A Pre-training Strategy for Zero-Resource Response Selection in Knowledge-Grounded Conversations

Chongyang Tao[1*], **Changyu Chen**[2*], **Jiazhan Feng**[1], **Jirong Wen**[2,3] and **Rui Yan**[2,3†]
[1]Peking University, Beijing, China
[2]Gaoling School of Artificial Intelligence, Renmin University of China
[3]Beijing Academy of Artificial Intelligence
[1]{chongyangtao,fengjiazhan}@pku.edu.cn
[2]{chen.changyu,jrwen,ruiyan}@ruc.edu.cn

Abstract

Recently, many studies are emerging towards building a retrieval-based dialogue system that is able to effectively leverage background knowledge (e.g., documents) when conversing with humans. However, it is non-trivial to collect large-scale dialogues that are naturally grounded on the background documents, which hinders the effective and adequate training of knowledge selection and response matching. To overcome the challenge, we consider decomposing the training of the knowledge-grounded response selection into three tasks including: 1) query-passage matching task; 2) query-dialogue history matching task; 3) multi-turn response matching task, and joint learning all these tasks in a unified pre-trained language model. The former two tasks could help the model in knowledge selection and comprehension, while the last task is designed for matching the proper response with the given query and background knowledge (dialogue history). By this means, the model can be learned to select relevant knowledge and distinguish proper response, with the help of ad-hoc retrieval corpora and a large number of ungrounded multi-turn dialogues. Experimental results on two benchmarks of knowledge-grounded response selection indicate that our model can achieve comparable performance with several existing methods that rely on crowd-sourced data for training.

1 Introduction

Along with the very recent prosperity of artificial intelligence empowered conversation systems in the spotlight, many studies have been focused on building human-computer dialogue systems (Wen et al., 2017; Zhang et al., 2020) with either retrieval-based methods (Wang et al., 2013; Wu et al., 2017; Whang et al., 2020) or generation-based methods (Li et al., 2016; Serban et al., 2016; Zhang et al., 2020), which both predict the response with only the given context. In fact, unlike a person who may associate the conversation with the background knowledge in his or her mind, the machine can only capture limited information from the query message itself. As a result, it is difficult for a machine to properly comprehend the query, and to predict a proper response to make it more engaging. To bridge the gap of the knowledge between the human and the machine, researchers have begun to simulating this motivation by grounding dialogue agents with background knowledge (Zhang et al., 2018; Dinan et al., 2019; Li et al., 2020), and lots of impressive results have been obtained.

In this paper, we consider the response selection problem in knowledge-grounded conversion and specify the background knowledge as unstructured documents that are common sources in practice. The task is that given a conversation context and a set of knowledge entries, one is required 1): to select proper knowledge and grasp a good comprehension of the selected document materials (knowledge selection); 2): to distinguish the true response from a candidate pool that is relevant and consistent with both the conversation context and the background documents (knowledge matching).

While there exists a number of knowledge documents on the Web, it is non-trivial to collect large-scale dialogues that are naturally grounded on the documents for training a neural response selection model, which hinders the effective and adequate training of knowledge selection and response matching. Although some benchmarks built upon crowd-sourcing have been released by recent works (Zhang et al., 2018; Dinan et al., 2019), the relatively small training size makes it hard for the dialogue models to generalize on other domains or topics (Zhao et al., 2020). Thus, in this work, we

[*] Equal Contribution.
[†] Corresponding author: Rui Yan (ruiyan@ruc.edu.cn).

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 4446–4457
August 1–6, 2021. ©2021 Association for Computational Linguistics

focus on a more challenging and practical scenario, learning a knowledge-grounded conversation agent without any knowledge-grounded dialogue data, which is known as zero-resource settings.

Since knowledge-grounded dialogues are unavailable in training, it raises greater challenges for learning the grounded response selection model. Fortunately, there exists a large number of unstructured knowledge (e.g., web pages or wiki articles), passage search datasets (e.g., query-passage pairs coming from ad-hoc retrieval tasks) (Khattab and Zaharia, 2020) and multi-turn dialogues (e.g., context-response pairs collected from Reddit) (Henderson et al., 2019), which might be beneficial to the learning of knowledge comprehension, knowledge selection and response prediction respectively. Besides, in multi-turn dialogues, the background knowledge and conversation history (excluding the latest query) are symmetric in terms of the information they convey, and we assume that the dialogue history can be regarded as another format of background knowledge for response prediction.

Based on the above intuition, in this paper, we consider decomposing the training of the grounded response selection task into several sub-tasks, and joint learning all those tasks in a unified model. To take advantage of the recent breakthrough on pre-training for natural language tasks, we build the grounded response matching models on the basis of a pre-trained language model (PLMs) (Devlin et al., 2019; Yang et al., 2019), which are trained with large-scale unstructured documents from the web. On this basis, we further train the PLMs with query-passage matching task, query-dialogue history matching task, and multi-turn response matching task jointly. The former two tasks could help the model not only in knowledge selection but also in knowledge (and dialogue history) comprehension, while the last task is designed for matching the proper response with the given query and background knowledge (dialogue history). By this means, the model can be learned to select relevant knowledge and distinguish proper responses, with the help of a large number of ungrounded dialogues and ad-hoc retrieval corpora. During the testing stage, we first utilize the trained model to select proper knowledge, and then feed the query, dialogue history, selected knowledge, and the response candidate into our model to calculate the final matching degree. Particularly, we design two strategies to compute the final matching score. In the first strategy, we directly concatenate the selected knowledge and dialogue history as a long sequence of background knowledge and feed into the model. In the second strategy, we first compute the matching degree between each query-knowledge and the response candidates, and then integrate all matching scores.

We conduct experiments with benchmarks of knowledge-grounded dialogue that are constructed by crowd-sourcing, such as the Wizard-of-Wikipedia Corpus (Dinan et al., 2019) and the CMU_DoG Corpus (Zhou et al., 2018a). Evaluation results indicate that our model achieves comparable performance on knowledge selection and response selection with several existing models trained on crowd-sourced benchmarks.

Our contributions are summarized as follows:

- To the best of our knowledge, this is the first exploration of knowledge-grounded response selection under the zero-resource setting.
- We propose decomposing the training of the grounded response selection models into several sub-tasks, so as to empower the model through these tasks in knowledge selection and response matching.
- We achieve a comparable performance of response selection with several existing models learned from crowd-sourced training sets.

2 Related Work

Early studies of retrieval-based dialogue focus on single-turn response selection where the input of a matching model is a message-response pair (Wang et al., 2013; Ji et al., 2014; Wang et al., 2015). Recently, researchers pay more attention to multi-turn context-response matching and usually adopt the representation-matching-aggregation paradigm to build the model. Representative methods include the dual-LSTM model (Lowe et al., 2015), the sequential matching network (SMN) (Wu et al., 2017), the deep attention matching network (DAM) (Zhou et al., 2018b), interaction-over-interaction network (IoI) (Tao et al., 2019) and multi-hop selector network (MSN) (Yuan et al., 2019). More recently, pre-trained language models (Devlin et al., 2019; Yang et al., 2019) have shown significant benefits for various NLP tasks, and some researchers have tried to apply them on multi-turn response selection. Vig and Ramea (2019) exploit BERT to represent each utterance-response pair and fuse these representations to calculate the matching score; Whang et al. (2020) and Xu et al. (2020) treat the context as a long sequence and conduct context-response matching with BERT. Besides, Gu et al. (2020a) integrate speaker embeddings into BERT to improve the utterance representation in multi-turn dialogue.

To bridge the gap of the knowledge between the human and the machine, researchers have investigated into grounding dialogue agents with unstructured background knowledge (Ghazvininejad et al., 2018; Zhang et al., 2018; Dinan et al., 2019). For example, Zhang et al. (2018) build a persona-based conversation data set that employs the interlocutor's profile as the background knowledge; Zhou et al. (2018a) publish a data where conversations are grounded in articles about popular movies; Dinan et al. (2019) release another document-grounded data with Wiki articles covering a wide range of topics. Meanwhile, several retrieval-based knowledge-grounded dialogue models are proposed, such as document-grounded matching network (DGMN) (Zhao et al., 2019) and dually interactive matching network (DIM) (Gu et al., 2019) which let the dialogue context and all knowledge entries interact with the response candidate respectively via the cross-attention mechanism. Gu et al. (2020b) further propose to pre-filter the context and the knowledge and then use the filtered context and knowledge to perform the matching with the response. Besides, with the help of gold knowledge index annotated by human wizards, Dinan et al. (2019) consider joint learning the knowledge selection and response matching in a multi-task manner or training a two-stage model.

3 Model

In this section, we first formalize the knowledge-grounded response matching problem and then introduce our method from preliminary to response matching with PLMs to details of three pre-training tasks.

3.1 Problem Formalization

We first describe a standard knowledge-grounded response selection task such as Wizard-of-Wikipedia. Suppose that we have a knowledge-grounded dialogue data set $\mathcal{D} = \{k_i, c_i, r_i, y_i\}_{i=1}^N$ where $k_i = \{p_1, p_2, \ldots, p_{l_k}\}$ represents a collection of knowledge with p_j the j-th knowledge entry (a.k.a., passage) and l_k is the number of entries; $c_i = \{u_1, u_2, \ldots, u_{l_c}\}$ denotes multi-turn dialogue context with u_j the j-th turn and l_c is the number of dialogue turns. It should be noted that in this paper we denote the latest turn u_{l_c} as dialogue query q_i, and dialogue context except for query is denoted as $h_i = c_i/\{q_i\}$. r_i stands for a candidate response. $y_i = 1$ indicates that r_i is a proper response for c_i and k_i, otherwise $y_i = 0$. N is the number of samples in data set. The goal knowledge-grounded dialogue is to learn a matching model $g(k, c, r)$ from $\mathcal{D}$, and thus for any new (k, c, r), $g(k, c, r)$ returns the matching degree between r and (k, c). Finally, one can collect the matching scores of a series of candidate responses and conduct response ranking.

Zero-resource grounded response selection then is formally defined as follows. There is a standard multi-turn dialogue dataset $\mathcal{D}_c = \{q_i, h_i, r_i\}_{i=1}^N$ and an ad-hoc retrieval dataset $\mathcal{D}_p = \{q_i, p_i, z_i\}_{i=1}^M$ where q_i is a query and p_i stands a candidate passage, $z_i = 1$ indicates that p_i is a relevant passage for q_i, otherwise $z_i = 0$. Our goal is to learn a model $g(k, h, q, r)$ from $\mathcal{D}_c$ and $\mathcal{D}_p$, and thus for any new input (k, h, q, r), our model can select proper knowledge $\hat{k}$ from k and calculate the matching degree between r and $(\hat{k}, q, h)$.

3.2 Preliminary: Response Matching with PLMs

Pre-trained language models have been widely used in many NLP tasks due to the strong ability of language representation and understanding. In this work, we consider building a knowledge-grounded response matching model with BERT.

Specifically, given a query q, a dialogue history $h = \{u_1, u_2, ..., u_{n_h}\}$ where u_i is the i-th turn in the history, a response candidate $r = \{r_1, r_2, ..., r_{l_r}\}$ with l_r words, we concatenate all sequences as a single consecutive tokens sequence with special tokens, which can be represented as $x = \{[\texttt{CLS}], u_1, [\texttt{SEP}], \ldots, [\texttt{SEP}], u_{l_h}, [\texttt{SEP}], q, [\texttt{SEP}], r, [\texttt{SEP}]\}$. [CLS] and [SEP] are classification symbol and segment separation symbol respectively. For each token in x, BERT uses a summation of three kinds of embeddings, including WordPiece embedding (Wu et al., 2016), segment embedding, and position embedding.

Then, the embedding sequence of x is fed into BERT, giving us the contextualized embedding sequence $\{E_{[\texttt{CLS}]}, E_2, \ldots, E_{l_x}\}$. $E_{[\texttt{CLS}]}$ is an aggregated representation vector that contains the

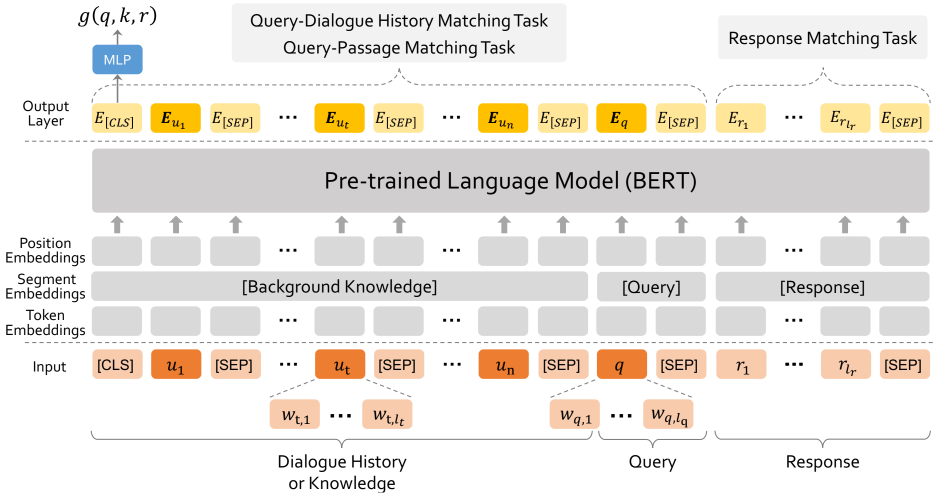

Figure 1: The overall architecture of our model.

semantic interaction information between the query, history, and response candidate. Finaly, $E_{\texttt{[CLS]}}$ is fed into a non-linear layer to calculate the final matching score, which is formulated as:

$$g(h,q,r) = \sigma(W_2 \cdot \tanh(W_1 E_{\texttt{[CLS]}} + b_1) + b_2) \quad (1)$$

where $W_{\{1,2\}}$ and $b_{\{1,2\}}$ is training parameters for response selection task, σ is a sigmoid function.

In knowledge-grounded dialogue, each dialogue is associated with a large collection of knowledge entries $k = \{p_1, p_2, \dots, p_{l_k}\}$[1]. The model is required to select $m (m \geq 1)$ knowledge entries based on semantic relevance between the query and each knowledge, and then performs the response matching with the query, dialogue history and the highly-relevant knowledge. Specifically, we denote $\hat{k} = (\hat{p}_1, \dots, \hat{p}_m)$ as the selected knowledge entries, and feed the input sequence $x = \{\texttt{[CLS]}, \hat{p}_1, \texttt{[SEP]}, \dots, \texttt{[SEP]}, \hat{p}_m, \texttt{[SEP]}, u_1, \texttt{[SEP]}, \dots, \texttt{[SEP]}, u_{l_h}, \texttt{[SEP]}, q, \texttt{[SEP]}, r, \texttt{[SEP]}\}$ to BERT. The final matching score $g(\hat{k}, h, q, r)$ can be computed based on $\texttt{[CLS]}$ representation.

3.3 Pre-training Strategies

On the basis of BERT, we further jointly train it with three tasks including *1) query-passage matching task*; *2) query-dialogue history matching task*; *3) multi-turn response matching task*. The former two tasks could help the model in knowledge selection and knowledge (and dialogue history) comprehension, while the last task is designed for matching the proper response with the given query and background knowledge (dialogue history). By this means, the model can be learned to select relevant knowledge and distinguish the proper response, with the help of a large number of ungrounded dialogues and ad-hoc retrieval corpora.

3.3.1 Query-Passage Matching

Although there exist a huge amount of conversation data on social media, it is hard to collect sufficient dialogues that are naturally grounded on knowledge documents. Existing studies (Dinan et al., 2019) usually extract the relevant knowledge before the response matching or jointly train the knowledge retrieval and response selection in a multi-task manner. However, both methods need in-domain knowledge-grounded dialogue data (with gold knowledge label) to train, making the model hard to generalize to a new domain. Fortunately, the ad-hoc retrieval task (Harman, 2005; Khattab and Zaharia, 2020) in the information retrieval area provides a potential solution to simulate the process of knowledge seeking. To take advantage of the parallel data in the ad-hoc retrieval task, we consider incorporating the query-passage matching task, so as to help the knowledge selection and knowledge comprehension for our task.

Given a query-passage pair (q, p), we first concatenate the query q and the passage p as a single consecutive token sequence with special tokens separating them, which is formulated as:

$$S^{qp} = \{\texttt{[CLS]}, w_1^p, \dots, w_{n_p}^p, \texttt{[SEP]}, w_1^q, \dots, w_{n_q}^q\} \quad (2)$$

where w_i^p, w_j^q denotes the i-th and j-th token of knowledge entry p and query q respectively. For each token in S_i^{qp}, *token*, *segment* and *position*

[1] The scale of the knowledge referenced by each dialogue usually exceeds the limitation of input length in PLMs.

embeddings are summated and fed into BERT. It is worth noting that here we set the segment embedding of the knowledge to be the same as the dialogue history. Finally, we feed the output representation of [CLS] $E^{\text{qp}}_{\texttt{[CLS]}}$ into a MLP to obtain the final query-passage matching score $g(q, p)$. The loss function of each training sample for query-passage matching task is defined by

$$\begin{aligned}&\mathcal{L}_{\text{p}}(q, p^{+}, p^{-}_{1}, \ldots, p^{-}_{n_p})\\ &= -\log(\frac{e^{g(q,p^{+})}}{e^{g(q,p^{+})} + \sum_{j=1}^{\delta_p} e^{g(q,p^{-}_{j})}})\end{aligned} \tag{3}$$

where p^{+} stands for the positive passage for q, p^{-}_{j} is the j-th negative passage and δ_p is the number of negative passage.

3.3.2 Query-Dialogue History Matching

In multi-turn dialogues, the conversation history (excluding the latest query) is a piece of supplementary information for the current query and can be regarded as another format of background knowledge during the response matching. Besides, due to the natural sequential relationship between dialogue turns, the dialogue query usually shows a strong semantic relevance with the previous turns in the dialogue history. Inspired by such characteristics, we design a query-dialogue history matching task with the multi-turn dialogue context, so as to enhance the capability of the model to comprehend the dialogue history with the given dialogue query and to rank relevant passages with these pseudo query-passage pairs.

Specifically, we first concatenate the dialogue history into a long sequence. The task requires the model to predict whether a query $q = \{w^{q}_{1}, \ldots, w^{q}_{n_q}\}$ and a dialogue history sequence $h = \{w^{h}_{1}, \ldots, w^{h}_{n_h}\}$ are consecutive and relevant. We concatenate two sequences into a single consecutive sequence with [SEP] tokens,

$$S^{qh} = \{\texttt{[CLS]}, w^{h}_{1}, \ldots, w^{h}_{n_h}, \texttt{[SEP]}, w^{q}_{1}, \ldots, w^{q}_{n_q}\} \tag{4}$$

For each word in S^{qh}, *token*, *segment* and *position* embeddings are summated and fed into BERT. Finally, we feed $E^{\text{qh}}_{\texttt{[CLS]}}$ into a MLP to obtain the final query-history matching score $g(q, h)$. The loss function of each training sample for query-history matching task is defined by

$$\begin{aligned}&\mathcal{L}_{\text{h}}(q, h^{+}, h^{-}_{1}, \ldots, h^{-}_{n_h})\\ &= -\log(\frac{e^{g(q,h^{+})}}{e^{g(q,h^{+})} + \sum_{j=1}^{\delta_h} e^{g(q,h^{-}_{j})}})\end{aligned} \tag{5}$$

where h^{+} stands for the true dialogue history for q, h^{-}_{j} is the j-th negative dialogue history randomly sampled from the training set and δ_h is the number of sampled dialogue history.

3.3.3 Multi-turn Response Matching

The above two tasks are designed for empowering the model to knowledge or history comprehension and knowledge selection. In this task, we aim at training the model to match reasonable responses based on dialogue history and query. Since we treat the dialogue history as a special form of background knowledge and they share the same segment embeddings in the PLMs, our model can acquire the ability to identify the proper response with either dialogue history or the background knowledge through the multi-turn response matching task.

Specifically, we format the multi-turn dialogues as query-history-response triples and requires the model to predict whether a response candidate $r = \{w^{r}_{1}, \ldots, w^{r}_{n_r}\}$ is appropriate for a given query $q = \{w^{q}_{1}, \ldots, w^{q}_{n_q}\}$ and a concatenated dialogue history sequence $h = \{w^{h}_{1}, \ldots, w^{h}_{n_h}\}$. Concretely, we concatenate three input sequences into a single consecutive tokens sequence with [SEP] tokens,

$$\begin{aligned}S^{hqr} = \{&\texttt{[CLS]}, w^{h}_{1}, \ldots, w^{h}_{n_h}, \texttt{[SEP]},\\ &w^{q}_{1}, \ldots, w^{q}_{n_q}, \texttt{[SEP]}, w^{r}_{1}, \ldots, w^{r}_{n_r}\}\end{aligned} \tag{6}$$

Similarly, we feed an embedding sequence of which each entry is a summation of *token*, *segment* and *position* embeddings into BERT. Finally, we feed $E^{\text{hqr}}_{\texttt{[CLS]}}$ into a MLP to obtain the final response matching score $g(h, q, r)$.

The loss function of each training sample for multi-turn response matching task is defined by

$$\begin{aligned}&\mathcal{L}_{\text{r}}(h, q, r^{+}, r^{-}_{1}, \ldots, r^{-}_{\delta_r})\\ &= -\log(\frac{e^{g(h,q,r^{+})}}{e^{g(h,q,r^{+})} + \sum_{i=j}^{n_r} e^{g(h,q,r^{-}_{j})}})\end{aligned} \tag{7}$$

where r^{+} is the true response for a given q and h, r^{-}_{j} is the j-th negative response candidate randomly sampled from the training set and δ_r is the number of negative response candidate.

3.3.4 Joint Learning

We adopt a multi-task learning manner and define the final objective function as:

$$\mathcal{L}_{\texttt{final}} = \mathcal{L}_{\text{p}} + \mathcal{L}_{\text{h}} + \mathcal{L}_{\text{r}} \tag{8}$$

In this way, all tasks are jointly learned so that the model can effectively leverage two training corpus and learn to select relevant knowledge and distinguish the proper response.

3.4 Calculating Matching Score

After learning model from $\mathcal{D}_c$ and $\mathcal{D}_p$, we first rank $\{p_i\}_{i=1}^{n_k}$ according to $g(q, k_i)$ and then select top m knowledge entries $\{p_1, \ldots, p_m\}$ for the subsequent response matching process. Here we design two strategies to compute the final matching score $g(k, h, q, r)$. In the first strategy, we directly concatenate the selected knowledge and dialogue history as a long sequence of background knowledge and feed into the model to obtain the final matching score, which is formulated as,

$$g(k, h, q, r) = g(p_1 \oplus \ldots \oplus p_m \oplus c, q, r) \quad (9)$$

where $\oplus$ denotes the concatenation operation.

In the second strategy, we treat each selected knowledge entry and the dialogue history equally as the background knowledge, and compute the matching degree between each query, background knowledge, and the response candidates with the trained model. Consequently, the matching score is defined as an integration of a set of knowledge-grounded response matching scores, formulated as,

$$g(k, h, q, r) = g(h, q, r) + \max_{i \in (0,m)} g(p_i, q, r) \quad (10)$$

where m is the number of selected knowledge entries. We name our model with the two strategies as PTKGC$_{\texttt{cat}}$ and PTKGC$_{\texttt{sep}}$ respectively. We compare the two learning strategies through empirical studies, as will be reported in the next section.

4 Experiments

4.1 Datasets and Evaluation Metrics

Training Set. We adopt MS MARCO passage ranking dataset (Nguyen et al., 2016) built on Bing's search for query-passage matching task. The dataset contains 8.8M passages from Web pages gathered from Bing's results to real-world queries and each passage contains an average of 55 words. Each query is associated with sparse relevance judgments of one (or very few) passage marked as relevant. The training set contains about 500k pairs of query and relevant passage, and another 400M pairs of query and passages that have not been marked as relevant, from which the negatives are sampled in our task.

For the query-dialogue history matching task and multi-turn response matching task, we use the multi-turn dialogue corpus constructed from the Reddit (Dziri et al., 2018). The dataset contains more than 15 million dialogues and each dialogue has at least 3 utterances. After the pre-processing, we randomly sample 2.28M/20K dialogues as the training/validation set. For each dialogue session, we regard the last turn as the response, the last but one as the query, and the rest as the positive dialogue history. The negative dialogue histories are randomly sampled from the whole dialogue set. On average, each dialogue contains 4.3 utterances, and the average length of the utterances is 42.5.

Test Set. We tested our proposed method on the Wizard-of-Wikipedia (WoW) (Dinan et al., 2019) and CMU_DoG (Zhou et al., 2018a). Both datasets contain multi-turn dialogues grounded on a set of background knowledge and are built with crowd-sourcing on Amazon Mechanical Turk. In WoW, the given knowledge collection is obtained from Wikipedia and covers a wide range of topics or domains, while in CMU_DoG, the underlying knowledge focuses on the movie domain. Unlike CMU_DoG where the golden knowledge index for each turn is unknown, the golden knowledge index for each turn is provided in WoW. Two configurations (e.g., test-seen and test-unseen) are provided in WoW. Following existing works (Dinan et al., 2019; Zhao et al., 2019), positive responses are true responses from humans and negative ones are randomly sampled. The ratio between positive and negative responses is $1 : 99$ for WoW and $1 : 19$ for CMU_DoG. More details of the two benchmarks are shown in Appendix A.1.

Evaluation Metrics. Following previous works on knowledge-grounded response selection (Gu et al., 2020b; Zhao et al., 2019), we also employ recall n at k R$_n$@k (where $n = 100$ for WoW and $n = 20$ for CMU_DoG and $k = \{1, 2, 5\}$) as the evaluation metrics.

4.2 Implementation Details

Our model is implemented by PyTorch (Paszke et al., 2019). Without loss of generality, we select English uncased BERT$_{\text{base}}$ (110M) as the matching model. During the training, the maximum lengths of the knowledge (a.k.a., passage), the dialogue history, the query, and the response candidate were set to 128, 120 60, and 40. Intuitively, the last tokens in the dialogue history and the previous

Models	Test Seen			Test Unseen		
	R@1	R@2	R@5	R@1	R@2	R@5
IR Baseline	17.8	-	-	14.2	-	-
BoW MemNet	71.3	-	-	33.1	-	-
Two-stage Transformer	84.2	-	-	63.1	-	-
Transformer MemNet	87.4	-	-	69.8	-	-
DIM (Gu et al., 2019)	83.1	91.1	95.7	60.3	77.8	92.3
FIRE (Gu et al., 2020b)	88.3	95.3	97.7	68.3	84.5	95.1
PTKGC$_{\texttt{cat}}$	85.7	94.6	98.2	65.5	82.0	94.7
PTKGC$_{\texttt{sep}}$	89.5	96.7	98.9	69.6	85.8	96.3

Table 1: Evaluation results on the test set of WoW.

Models	R@1	R@2	R@5
Starspace (Wu et al., 2018)	50.7	64.5	80.3
BoW MemNet (Zhang et al., 2018)	51.6	65.8	81.4
KV Profile Memory (Zhang et al., 2018)	56.1	69.9	82.4
Transformer MemNet (Mazaré et al., 2018)	60.3	74.4	87.4
DGMN (Zhao et al., 2019)	65.6	78.3	91.2
DIM (Gu et al., 2019)	78.7	89.0	97.1
FIRE (Gu et al., 2020b)	81.8	90.8	97.4
PTKGC$_{\texttt{cat}}$	61.6	73.5	86.1
PTKGC$_{\texttt{sep}}$	66.1	77.8	88.7

Table 2: Evaluation results on the test set of CMU_DoG.

tokens in the query and response candidate are more important, so we cut off the previous tokens for the context but do the cut-off in the reverse direction for the query and response candidate if the sequences are longer than the maximum length. We set a batch size of 32 for multi-turn response matching and query-dialogue history matching, and 8 for query-document matching in order to train these tasks jointly under the circumstance of training examples inequality. We set $\delta_p = 6$, $\delta_h = 1$ and $\delta_r = 12$ for the query-passage matching, the query-dialogue history matching and the multi-turn response matching respectively. Particularly, the negative dialogue histories are sampled from other training instances in a batch. The model is optimized using Adam optimizer with a learning rate set as $5e-6$. The learning rate is scheduled by warmup and linear decay. A dropout rate of 0.1 is applied for all linear transformation layers. The gradient clipping threshold is set as 10.0. Early stopping on the corresponding validation data is adopted as a regularization strategy. During the testing, we vary the number of selected knowledge-entries $m \in \{1, \ldots, 15\}$ and set $m = 2$ for PTKGC$_{\texttt{cat}}$ and set $m = 14$ for PTKGC$_{\texttt{sep}}$ because they achieve the best performance.

4.3 Baselines

Since the characteristics of the two data sets are different (only WoW provides the golden knowledge label), we compare the proposed model with the baselines on both data sets individually.

Baselines on WoW. 1) *IR Baseline* (Dinan et al., 2019) uses simple word overlap for response selection; 2) *BoW MemNet* (Dinan et al., 2019) is a memory network where knowledge entries are embedded via bag-of-words representation, and the model learns the knowledge selection and response matching jointly; 3) *Transformer MemNet* (Dinan et al., 2019) is an extension of BoW MemNet, and the dialogue history, response candidate and knowledge entries are encoded with Transformer encoder (Vaswani et al., 2017) pre-trained on a large data set. 4) *Two-stage Transformer* (Dinan et al., 2019) trains two separately models for knowledge selection and response retrieval respectively. A best-performing model on the knowledge selection task is used for the dialogue retrieval task.

Baselines on CMU_DoG 1) *Starspace* (Wu et al., 2018) selects the response by the cosine similarity between a concatenated sequence of dialogue context, knowledge, and the response candidate represented by StarSpace (Wu et al., 2018); 2) *BoW MemNet* (Zhang et al., 2018) is a memory network with the bag-of-words representation of knowledge entries as the memory items; 3) *KV Profile Memory* (Zhang et al., 2018) is a key-value memory network grounded on knowledge profiles; 4) *Transformer MemNet* (Mazaré et al., 2018) is similar to BoW MemNet and all utterances are encoded with a pre-trained Transformer; 5) *DGMN* (Zhao et al., 2019) lets the dialogue context and all knowledge entries interact with the response candidate respectively via the cross-attention; 6) *DIM* (Gu et al., 2019) is similar to DGMN and all utterance are encoded with BiLSTMs; 7) *FIRE* (Gu et al., 2020b) first filters the context and knowledge and then use the filtered context and knowledge to perform the iterative response matching process.

4.4 Evaluation Results

Performance of Response Selection. Table 1 and Table 2 report the evaluation results of response selection on WoW and CMU_DoG where PTKGC$_{\texttt{cat}}$ and PTKGC$_{\texttt{sep}}$ represent the final matching score computed with the first strategy (Equation 9) and the second strategy (Equation 10) respectively. We can see that PTKGC$_{\texttt{sep}}$ is

Models	Wizard of Wikipedia						CMU_DoG		
	Test Seen			Test Unseen					
	R@1	R@2	R@5	R@1	R@2	R@5	R@1	R@2	R@5
PTKGC$_{\texttt{sep}}$	89.5	96.7	98.9	69.6	85.8	96.3	66.1	77.8	88.7
PTKGC$_{\texttt{sep}}$ (q)	70.6	79.7	86.8	55.9	70.8	83.4	47.3	58.8	75.0
PTKGC$_{\texttt{sep}}$ (q+h)	84.9	93.9	97.8	64.9	81.7	94.3	59.5	72.3	86.1
PTKGC$_{\texttt{sep}}$ (q+k)	89.5	96.4	98.6	67.0	84.0	96.0	62.7	73.8	84.8
PTKGC$_{\texttt{sep,m=1}}$	85.6	94.4	97.9	66.7	82.8	94.3	60.4	72.5	86.0
PTKGC$_{\texttt{sep,m=1}}$ - $\mathcal{L}_{\mathrm{p}}$	84.7	93.5	97.5	63.4	80.5	94.0	58.7	70.8	85.6
PTKGC$_{\texttt{sep,m=1}}$ - $\mathcal{L}_{\mathrm{h}}$	84.9	93.7	97.6	65.5	81.7	94.1	59.4	71.4	85.3

Table 3: Ablation study.

Models	Wizard Seen			Wizard Unseen		
	R@1	R@2	R@5	R@1	R@2	R@5
Random	2.7	-	-	2.3	-	-
IR Baseline	5.8	-	-	7.6	-	-
BoW MemNet	23.0	-	-	8.9	-	-
Transformer	22.5	-	-	12.2	-	-
Transformer (w/ pretrain)	25.5	-	-	22.9	-	-
Our Model	22.0	31.2	48.8	23.1	32.1	50.7
Our Model - $\mathcal{L}_{\mathrm{p}}$	12.8	22.6	45.2	13.3	23.3	45.5
Our Model - $\mathcal{L}_{\mathrm{h}}$	21.2	29.9	47.6	22.7	31.2	49.2

Table 4: The performance of knowledge selection on the test sets of WoW data. All baselines come from Dinan et al. (2019). The details for all baselines are shown in Appendix A.2.

consistently better than PTKGC$_{\texttt{cat}}$ over all metrics on two data sets, demonstrating that individually representing each knowledge-query-response triple with BERT can lead to a more optimal matching signal than representing a single long sequence. Our explanation to the phenomenon is that there is information loss when a long sequence composed of the knowledge and dialogue history passes through the deep architecture of BERT. Thus, the earlier different knowledge entries and dialogue history are fused together, the more information of dialogue history or background knowledge will be lost in matching. Particularly, on the WoW, in terms of R@1, our PTKGC$_{\texttt{sep}}$ achieves a comparable performance with the existing state-of-the-art models that are learned from the crowd-sourced training set, indicating that the model can effectively learn how to leverage external knowledge feed for response selection through the proposed pre-training approach.

Notably, we can observe that our PTKGC$_{\texttt{sep}}$ performs worse than DIM and FIRE on the CMU_DoG. Our explanation to the phenomenon is that the dialogue and knowledge in CMU_DoG focus on the movie domain while our train data including ad-hoc retrieval corpora and multi-turn dialogues come from the open domain. Thus, our model may not select proper knowledge entries and can not well recognize the semantics clues for response matching due to the domain shift. Despite this, PTKGC$_{\texttt{sep}}$ can still show better performance than several existing models, such as Transformer MemNet and DGMN, though PTKGC$_{\texttt{sep}}$ does not access any training examples in the benchmarks.

Performance of Knowledge Selection. We also assess the ability of models to predict the knowledge selected by human wizards in WoW data. The results are shown in Table 4. We can find that the performance of our method is comparable with various supervised methods trained on the gold knowledge index. In particular, on the test-seen, our model is slightly worse than Transformer (w/ pretrain), while on the test-unseen, our model achieves slightly better results. The results demonstrate the advantages of our pretraining tasks and the good generalization ability of our model.

4.5 Discussions

Ablation Study. We conduct a comprehensive ablation study to investigate the impact of different inputs and different tasks. First, we remove the dialogue history, knowledge, and both of them from the model, which is denoted as PTKGC$_{\texttt{sep}}$(q+k), PTKGC$_{\texttt{sep}}$(q+h) and PTKGC$_{\texttt{sep}}$(q) respectively. According to the results of the first four rows in Table 3, we can find that both the dialogue history and knowledge are crucial for response selection as removing anyone will generally cause a performance drop on the two data. Besides, the background knowledge is more critical for response selection as removing the background knowledge causes more significant performance degradation than removing the dialogue history.

Then, we remove each training task individually from PTKGC$_{\texttt{sep}}$, and denote the models

Models	Wizard Seen			Wizard Unseen		
	R@1	R@2	R@5	R@1	R@2	R@5
PTKGC$_{\mathtt{sep}}$ (q+h)	84.9	93.9	97.8	64.9	81.7	94.3
PTKGC$_{\mathtt{sep}}$ (q+h) -$\mathcal{L}_{\mathrm{h}}$	84.1	93.7	97.7	64.3	81.9	93.8
PTKGC$_{\mathtt{sep}}$ (q+h) -$\mathcal{L}_{\mathrm{p}}$	83.4	93.5	97.9	60.9	80.2	93.5
PTKGC$_{\mathtt{sep}}$ (q+h) -$\mathcal{L}_{\mathrm{h}}$-$\mathcal{L}_{\mathrm{p}}$	83.2	93.8	97.6	60.9	80.1	93.8

Table 5: Ablation study of our model without considering the grounded knowledge.

as PTKGC$_{\mathtt{sep}}$-X, where X $\in \{\mathcal{L}_{\mathrm{p}}, \mathcal{L}_{\mathrm{h}}\}$ meaning query-passage matching task and query-dialogue history matching task respectively. Table 4 shows the ablation results of knowledge selection. We can find that both tasks are useful in the learning of knowledge selection, and query-passage matching plays a dominant role since the performance of knowledge selection drops dramatically when the task is removed from the pre-training process. The last two rows in Table 3 show the ablation results of response selection. We report the ablation results when only 1 knowledge is provided since the knowledge recalls for different ablated models and the full model are very close when m is large ($m = 14$). We can see that both tasks are helpful and the performance of response selection drops more when removing the query-passage matching task. Particularly, $\mathcal{L}_{\mathrm{p}}$ plays a more important role and the performance on test-unseen of WoW drops more obvious when removing each training task.

To further investigate the impact of our pre-training tasks on the performance of the multi-turn response selection (without considering the grounded knowledge), we conduct an ablation study and the results are shown in Table 5. We can observe that the performance of the response matching model (no grounded knowledge) drops obviously when removing one of the pretraining tasks or both tasks. Particularly, the query-passage matching task contributes more to the response selection.

The impact of the number of selected knowledge. We further study how the number of selected knowledge (m) influences the performance of PTKGC$_{\mathtt{sep}}$. Figure 2 shows how the performance of our model changes with respect to different numbers of selected knowledge. We observe that the performance increases monotonically until the knowledge number reaches a certain value, and then stable when the number keeps increasing. The results are rational because more knowledge entries can provide more useful

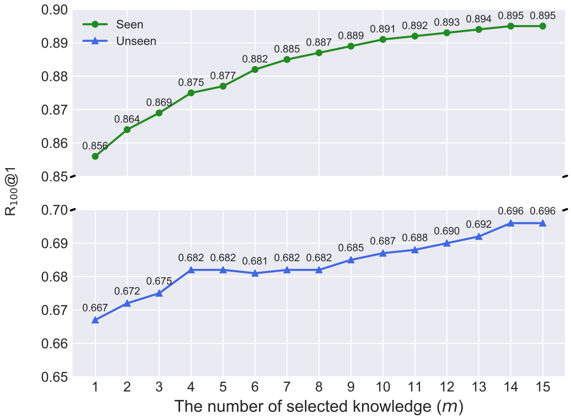

Figure 2: The performance of response selection across different number of selected knowledge.

information for response matching, but when the knowledge becomes enough, the noise will be brought to matching.

5 Conclusion

In this paper, we study response matching in knowledge-grounded conversations under a zero-resource setting. In particular, we propose decomposing the training of the knowledge-grounded response selection into three tasks and joint train all tasks in a unified pre-trained language model. Our model can be learned to select relevant knowledge and distinguish proper response, with the help of ad-hoc retrieval corpora and amount of multi-turn dialogues. Experimental results on two benchmarks indicate that our model achieves a comparable performance with several existing methods trained on crowd-sourced data. In the future, we would like to explore the ability of our proposed method in retrieval-augmented dialogues.

Acknowledgement

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (No. 2020YFB1406702), the National Science Foundation of China (NSFC No. 61876196) and Beijing Outstanding Young Scientist Program (No. BJJWZYJH012019100020098). Rui Yan is the corresponding author, and is supported as a young fellow at Beijing Academy of Artificial Intelligence (BAAI).

References

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

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

Nouha Dziri, Ehsan Kamalloo, Kory W Mathewson, and Osmar R Zaiane. 2018. Augmenting neural response generation with context-aware topical attention. *arXiv preprint arXiv:1811.01063*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *The Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5110–5117.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020a. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20, pages 2041–2044. ACM.

Jia-Chen Gu, Zhen-Hua Ling, Xiaodan Zhu, and Quan Liu. 2019. Dually interactive matching network for personalized response selection in retrieval-based chatbots. 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 1845–1854, Hong Kong, China.

Jia-Chen Gu, Zhenhua Ling, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2020b. Filtering before iteratively referring for knowledge-grounded response selection in retrieval-based chatbots. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1412–1422, Online. Association for Computational Linguistics.

Donna K Harman. 2005. The trec ad hoc experiments.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10, Florence, Italy.

Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

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

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. In *Proceedings of the 34th Conference on Neural Information Processing Systems*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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 *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, volume 16, pages 3776–3784.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1–11.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jesse Vig and Kalai Ramea. 2019. Comparison of transfer-learning approaches for response selection in multi-turn conversations. In *Workshop on DSTC7*.

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945. Association for Computational Linguistics.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2015. Syntax-based deep matching of short texts. In *IJCAI*, pages 1354–1361.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 438–449. Association for Computational Linguistics.

Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and HeuiSeok Lim. 2020. An effective domain adaptive post-training method for bert in response selection. In *Proceedings of INTERSPEECH 2020*, pages 1585–1589.

Ledell Yu Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2018. Starspace: Embed all the things! In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5569–5577.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 496–505. Association for Computational Linguistics.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2020. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 111–120. Association for Computational Linguistics.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Xueliang Zhao, Chongyang Tao, Wei Wu, Can Xu, Dongyan Zhao, and Rui Yan. 2019. A document-grounded matching network for response selection in retrieval-based chatbots. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5443–5449.

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

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

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018b. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127. Association for Computational Linguistics.

A Appendices

A.1 Details of Test Sets

Statistics	Wizard of Wikipedia		CMU_DoG
	Test Seen	Test Unseen	Test
Avg. # turns	9.0	9.1	12.4
Avg, # words per turn	16.4	16.1	18.1
Avg. # knowledge entries	60.8	61.0	31.8
Avg. # words per knowledge	36.9	37.0	27.0

Table 6: The statistics of test sets of two benchmarks.

We tested our proposed method on the Wizard-of-Wikipedia (WoW) (Dinan et al., 2019) and CMU_DoG (Zhou et al., 2018a). Both datasets contain multi-turn dialogues grounded on a set of background knowledge and are built with crowd-sourcing on Amazon Mechanical Turk.

In the WoW dataset, one of the paired speakers is asked to play the role of a knowledgeable expert with access to the given knowledge collection obtained from Wikipedia, while the other of a curious learner. The dataset consists of 968 complete knowledge-grounded dialogues for testing. *It is worth noting that the golden knowledge index for each turn is available in the dataset.* Response selection is performed at every turn of a complete dialogue, which results in 7512 for testing in total. Following the setting of the original paper, positive responses are true responses from humans and negative ones are randomly sampled. The ratio between positive and negative responses is 1 : 99 in testing sets. Besides, the test set is divided into two subsets: Test Seen and Test Unseen. The former shares 533 common topics with the training set, while the latter contains 58 new topics uncovered by the training or validation set.

The CMU_DoG data contains knowledge-grounded human-human conversations where the underlying knowledge comes from wiki articles and focuses on the movie domain. Similar to Dinan et al. (2019), the dataset was also built in two scenarios. In the first scenario, only one worker can access the provided knowledge collections, and he/she is responsible for introducing the movie to the other worker; while in the second scenario, both workers know the knowledge and they are asked to discuss the content. *Different from WoW, the golden knowledge index for each turn is unknown for both scenarios.* Since the data size for an individual scenario is small, we merge the data of the two scenarios following the setting with Zhao et al. (2019). Finally, there are 537 dialogues for testing. We evaluate the performance of the response selection at every turn of a dialogue, which results in 6637 samples for testing. We adopted the version shared in Zhao et al. (2019), where 19 negative candidates were randomly sampled for each utterance from the same set. More details about the two benchmarks can be seen in Table 6.

A.2 Baselines for Knowledge Selection

To compare the performance of knowledge selection, we choose the following baselines from Dinan et al. (2019) including (1) Random: the model randomly selects a knowledge entry from a set of knowledge entries; (2) IR Baseline: the model uses simple word overlap between the dialogue context and the knowledge entry to select the relevant knowledge; (3) BoW MemNet: the model is based on memory network where each memory item is a bag-of-words representation of a knowledge entry, and the gold knowledge labels for each turn are used to train the model; (4) Transformer: the model trains a context-knowledge matching network based on Transformer architecture; (5) Transformer (w/ pretrain): the model is similar to the former model, but the transformer is pre-trained on Reddit data and fine-tuned for the knowledge selection task.

A.3 Results of Low-Resource Setting

Ration (t)	Wizard Seen			Wizard Unseen		
	R@1	R@2	R@5	R@1	R@2	R@5
0%	89.5	96.7	98.9	69.6	85.8	96.3
10%	90.8	97.1	99.4	73.2	86.9	96.8
50%	91.5	97.1	99.3	73.9	87.9	96.9
100%	92.2	97.6	99.4	74.3	88.1	97.1

Table 7: Evaluation results of our model in the low-resource setting on the Wizard of Wikipedia data.

As an additional experiment, we also evaluate the proposed model for a low-resource setting. We randomly sample $t \in \{10\%, 50\%, 100\%\}$ portion of training data from WoW, and use the data to fine-tune our model. The results are shown in Table 7. We can find that with only 10% training data, our model can significantly outperform existing models, indicating the advantages of our pre-training tasks. With 100% training data, our model can achieve 2.7% improvement in terms of R@1 on the test-seen and 4.7% improvement on the test-unseen.