewis et al., 2019), and therefore are revolutionizing almost the full spectrum of NLP applications (Raffel et al., 2019; Sun et al., 2019b; Qiao et al., 2019; Zhang et al., 2019b; Lample and Conneau, 2019) and some interdisciplinary applications in NLP and computer vision (Lu et al., 2019; Su et al., 2019; Sun et al., 2019a). In the context of dialogue generation, by fine-tuning GPT-2 (Radford et al., 2019) in different sizes on social media data, recent work has (Zhang et al., 2019c; Wolf et al., 2019) shown promising progress on conversation engagement and commonsense question-answering. In this work, we further explore the application of pre-training to the task of open domain dialogue generation by equipping the pre-trained language models with external knowledge. Different from a very recent paper on pre-training for low-resource knowledge-grounded dialogue generation (Zhao et al., 2020), the work presents an in-depth investigation on how to release the power of the existing pre-trained language models on the task when input exceeds the capacity of the models.

3 Preliminary

3.1 Problem Formalization

Suppose that we have a dataset $\mathcal{D} = \{(U_i, D_i, r_i)\}_{i=1}^{N}$, where $\forall i \in \{1, \ldots, N\}$, U_i is a dialogue context, D_i is a document that contains relevant knowledge regarding to U_i, and r_i is a response to U_i based on D_i. The goal is to learn a generation model $P(r|U, D; \theta)$ (θ denotes the parameters of the model) from $\mathcal{D}$, and thus given a new dialogue context U associated with a document D, one can generate a response r following $P(r|U, D; \theta)$.

3.2 Pre-trained Language Models

We define $P(r|U, D; \theta)$ on the basis of GPT-2 from OpenAI (Radford et al., 2019). GPT-2 are transformer language models with a stack of masked multi-head self-attention layers, and are learned from large scale web text. To apply GPT-2 to the task of knowledge-grounded dialogue generation, we formulate the generation problem as

$$\begin{aligned} P(r|U, D; \theta) &= P(r|g(U, D); \theta) \\ &= \prod_{t=1}^{l_r} P(r_t|g(U, D), r_{1:t-1}; \theta), \end{aligned} \tag{1}$$

where $g(U, D)$ tailors $U \cup D$ to meet the length constraint of a GPT-2 model as the input of generation, and r_t refers to the t-th token of r whose length is supposed to be l_r. The problem then boils down to (1) how to define $g(U, D)$; and (2) how to fine-tune θ (and probably learn $g(U, D)$) with $\mathcal{D}$.

In this work, we assume that labels that indicate the ground-truth knowledge are not available, which is practical but makes the problem even more challenging. Since D could be rather redundant with a lot of information irrelevant with the topic or the context of the conversation, simply truncating the concatenation of sentences of U and D as $g(U, D)$ may cut the relevant knowledge and introduce noise into response generation, which hurts the performance of the GPT-2 model, as will be demonstrated in the experiments. Therefore, we consider learning a $g(U, D)$ that can distill useful

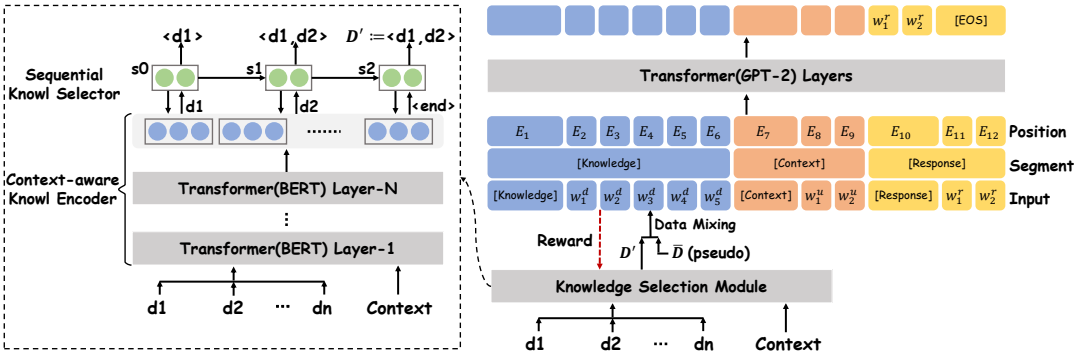

Figure 1: Architecture of the proposed model.

information from D for the GPT-2 model, as will be elaborated in the next section.

4 Approach

Heading for learning a $g(U, D)$ for applying GPT-2 to the task of knowledge-grounded dialogue generation, we need to deal with several challenges: (1) how to model the correlation between a context and the external knowledge; (2) how to learn $g(U, D)$ when labels of ground-truth knowledge are absent; and (3) how to jointly optimize $g(U, D)$ and the GPT-2 model with $\mathcal{D}$, and thus the two can boost each other. Figure 1 illustrates the architecture of the model. On the basis of the transformer architecture, the knowledge selection module is made up of a context-aware knowledge encoder and a sequential knowledge selector. The former captures interaction patterns between a context U and each sentence in D through a stack of self-attention layers, and the patterns are then fed to the latter to decode useful knowledge one sentence per step. Since human annotations are not accessible, the learning method begins with pseudo ground-truth constructed by making full use of responses, and optimization of $g(U, D)$ and optimization of the GPT-2 generation model are alternatively conducted with a reinforcement learning approach and a curriculum learning approach respectively.

4.1 Context-Aware Knowledge Encoder

We choose BERT (Devlin et al., 2018) as the backbone of the encoder. Thus, the encoder can take advantage of pre-training, and the multi-layer bi-directional attention mechanism in BERT allows a dialogue context and the associated knowledge to sufficiently interact with each other, resulting in context-aware knowledge representations. Specifically, let $U = (u_1, \ldots, u_n)$ and $D = (d_1, \ldots, d_m)$ be the context and the knowledge respectively, then we concatenate $\{u_i\}_{i=1}^{n}$ as $(w_1^u, \cdots, w_{l_u}^u)$ with w_i^u the i-th word and l_u the length of the sequence, and define the input of the encoder as $\mathcal{S} = (S_1, \ldots, S_m)$ with S_i formulated as

$$S_i = [\text{CLS}]w_1^u \ldots w_{l_u}^u[\text{SEP}]w_{i,1}^d \ldots w_{i,j}^d \ldots w_{i,l_d}^d[\text{SEP}], \quad (2)$$

where $w_{i,j}^d$ refers to the j-th word of $d_i \in D$, and l_d is the length of d_i. Each $S_i \in \mathcal{S}$ passes through the stacked self-attention layers, and is finally represented as $e_i = \text{CLS}(\text{BERT}(S_i))$ where $\text{BERT}(S_i)$ refers to the sequence of vectors from the last layer of the encoder and $\text{CLS}(\cdot)$ is a function that returns the first vector of the sequence (i.e., the vector corresponding to the [CLS] token). The output of the encoder is given by $E = (e_1, \ldots, e_m)$.

4.2 Sequential Knowledge Selector

With E as input, the sequential knowledge selector determines a subset of D (denoted as D') as the relevant knowledge and exploits D' to construct $g(U, D)$. Since there may exist one-to-many relations between a context and the relevant knowledge (Kim et al., 2020), the size of D' could vary from context to context. Therefore, we regard the construction of D' as a sequence prediction process in which D' starts from an empty set and gradually expands by adding one sentence from D per step. By this means, the size of D' can also be viewed as a parameter and is dynamically determined according to the given context. Formally, we maintain a sequence of hidden states $\{s_t\}_{t=0}^{T_{U,D}}$ with the initial state s_0 a trainable parameter, and weight $\{d_i\}_{i=1}^{m}$ by an attention mechanism which

can be formulated as

$$P(d_i|U, d_{j_{1:t-1}}) = \exp(\alpha_{t,i}) / \sum_i \exp(\alpha_{t,i})$$
$$\alpha_{t,i} = v^\top \tanh(W_e e_i + W_s s_t + b), \tag{3}$$

where W_e, W_s, b and v are trainable parameters. Then d_{j_t} will be added to D' if $j_t = \text{argmax}_{i \in \{1,\dots,m\}} P(d_i|U, d_{j_{1:t-1}})$. After that, s_{t+1} is calculated by

$$s_{t+1} = \text{LSTM}(e_{j_t}, s_t) \tag{4}$$

To determine $T_{U,D}$, we introduce a special embedding e_{spe} into E, and terminate the prediction process if e_{spe} is selected or an upper bound T_{max} is reached. Finally, $g(U, D)$ is defined as the concatenation of the sentences in $U \cup D'$.

4.3 Learning Method

Learning a $g(U, D)$ without human annotations is not trivial. For example, in a recent paper (Kim et al., 2020), when human labels are removed, the accuracy of knowledge selection drops from 27% to 0.3%. Moreover, since knowledge selection and response generation are entangled, ideally we hope $g(U, D)$ and the GPT-2 model can enhance each other in learning. However, as the parameters of $g(U, D)$ are far from optimal at the early stage, it is very possible that noise from $g(U, D)$ will be fed to the GPT-2 model and then flows back to the learning procedure of $g(U, D)$, resulting in inferior models on both sides. To cope with the challenges, we propose a joint optimization strategy with weak supervision as follows. The learning algorithm is summarized in Algorithm 1.

Pseudo Ground-Truth Construction. To alleviate error accumulation in joint optimization, we consider constructing weak supervision and utilize the signals to warm up the learning of $g(U, D)$ and the fine-tuning of GPT-2 beforehand. The intuition is that responses from humans carry clues to relevance of the knowledge candidates, and thus can be used to construct pseudo ground-truth. To be specific, we first sort $D = \{d_t\}_{t=1}^m$ in a descending order as $\{d_{j_t}\}_{t=1}^m$ according to $\{\text{Sim}(d_t, r)\}_{t=1}^m$ where $\text{Sim}(\cdot, \cdot)$ denotes a similarity function, and then build a subset of D by

$$\bar{D} = \{d_{j_1}, \dots, d_{j_{\bar{m}}}\},$$
$$\bar{m} = \text{argmax}_t(\text{Sim}(d_{j_{1:t}}, r)), \tag{5}$$

where $d_{j_{1:t}}$ refers to the concatenation of $\{d_{j_i}\}_{i=1}^t$. With $\bar{D}$, $g(U, D)$ and the GPT-2 model are optimized via maximum likelihood estimation (MLE) on $\mathcal{D}_K = \{(U_i, D_i, \bar{D}_i)\}_{i=1}^N$ and $\mathcal{D}_G = \{(U_i, \bar{D}_i, r_i)\}_{i=1}^N$ respectively.

Joint Optimization: the Reinforcement Step. We exploit the policy-gradient method (Sutton et al., 2000) to continue-train $g(U, D)$ by which $g(U, D)$ is further "supervised" by the GPT-2 model and is directly optimized for a target metric (e.g., F1 in the experiments). Specifically, we sample a $\tilde{D}$ according to $P(d_i|U, d_{j_{1:t-1}})$ (in Eq.3.) under a termination criterion similar to $\bar{D}$ at each time step, and define the loss function as

$$\mathcal{L}_K = -\frac{1}{N} \sum_{i=1}^N \left(\tilde{R}_i \sum_{t=1}^{|\tilde{D}_i|} \log P(d_{i,j_t}|U_i, d_{i,j_{1:t-1}}) \right),$$
$$\tilde{R}_i = R(\tilde{D}_i) - b, \tag{6}$$

where $R(\tilde{D}_i) = \text{Sim}(r'_i, r_i)$ with r'_i the response generated by the GPT-2 model given U_i and $\tilde{D}_i$, and $b = \sum_{i=1}^N R(\tilde{D}_i)/N$ is the baseline that is used to reduce the variance of gradient estimation(Clark and Manning, 2016). We can see that minimizing $\mathcal{L}_K$ is equivalent to maximizing the conditional likelihood of $\tilde{D}_i$ if it obtains a higher reward than the baseline.

Joint Optimization: the Curriculum Step. Though $g(U, D)$ has been pre-trained with the pseudo ground-truth $\bar{D}$, the relevant knowledge provided by the model (i.e., D') may still be worse than $\bar{D}$ at the beginning of fine-tuning. Therefore, we mix D' and $\bar{D}$ and exploit a curriculum learning strategy to fine-tune the GPT-2 model where D' and $\bar{D}$ are regarded as hard materials and easy materials respectively and fine-tuning gradually moves from $\bar{D}$ to D'. Formally, the loss function for fine-tuning the GPT-2 model is defined by

$$\mathcal{L}_G = -\frac{1}{N} \sum_{i=1}^N \left(z_i \sum_{t=1}^{l_r} \log P(r_{i,t}|U_i, \bar{D}_i, r_{i,1:t-1}) \right.$$
$$\left. + (1 - z_i) \sum_{t=1}^{l_r} \log P(r_{i,t}|U_i, D'_i, r_{i,1:t-1}) \right), \tag{7}$$

where $\{z_i\}$ are sampled from a Bernoulli distribution parameterized by p. By gradually shrinking p, the generation model will be exposed to more hard materials with the learning procedure going on.

Algorithm 1 Optimization Algorithm

Input: Training data $\mathcal{D}$, pre-trained GPT-2, initial curriculum rate p_0, exponential decay constant λ, maximum step M.
Construct $\mathcal{D}_K$ and $\mathcal{D}_G$.
Optimize $g(U, D)$ and GPT-2 using MLE on $\mathcal{D}_K$ and $\mathcal{D}_G$ respectively.
for $m \leftarrow 1$ to M **do**
    Sample a mini-batch $\{(U_i, D_i, r_i)\}$ from $\mathcal{D}$.
    Update the parameters of $g(U, D)$ based on Eq.6. ▷ the Reinforcement Step.
    Sample $\{z_i\}$ from a Bernoulli distribution parameterized by p, where $p = p_0 e^{-\lambda m}$.
    Update the parameters of the GPT-2 model based on Eq.7. ▷ the Curriculum Step.
end for
return $g(U, D)$ and GPT-2.

5 Experiments

We conduct experiments on Wizard of Wikipedia (Wizard) and CMU Document Grounded Conversations (CMU_DoG) (Zhou et al., 2018b).

5.1 Datasets and Evaluation Metrics

Both datasets are built with crowd-sourcing on Amazon Mechanical Turk, employ Wikipedia as the knowledge base, and are split into training sets, validation sets, and test sets by the data owners. Topics in Wizard cover a wide range ($1,365$ in total), and each conversation happens between a wizard who has access to the knowledge about a specific topic and an apprentice who is just eager to learn from the wizard about the topic. The test set is split into two subsets: Test Seen and Test Unseen. Test Seen contains new dialogues with topics appearing in the training set, while topics in Test Unseen never appear in the training set and the validation set. We follow (Dinan et al., 2019) and conduct the pre-processing with the code published on ParlAI[2]. Different from Wizard, CMU_DoG focuses on movie domain, and besides wizard-apprentice conversations, the data also contain conversations between two workers who know the document and try to discuss the content in depth. To better compare with the baselines, we adopt the version shared at https://github.com/lizekang/ITDD. In both data, only the turns where knowledge is accessible are considered in response generation. More details are described in supplementary material.

We choose perplexity (PPL) of the ground-truth responses, BOW Embedding (Liu et al., 2016), and unigram F1 (Dinan et al., 2019) as metrics, where Embedding-based metrics are computed with an NLG evaluation open source available at https://github.com/Maluuba/nlg-eval, and F1 is calculated with the code published at https://github.com/facebookresearch/ParlAI/blob/master/parlai/core/metrics.py.

Besides automatic evaluation, we randomly sample 300 examples from Test Seen, Test Unseen, and the test set of CMU_DoG respectively, and recruit 3 well-educated native speakers as annotators for human evaluation. To each annotator, an example is presented with a context, the associated external knowledge[3], and model responses (top 1 in greedy search) that are randomly shuffled to hide their sources. The annotators then judge the quality of the responses from three aspects, including *fluency*, *context coherence* and *knowledge relevance*, and assign a score in $\{0, 1, 2\}$ (representing "bad", "fair", and "good") to each response for each aspect. Each response receives 3 scores per aspect, and the agreement among the annotators is measured via Fleiss' kappa (Fleiss, 1971).

5.2 Baselines

The following models are selected as baselines:

Transformer Memory Network (TMN): the model proposed in (Dinan et al., 2019) along with the release of the Wizard data. We implement it using the code shared at https://github.com/facebookresearch/ParlAI/blob/master/projects/wizard_of_wikipedia.

Incremental Transformer with Deliberation Decoder (ITDD): a transformer-based model (Li et al., 2019) that incrementally encodes multi-turn dialogues and knowledge and decodes responses with a deliberation technique. We implement it using the code shared at https://github.com/lizekang/ITDD.

Sequential Knowledge Transformer (SKT): a sequential latent variable model with state-of-the-art performance on knowledge selection published in a very recent paper (Kim et al., 2020). Since human labels that indicate ground-truth knowl-

[2] https://github.com/facebookresearch/ParlAI/blob/master/projects/wizard_of_wikipedia

[3] For ease of labeling, only the ground-truth knowledge is shown to the annotators in Wizard.

Models	Test Seen					Test Unseen				
	PPL	F1	Average	Extrema	Greedy	PPL	F1	Average	Extrema	Greedy
TMN (Dinan et al., 2019)	66.5	15.9	0.844	0.427	0.658	103.6	14.3	0.839	0.408	0.645
ITDD (Li et al., 2019)	17.8	16.2	0.841	0.425	0.654	44.8	11.4	0.826	0.364	0.624
SKT* (Kim et al., 2020)	52.0	19.3	0.846	0.440	0.665	81.4	16.1	0.839	0.418	0.652
DRD (Zhao et al., 2020)	19.4	19.3	0.852	0.452	0.674	23.0	17.9	0.849	0.439	0.664
SKT+GPT-2*	17.6	20.3	0.866	0.460	0.679	23.7	17.8	0.860	0.437	0.664
GPT-2$_{trunc}$	14.6(2.2)	18.7(0.7)	0.864(0.002)	0.451(0.006)	0.674(0.004)	16.9(3.1)	18.3(0.6)	0.862(0.002)	0.444(0.005)	0.668(0.003)
KnowledGPT	19.2	**22.0**	**0.872**	**0.463**	**0.682**	22.3	**20.5**	**0.870**	0.452	**0.674**

Table 2: Evaluation results on Wizard. Models that leverage human labels are marked with *. Numbers in bold mean that the improvement to the best baseline is statistically significant (t-test with p-value < 0.01).

Models	PPL	F1	Average	Extrema	Greedy
TMN (Dinan et al., 2019)	75.2	9.9	0.789	0.399	0.615
ITDD (Li et al., 2019)	26.0	10.4	0.748	0.390	0.587
DRD (Zhao et al., 2020)	46.1	10.8	0.791	0.406	0.613
GPT-2$_{trunc}$	18.6	10.8	0.730	0.419	0.597
KnowledGPT	20.6	**13.5**	**0.837**	**0.437**	**0.654**

Table 3: Evaluation results on CMU_DoG. Numbers in bold mean that the improvement to the best baseline is statistically significant (t-test with p-value < 0.01).

edge are crucial to the performance of the model, we only involve it as a baseline on the Wizard data. The model is implemented with the code shared at https://github.com/bckim92/sequential-knowledge-transformer.

Disentangled Response Decoder (DRD): a model that tackles the low-resource challenge with pre-training techniques (Zhao et al., 2020). We choose the one in which all parameters are fine-tuned with the full training data after pre-training as the baseline, since such a configuration results in state-of-the-art performance on Wizard, as reported in (Zhao et al., 2020).

We name our model **KnowledGPT**. Besides the baselines described above, the following pre-trained models are also included in comparison in order to have a thorough understanding towards the proposed method: (1) **GPT-2**$_{trunc}$. We concatenate a context and the associated knowledge as a long document, and then truncate the document to meet the length constraint of the GPT-2 model. This is to check if the simple heuristics work for the task. Note that in Wizard, we randomly mix the ground-truth knowledge with others and repeat the procedure 8 times. The means with standard deviation (i.e., numbers in "( )") are reported to remove randomness; and (2) **SKT+GPT-2**. We feed the candidate selected by SKT to GPT-2 for response generation. This is to examine if we can simply replace the proposed knowledge selection module as well as the learning approach with an off-the-shelf knowledge selection model. Similar to SKT, the comparison is only conducted on Wizard.

5.3 Implementation Details

In both Wizard and CMU_DoG, we set the hidden size and the number of layers of the sequential knowledge selector as 256 and 1 respectively. T_{max} for D' is set as 1 in Wizard, and 2 in CMU_DoG. We choose BERT (110M) and GPT-2 (117M) as the pre-trained language models in KnowledGPT, and implement the models with the code in https://github.com/huggingface/transformers. We employ greedy search in response decoding. All models are learned with Adam (Kingma and Ba, 2015) optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. In warming up, we define $\mathrm{Sim}(\cdot,\cdot)$ as unigram F1, and optimize $g(U, D)$ and the GPT-2 model with the pseudo ground-truth for 1000 steps with a batch size of 64. In joint optimization, the batch size is set as 128, and the learning rates for $g(U, D)$ and GPT-2 are set as $5e-6$ and $5e-5$ respectively. The learning rate will be halved if there is no improvement in terms of PPL on the validation sets. The parameter p of the Bernoulli distribution in the curriculum step is initially set as 1.0 and anneals with a rate of $1e-5$. Early stopping on validation is adopted as a regularization strategy.

5.4 Evaluation Results

Table 2 and Table 3 report evaluation results on Wizard and CMU_DoG respectively. KnowledGPT achieves new state-of-the-art on most metrics in both datasets, which demonstrates the effectiveness of large-scale pre-trained language models on the task of knowledge-grounded dialogue generation. GPT-2$_{trunc}$ is worse than KnowledGPT, due to (1) knowledge loss: we find that in 53% test examples (Test Seen+Test Unseen), the ground-truth knowledge is cut. In this case, GPT-2$_{trunc}$ only relies on the context, the related knowledge in other candidates (thanks to the one-to-many relations between a context and knowledge), and the knowledge packed in the parameters of GPT-2 for responding, which explains the comparable per-

Models	Wizard								CMU_DoG			
	Test Seen				Test Unseen							
	Fluency	Context Coherence	Knowledge Relevance	Kappa	Fluency	Context Coherence	Knowledge Relevance	Kappa	Fluency	Context Coherence	Knowledge Relevance	Kappa
DRD	1.71	1.50	1.26	0.67	1.64	1.44	1.18	0.69	1.58	1.48	1.07	0.60
GPT-2$_{trunc}$	1.86	1.54	1.22	0.71	1.84	1.47	1.20	0.59	1.83	1.58	1.06	0.64
KnowledGPT	1.89	1.67	1.71	0.70	1.88	1.60	1.68	0.73	1.83	1.65	1.50	0.77

Table 4: Human evaluation results on Wizard and CMU_DoG.

Models	Wizard										CMU_DoG				
	Test Seen					Test Unseen									
	PPL	F1	Average	Extrema	Greedy	PPL	F1	Average	Extrema	Greedy	PPL	F1	Average	Extrema	Greedy
KnowledGPT	19.2	22.0	0.872	0.463	0.682	22.3	20.5	0.870	0.452	0.674	20.6	13.5	0.837	0.437	0.654
-pseudo	22.3	18.3	0.857	0.436	0.662	24.1	17.9	0.854	0.430	0.655	23.2	12.9	0.815	0.440	0.639
-joint	20.0	20.4	0.863	0.457	0.675	21.8	19.5	0.861	0.451	0.669	22.6	11.7	0.806	0.438	0.635
-curriculum	19.4	21.2	0.867	0.457	0.677	21.5	20.3	0.866	0.451	0.672	21.9	12.4	0.816	0.443	0.644
-reinforcement	19.4	21.3	0.866	0.459	0.677	21.9	20.2	0.863	0.449	0.670	20.3	12.6	0.817	0.437	0.643

Table 5: Ablation study on Wizard and CMU_DoG

formance with SKT and DRD; and (2) noisy input: even though the ground-truth knowledge is kept, the redundant and irrelevant information in the knowledge candidates are still harmful. Evidence is that GPT-2$_{trunc}$ is worse than KnowledGPT on CMU_DoG even though we do not cut anything on the knowledge (the maximum length of the knowledge input is 502, and thus is within the constraint of GPT-2). KnowledGPT also outperforms SKT+GPT-2 on Wizard, because (1) KnowledGPT is more accurate than SKT on knowledge selection, even though it does not leverage any human annotations in learning. In fact, the accuracy scores of knowledge selection for SKT are 26.8 and 18.3 on Test Seen and Test Unseen respectively, while the two numbers are 28.0 and 25.4 respectively for KnowledGPT; and (2) in KnowledGPT, knowledge selection and response generation are jointly optimized.

Table 4 shows human evaluation results. While the three models are comparable on *fluency*, KnowledGPT is superior to the others on both *context coherence* and *knowledge relevance*, which is consistent with the results on automatic metrics. All kappa values are no less than 0.6, indicating substantial agreement among the annotators. We present a case study in supplementary material.

5.5 Discussions

Ablation study. To understand the impact of the learning strategies on model performance, we compare the full KnowledGPT with the following variants: (1) *-pseudo*: the warming up stage is removed; (2) *-joint*: the joint optimization stage is removed; (3) *-reinforcement*: $g(U, D)$ is fixed after it is optimized with MLE on $\mathcal{D}_K$; and (4) *-curriculum*: GPT-2 is fixed after it is optimized with MLE on $\mathcal{D}_G$. Table 5 reports the evaluation results. We can conclude that (1) the pseudo ground-truth plays a crucial role in Wizard, as removing the step causes dramatic performance drop. This is because in Wizard, there is a strong correlation between the knowledge and human responses. The results indicate that though the pseudo ground-truth is constructed with heuristics, it still contains valuable information and thus allows the following joint optimization to start from a good point. On the other hand, in CMU_DoG, the crowd-workers do not refer to the external knowledge as much as those workers do in Wizard when they form the responses; (2) the reinforcement step and curriculum step are useful because the reinforcement step allows the knowledge selection module to make better use of GPT-2's feedback, and through the curriculum step GPT-2 can take advantage of the output of knowledge selection module progressively; (3) joint optimization is meaningful, as removing this stage results in performance drop.

Models	Wizard				CMU_DoG	
	Test Seen		Test Unseen			
	PPL	F1	PPL	F1	PPL	F1
T_{max}=1	19.2	22.0	22.3	20.5	20.6	12.6
T_{max}=2	18.2	21.3	21.0	20.3	20.6	13.5
T_{max}=3	17.2	21.1	20.2	20.3	19.7	11.2

Table 6: Performance of KnowledGPT under different T_{max}s.

Impact of T_{max} (i.e., the upper bound in knowledge selection). Besides the learning strategies, we are also curious about how T_{max}, as part of the termination criterion in knowledge selection described at the end of Section 4.2, influences the

performance of KnowledGPT. To this end, we vary the value of T_{max} in $\{1, 2, 3\}$ and report the evaluation results in Table 6. The larger T_{max} is, the more chances KnowledGPT has to involve the ground-truth candidate into generation, and the lower PPL is. This also explains why the PPL of GPT-2_{trunc} is lower than that of KnowledGPT in Table 2 and Table 3. On the other hand, a larger T_{max} also means more noise in generation. That is why when T_{max} exceeds a value, F1 begins to drop.

6 Conclusions

We apply large-scaled pre-trained language models to the task of knowledge-grounded dialogue generation. To this end, we devise a knowledge selection module, and propose an unsupervised approach to jointly optimizing knowledge selection and response generation. Evaluation results on two benchmarks indicate that our model can significantly outperform state-of-the-art methods.

Acknowledgments

We would like to thank the reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (No. 2020AAA0105200), the National Science Foundation of China (NSFC No. 61876196 and NSFC No. 61672058). Rui Yan was sponsored as the young fellow of Beijing Academy of Artificial Intelligence (BAAI). Rui Yan is the corresponding author.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Kevin Clark and Christopher D Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262.

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 *ICLR*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13042–13054.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. *Proc. Interspeech 2019*, pages 1891–1895.

Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional dialogue generation using image-grounded language models. In *CHI*, page 277. ACM.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. *arXiv preprint arXiv:2002.07510*.

Seonhoon Kim, Jin-Hyuk Hong, Inho Kang, and Nojun Kwak. 2018. Semantic sentence matching with densely-connected recurrent and co-attentive information. *arXiv preprint arXiv:1805.11360*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *NAACL*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *ACL*, pages 994–1003.

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.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. *arXiv preprint arXiv:1902.04911*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472.

Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the behaviors of bert in ranking. *arXiv preprint arXiv:1904.07531*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam 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 preprint arXiv:1910.10683*.

Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2019. Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. *arXiv preprint arXiv:1908.09528*.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pages 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *ACL*, pages 1577–1586.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Engaging image chat: Modeling personality in grounded dialogue. *arXiv preprint arXiv:1811.00945*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019a. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019b. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of NAACL-HLT*, pages 380–385.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *IJCAI*, pages 4418–4424.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1855–1865.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie. 2018. Learning to ask questions in open-domain conversational systems with typed decoders. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2193–2203.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Chen Xing, Wei Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017a. Topic aware neural response generation. In *AAAI*, pages 3351–3357.

Chen Xing, Wei Wu, Yu Wu, Ming Zhou, Yalou Huang, and Wei-Ying Ma. 2017b. Hierarchical recurrent attention network for response generation. *arXiv preprint arXiv:1701.07149*.

Can Xu, Wei Wu, Chongyang Tao, Huang Hu, Matt Schuerman, and Ying Wang. 2019. Neural response generation with meta-words. *arXiv preprint arXiv:1906.06050*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019a. Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2018a. Learning to control the specificity in neural response generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1108–1117.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019b. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019c. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*, pages 654–664.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020. Low-resource knowledge-grounded dialogue generation. *arXiv preprint arXiv:2002.10348*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018b. A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358*.

A Details of Datasets

Table 7 reports the statistics of the Wizard data and the CMU_DoG data.

	Wizard of Wikipedia				CMU_DoG		
	Train	Valid	Test Seen	Test Unseen	Train	Valid	Test
# Utterances	166,787	17,715	8,715	8,782	74,717	4,993	13,646
# Conversations	18,430	1,948	965	968	3,373	229	619
# Topics/Documents	1,247	599	533	58	30	30	30
Avg. # of Turns	9.0	9.1	9.0	9.1	22.2	21.8	22.0

Table 7: Statistics of the two datasets.

B Comparison with DialoGPT

We compare KnowledGPT and with DialoGPT in order to learn if a pre-trained generation model with state-of-the-art performance on open domain dialogues is already good enough when it is fine-tuned with knowledge-grounded dialogues. We discard the associated knowledge and fine-tune DialoGPT on the knowledge-grounded dialogues. We choose the model trained from OpenAI GPT-2 with 345M parameters, as it shows the best performance in the evaluation in the original paper. The model is implemented based on the code shared at https://github.com/microsoft/DialoGPT.

Table 8 shows the results, indicating that external knowledge is necessary even though one has exploited a powerful pre-trained language model for dialogue generation. In CMU_DoG the gap between DialoGPT and KnowledGPT is narrowed because about 35% of the conversation has a weak correlation with the document (e.g. BLEU < 0.1).

Models	Wizard				CMU_DoG	
	Test Seen		Test Unseen			
	PPL	F1	PPL	F1	PPL	F1
DialoGPT	16.0	17.9	20.0	16.8	16.9	12.3
KnowledGPT	19.2	22.0	22.3	20.5	20.6	13.5

Table 8: Comparison with DialoGPT on Wizard and CMU_DoG

C Impact of Maximum Tokens of GPT-2

To further justify our claims on why GPT-2$_{trunc}$ is worse than KnowledGPT, we keep the ground-truth knowledge in the input sequence of GPT-2 and gradually increase the constraint of the maximum number of tokens on Wizard. As the maximum token limit increases, more irrelevant knowledge is introduced. Note that in practice, one has no way to perfectly locate the ground-truth, and this experiment is only to provide more insights to GPT-2$_{trunc}$. Table 9 shows the performance of GPT-2$_{trunc}$ with the increase of the maximum number of tokens where Ground-truth Percentage indicates the percentage of ground-truth in the input knowledge. First, when the ground-truth is forced to be kept, GPT-2$_{trunc}$ is always better than the one where the ground-truth is randomly mixed with other candidates and bears the risk to be cut. This echoes our claim that knowledge loss is one of the reasons for the poor performance of GPT-2$_{trunc}$ used with the practical setting. Second, even if ground-truth is retained, once more noise is introduced, the performance of GPT-2$_{trunc}$ will become worse. When the length is limited to 128 tokens, the PPL of the model is not good, mainly because under this limitation, the input sequence of some cases only contains the dialogue context and response.

Maximum Tokens	Test Seen		Test Unseen		Ground-truth Percentage
	PPL	F1	PPL	F1	
128	10.8	30.9	11.6	30.4	62.3%
256	9.3	25.6	10.0	24.6	20.3%
512	9.7	21.8	10.5	21.2	8.5%
768	10.1	20.6	10.7	20.2	5.5%
1024	10.7	19.7	11.3	19.4	4.1%

Table 9: Performance of GPT-2$_{trunc}$ under different maximum tokens with ground-truth knowledge involved.

Models	Wizard of Wikipedia				CMUDoG	
	Test Seen		Test Unseen			
	PPL	F1	PPL	F1	PPL	F1
KnowledGPT (117M)	19.2	22.0	22.3	20.5	20.6	13.5
KnowledGPT (345M)	16.1	22.0	17.9	20.6	18.1	13.4

Table 10: Performance of KnowledGPT under different sizes of GPT-2.

D Impact of the Size of GPT-2

We further check if the performance of KnowledGPT can be further improved when the GPT-2 model is replaced with a larger one. Table 10 shows the results. Though GPT-2 (345M) can further reduce PPL, it does not bring significant improvement to F1 over GPT-2 (117M), probably because the larger model can not provide more accurate feedback to the knowledge selection module in learning. Therefore, to balance efficacy and cost, GPT-2 (117M) is still favored in practice.

E Case Study

Table 11 and Table 12 show the examples from Test Seen and Test Unseen of Wizard, each example contains the dialogue context and the background knowledge which is retrieved from Wikipedia given

the last two turns of dialogue and the original topic. We can see that KnowledGPT can locate the knowledge more accurately due to its knowledge selection module and reinforcement learning, and make better use of the associated knowledge with the help of curriculum learning.

Knowledge (Topic: Cinematography)	
...	
Cinematography (also called ”direction of photography”) is the science or art of motion-picture photography by recording light or other electromagnetic radiation, either electronically by means of an image sensor, or chemically by means of a light-sensitive material such as film stock.	
Typically, a lens is used to repeatedly focus the light reflected from objects into real images on the light-sensitive surface inside a camera during a questioned exposure, creating multiple images.	
With an electronic image sensor, this produces an electrical charge at each pixel, which is electronically processed and stored in a video file for subsequent display or processing.	
The result with photographic emulsion is a series of invisible latent images on the film stock, which are later chemically ”developed” into a visible image.	
The images on the film stock are played back at a rapid speed and projected onto a screen, creating the illusion of motion.	
Cinematography finds uses in many fields of science and business as well as for entertainment purposes and mass communication.	
The word ”cinematography” was created from the greek words (”kinema”), meaning ”movement, motion” and (”graphein”) meaning ”to record”, together meaning ”recording motion.”	
...	
Context	
B	Thinking about taking a class on cinematography, the direction of photography .
A	I have always wanted to learn that .
B	Me too, I have learned its the science and or art of recording motion picture photography .
A	That is cool. did you enjoy it .
Response	
TMN	I think it is a great idea . It is a science fiction film that combines computer and computer science .
ITDD	I did . It was developed in the late 19th century .
SKT	Yes , I did . I loved it . It was used to focus the light reflected from objects into real images .
DRD	I did enjoy it . It is also called “ direction of photography ” .
SKT+GPT-2	I did! I love how it’s so easy to focus light reflected from objects into real images on the light-sensitive surface inside a camera.
GPT-2$_{trunc}$	Yes, I did, I love the fact that it is the science and art of recording motion picture photography
KnowledGPT	Yes, I did, I love cinematography, its used in many fields of science and business as well as for entertainment purposes and mass communication
Human	Yes and cinematography skills can also be used in mass marketing and other fields as well

Table 11: A case from Test Seen of Wizard of Wikipedia

Knowledge(Topic: Green)	
...	
Green is the color between blue and yellow on the visible spectrum.	
It is evoked by light which has a dominant wavelength of roughly 495570 nm.	
In subtractive color systems, used in painting and color printing, it is created by a combination of yellow and blue, or yellow and cyan; in the rgb color model, used on television and computer screens, it is one of the additive primary colors, along with red and blue, which are mixed in different combinations to create all other colors.	
By far the largest contributor to green in nature is chlorophyll, the chemical by which plants photosynthesize and convert sunlight into chemical energy.	
Many creatures have adapted to their green environments by taking on a green hue themselves as camouflage.	
Several minerals have a green color, including the emerald, which is colored green by its chromium content.	
In the middle ages and renaissance, green was the color commonly associated with merchants, bankers and the gentry, while red was reserved for the nobility.	
For this reason, the costume of the ”mona lisa” by leonardo da vinci and the benches in the british house of commons are green while those in the house of lords are red.	
It also has a long historical tradition as the color of ireland and of gaelic culture.	
...	
Context	
B	Green is my favorite color. Many minerals are a green color, including emeralds. I like emeralds too!
A	I like green too. My eyes are green.
B	So are mine! In gaelic cultures green is very popular, with a long historical tradition. It is known as the color of ireland.
A	Yes I do know it is a very popular color in ireland.
Response	
TMN	I think it ’ s a very interesting subject . It ’ s a culture of culture.
ITDD	Yes , it is one of the primary colors .
SKT	It is a popular color in the us
DRD	It is common in nature , especially in plants .
SKT+GPT-2	Yes, it is very popular in gaelic culture.
GPT-2$_{trunc}$	It is! It is also the second largest island in the british isles.
KnowledGPT	Yes, it is! In the middle ages and renaissance, green was the color commonly associated with merchants, bankers and the gentry.
Human	Back in the middle ages green was often associated with merchants and bankers. Red was nobility’s color of choice.

Table 12: A case from Test Uneen of Wizard of Wikipedia