Reciprocal Learning of Knowledge Retriever and Response Ranker for Knowledge-Grounded Conversations

Jiazhan Feng[1,4], **Chongyang Tao**[1], **Zhen Li**[1,4], **Chang Liu**[1,2],
Tao Shen[3] and **Dongyan Zhao**[1,2,5*]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Center for Data Science, Peking University [3]University of Technology Sydney
[4]School of Artificial Intelligence, Peking University
[5]State Key Laboratory of Media Convergence Production Technology and Systems
{fengjiazhan,chongyangtao,lizhen63,liuchang97,zhaody}@pku.edu.cn
tao.shen@uts.edu.au

Abstract

Grounding dialogue agents with knowledge documents has sparked increased attention in both academia and industry. Recently, a growing body of work is trying to build retrieval-based knowledge-grounded dialogue systems. While promising, these approaches require collecting pairs of dialogue context and the corresponding ground-truth knowledge sentences that contain the information regarding the dialogue context. Unfortunately, hand-labeling data to that end is time-consuming, and many datasets and applications lack such knowledge annotations. In this paper, we propose a reciprocal learning approach to jointly optimize a knowledge retriever and a response ranker for knowledge-grounded response retrieval without ground-truth knowledge labels. Specifically, the knowledge retriever uses the feedback from the response ranker as pseudo supervised signals of knowledge retrieval for updating its parameters, while the response ranker also receives the top-ranked knowledge sentences from knowledge retriever for optimization. Evaluation results on two public benchmarks show that our model can significantly outperform previous state-of-the-art methods.

1 Introduction

Human-machine communication is one of the ultimate goals of artificial intelligence. Recently, building a dialogue system with intelligence has sparked increased attention in both academia and industry. Advanced work includes retrieval-based methods (Zhou et al., 2018b; Tao et al., 2019; Han et al., 2021) and generation-based methods (Li et al., 2016; Serban et al., 2016; Zhang et al., 2020). In this paper, we focus on the retrieval-based approaches since they are superior in providing informative and fluent responses to a human input by selecting a proper response from a pre-built index. However, such models are still limited in their ability to fully understand the human query and predict a more engaging response. To this end, some researchers have begun to ground dialogue agents with knowledge (Dinan et al., 2019; Gopalakrishnan et al., 2019; Gunasekara et al., 2019) since humans can naturally associate the content of the conversation with the background knowledge in his/her mind, which has led to improved performance.

Two prominent lines of research have evolved for this task. One is to build retrieval-based knowledge-grounded dialogue models by directly attending to all available knowledge entries (Gu et al., 2019; Zhao et al., 2019; Gu et al., 2020; Hua et al., 2020). The other is to separate the knowledge-grounded response retrieval process into two stages: knowledge retrieving and response ranking (Dinan et al., 2019; Gopalakrishnan et al., 2019; Tao et al., 2021), in which a knowledge retriever first selects relevant knowledge sentences from grounded documents, and then a response ranker incorporates the retrieved knowledge sentences from the knowledge retriever and ranks the candidate responses regarding the dialogue context. However, a long-standing issue on this task is that it is non-trivial to collect large-scale dialogues that are naturally grounded on a small set of knowledge sentences. To train such models, one should first collect pairs of dialogue context and the corresponding list of knowledge sentences that contains the information corresponding to the dialogue context. Unfortunately, hand-labeling data to that end is time-consuming, and many data sets and applications lack such knowledge annotations[1]. Therefore,

[*] Corresponding author: Dongyan Zhao.

[1] While some data sets, e.g., Wizard of Wikipedia (Dinan et al., 2019), have ground-truth knowledge labels, many other data sets do not, e.g., CMU_DoG (Zhou et al., 2018a).

Proceedings of the 29th International Conference on Computational Linguistics, pages 389–399
October 12–17, 2022.

the above two research lines both suffer from insufficient knowledge supervision. The former is prone to be affected by noise from irrelevant and redundant knowledge when conducting response retrieval, and the knowledge retrieving process of the latter suffers from the lack of labels indicating the ground-truth knowledge sentences. Hence, the challenge we consider is: *How to better optimize the knowledge retriever and response ranker jointly without ground-truth knowledge labels?*

To address the challenge, we follow the two-stage paradigm and propose a **Rec**iprocal learning approach to jointly optimize knowledge retriever and response ranker for response retrieval in **K**nowledge-**G**rounded **C**onversations. We name our model as RECKGC. In reciprocal learning, the knowledge retriever uses the feedback from the response ranker as pseudo supervised signals of knowledge retrieval for updating its parameters, while the response ranker also receives the top-ranked knowledge sentences from the knowledge retriever to optimize itself. We use the posterior estimate to train the knowledge retriever, and use the prior information to train the response ranker. By this means, the knowledge retriever and response ranker can be jointly optimized without ground-truth knowledge labels.

We conduct experiments on two public benchmarks including Wizard of Wikipedia (Dinan et al., 2019) and CMU_DoG (Zhou et al., 2018a). Evaluation results indicate that our model can significantly outperform the existing methods, and achieve new state-of-the-art performance on both data sets. Our contributions in this paper are two-fold: (1) proposal of a reciprocal learning of knowledge retriever and response ranker for knowledge-grounded response retrieval without ground-truth knowledge label; (2) Empirical verification of the effectiveness of the proposed learning approach on two public benchmarks.

2 Related Work

Early studies of retrieval-based dialogue systems focused on building single-turn context-response matching models that consider only a single utterance or several utterances in the context that are concatenated into a long sequence for response selection (Wang et al., 2013, 2015). Recently, more emphasis has been placed on response retrieval with multi-turn dialogue context and lots of impressive results have been obtained, including 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), the multi-hop selector network (MSN) (Yuan et al., 2019). With advances in pre-trained language models (Devlin et al., 2019a; Liu et al., 2019), some researchers also attempt to apply them on response selection: to represent each utterance-response pair with BERT and fuse these representations to compute the context-response matching score (Vig and Ramea, 2019); to directly treat the context as a long sequence and conduct context-response matching with BERT (Whang et al., 2020); to leverage fine-grained post-training for improving retrieval-based dialogue systems (Han et al., 2021).

Inspired by the ability of human beings to associate dialogue content with background knowledge in his/her mind, researchers have begun to ground dialogue agents with knowledge. Zhang et al. (2018) collect a persona-based dialogue corpus which utilizes the interlocutor's profile as background knowledge; Zhou et al. (2018a) publish a corpus which contains conversations grounded in articles about popular movies; Dinan et al. (2019) release another corpus with Wiki articles as grounded documents which cover a wide range of topics. At the same time, lots of representative models have been obtained. Zhao et al. (2019); Gu et al. (2019); Hua et al. (2020) successively put forward document-grounded matching network (DGMN), dually interactive matching network (DIM), and RSM-DCK which let the dialogue context and all knowledge sentences interact with candidate responses respectively with cross-attention mechanism. Gu et al. (2020) propose a document-grounded model named FIRE which first compute the importance score for each context turn and knowledge sentence, then further use them to weigh the corresponding representation. Dinan et al. (2019) also propose to joint learn the knowledge selection and response matching in a multi-task manner or a two-stage training procedure. This strategy, however, requires ground-truth knowledge labels annotated by human wizards, which is presumed absent in our paper. Recently, Tao et al. (2021) study response matching in knowledge-grounded conversations under a zero-resource setting. In particular, they propose decomposing the training of the knowledge-grounded response selection into three tasks and jointly training all tasks in a unified pre-trained language model.

3 Methodology: RecKGC

In this section, we first formalize the task of knowledge-grounded response retrieval and then introduce our model from overview to several components to reciprocal learning of them.

3.1 Problem Formalization

Suppose that we have a knowledge-grounded dialogue data set $\mathcal{D} = \{C_i, K_i, r_i, y_i\}_{i=1}^{N}$, where C_i is a dialogue context that is the concatenated token sequence of multi-turn utterances, $K_i = \{k_1, k_2, \dots, k_{n_k}\}$ is a collection of background knowledge for conversation with k_j the j-th knowledge sentence and n_k is the number of knowledge sentences; r_i is a candidate response; $y_i = 1$ indicates that r_i is a proper response for C_i and K_i, otherwise, $y_i = 0$. The goal is to learn a matching model $g(C, K, r)$ from $\mathcal{D}$, and thus for any new context-knowledge-response triple (C, K, r), $g(C, K, r)$ returns the matching degree between r and (C, K). Finally, given a series of candidate responses regarding the same (C, K), one can collect the matching scores and conduct response ranking.

3.2 Model Overview

Our model is composed of two modules, the knowledge retriever and the response ranker. Given an input dialogue context and a collection of background knowledge sentences, these modules are used in a two-step process to predict a response. First, the *knowledge retriever* selects a small subset of knowledge sentences from the knowledge collection where some of them contain relevant information regarding the dialogue context. Then these extracted knowledge sentences are processed by the *response ranker*, along with the dialogue context, to thoroughly examine the selected knowledge and contexts, and predict the matching degree of a candidate response. Figure 1 shows an illustration of our model and reciprocal learning procedure. For the knowledge retriever, we use a dual-encoder architecture (Bromley et al., 1993), which is efficient for processing potential massive of knowledge sentences. For the response ranker, we leverage the standard transformer architecture, which performs full attention over the inputs and gives considerable natural language understanding performance. Both of the modules can be initialized from pre-trained language models such as BERT (Devlin et al., 2019b) or RoBERTa (Liu et al., 2019).

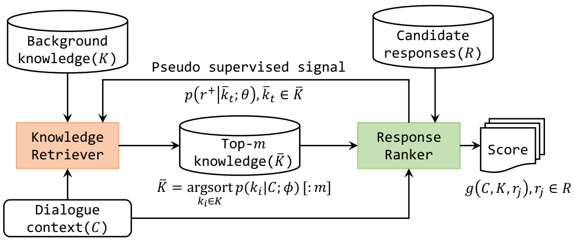
Background knowledge(K)
Candidate responses(R)
Pseudo supervised signal
$p(r^+|\bar{k}_t;\theta), \bar{k}_t \in \bar{K}$
Knowledge Retriever
Top-m knowledge($\bar{K}$)
Response Ranker
Score
$\bar{K} = \operatorname{argsort}_{k_i \in K} p(k_i|C;\phi)\,[:m]$
$g(C, K, r_j), r_j \in R$
Dialogue context(C)

Figure 1: The illustration of our proposed model and reciprocal learning procedure.

The focus of this work is to train the knowledge retriever without ground-truth knowledge labels and conduct the reciprocal learning of knowledge retriever and response ranker in an end-to-end setting. We discuss each component and our training objective in detail below.

3.3 Knowledge Retriever

Given a dialogue context C and a collection of background knowledge $K = \{k_1, k_2, \dots, k_{n_k}\}$, we propose a knowledge retriever is to select a relevant subset of knowledge sentences for the context. For this purpose, the retriever performs a ranking of the knowledge sentences conditioned on the dialogue context and outputs the top-ranked knowledge sentences.

Following Dinan et al. (2019), we leverage a knowledge retriever model composed of an embedder function $E_{retr}(\cdot)$ that maps any knowledge sentence $k_i \in K$ or dialogue context C to a d-dimensional vector, such that the similarity score between dialogue context C and a knowledge sentence k_i can be defined as a scaled dot product of their representation vectors:

$$s(C, k_i; \phi) = \frac{E_{retr}(C)^T E_{retr}(k_i)}{\sqrt{d}} \qquad (1)$$

where $\sqrt{d}$ is a relevance score scaling referring to Sachan et al. (2021), and the retriever is parameterized by ϕ. Although in principle the embedder function $E_{retr}(\cdot)$ can be implemented by any neural networks, in this work we use BERT-Small (Turc et al., 2019), which is a smaller BERT (28M) compared to BERT-base (110M) to take advantage of pre-training while decreasing the number of network parameters. We take the representation at the `[CLS]` token as the output, thus $d = 512$. Differently from the traditional dual-encoder, we use the same encoding function $E_{retr}(\cdot)$ for the context and knowledge sentence by sharing parameters.

The probability of a knowledge sentence k_i being relevant to the context C is defined as:

$$p(k_i|C;\phi) = \frac{exp(s(C,k_i;\phi))}{\sum_{t=1}^{n_k} exp(s(C,k_t;\phi))} \qquad (2)$$

where $k_i \in K$. Through Eq. 2, we can obtain the top-m knowledge sentences with the highest individual score as $\bar{K} = \{\bar{k}_1, \bar{k}_2, \ldots, \bar{k}_m\}$.

3.4 Response Ranker

Besides the knowledge retriever, our model consists of a response ranker that outputs the matching degree of a candidate response r_j based on retrieved knowledge sentences $\bar{K}$ and dialogue context C. We consider fine-tuning the existing PLMs, which is BERT-base (110M) in our paper, to obtain more competent dialogue modeling performance. Concisely, we first concatenate retrieved knowledge sentences $\bar{K}$, dialogue context C and candidate response r_j as a consecutive token sequence with special tokens separating them as,

$$\begin{aligned} x_j = \{ & \texttt{[CLS]}, \bar{k}_1, \texttt{[SEP]}, \ldots, \bar{k}_m, \texttt{[SEP]}, \\ & C, \texttt{[SEP]}, r_j, \texttt{[SEP]} \} \end{aligned} \qquad (3)$$

Then token, position and segment embeddings of each word of x_j are summated and fed into another embedder function $E_{rank}(\cdot)$ (i.e. BERT-base). Finally, we achieve the contextualized embedding $E_{rank}(x_j)$ and feed it into a multi-layer perceptron (MLP) to obtain the final matching degree of a candidate response r_j as:

$$h(C,\bar{K},r_j;\theta) = W_2 \cdot f(W_1 \cdot E_{rank}(x_j) + b_1) + b_2 \qquad (4)$$

where W_1, W_2, b_1, b_2 are learnt parameters, $f(\cdot)$ is a `tanh` activation function, and the ranker is parameterized by θ. The probability of a candidate response r_j being proper to the context C and retrieved knowledge sentences $\bar{K}$ is calculated as

$$p(r_j|C,\bar{K};\theta) = \frac{exp(h(C,\bar{K},r_j;\theta))}{\sum_{t=1}^{n_r} exp(h(C,\bar{K},r_t;\theta))} \qquad (5)$$

where n_r is the number of candidate responses regarding the C and $\bar{K}$. We denote the collection of candidate responses for the context C and background knowledge K as R which contains a ground-truth candidate response r^+, hence the size of R is n_r and $r_j \in R$. Now we can obtain the top candidate response with the highest probability as the output of knowledge-grounded dialogue system from Eq. 5.

3.5 Reciprocal Learning of Knowledge Retriever and Response Ranker

Contrary to previous work on knowledge-grounded response retrieval, we propose a reciprocal learning approach to jointly optimize the knowledge retriever and the response ranker in an end-to-end differentiable fashion. While in this paper, we assume that there are no ground-truth labels for extracting relevant knowledge, which is practical but makes the problem even more challenging.

In reciprocal learning, the trainable components consist of the knowledge retriever (ϕ) and response ranker (θ) parameters. For the training objective of the overall model, we propose to find ϕ and θ that would maximize the likelihood of a ground-truth response r^+ as:

$$p(r^+|C,K;\phi,\theta) = \sum_{\bar{K}\subset K} p(r^+|C,\bar{K};\theta)p(\bar{K}|C;\phi) \qquad (6)$$

However, marginalizing over all possible values of $\bar{K}$, which is a subset of retrieved knowledge sentences, is intractable as it is essentially combinatorial. The log-likelihood of a particular value of $\bar{K}$ is:

$$\begin{aligned} & \log p(r^+|C,\bar{K};\theta)p(\bar{K}|C;\phi) = \\ & \quad \underbrace{\log p(r^+|C,\bar{K};\theta)}_{\text{response ranker}} + \underbrace{\log p(\bar{K}|C;\phi)}_{\text{knowledge retriever}} \end{aligned} \qquad (7)$$

where the first term is parameterized by response ranker θ and the second term is parameterized by knowledge retriever ϕ. We discuss how to optimize both components in detail below.

Optimization of the Response Ranker (θ) For updating θ, we maximize the first term of Eq. 7. Specifically, we first construct $\bar{K}$ by retrieving the top-m relevant knowledge sentences that have the highest similarity scores from retriever. The similarity score is computed by Eq. 1 based on the current value of knowledge retriever parameters ϕ. Since we already have the ground-truth response r^+, the training objective of the response ranker for each training sample can be defined as the negative log-likelihood loss:

$$\mathcal{L}_\theta = -\log p(r^+|C,\bar{K};\theta) \qquad (8)$$

where the probability of the ground-truth response $p(r^+|C,\bar{K};\theta)$ can be computed by Eq. 5.

Optimization of the Knowledge Retriever (ϕ) For updating ϕ, we maximize the second term of Eq. 7. However, since there are no ground-truth labels for extracting relevant knowledge, we can not simply optimize the knowledge retriever (ϕ) by the negative log-likelihood loss similar to Eq. 8. To solve the problem, we consider incorporating the posterior information to provide additional guidance on obtaining appropriate knowledge during training, and the posterior estimate of the second term is formulated as $\log p(\bar{K}|C, r^+; \theta, \phi)$. Since it is non-trivial to maximize a probability of a set, we instead maximize the sum of the probability of each knowledge sentence $\bar{k}_t$ in the set $\bar{K}$, i.e. $\log \sum_{t=1}^{m} p(\bar{k}_t|C, r^+; \theta, \phi)$. The probability of each knowledge sentence $p(\bar{k}_t|C, r^+; \theta, \phi)$ can be further rewritten using the Bayes Rule:

$$\begin{aligned} p(\bar{k}_t|C, r^+; \theta, \phi) &= \frac{p(r^+|C, \bar{k}_t; \theta) p(\bar{k}_t|C; \phi)}{p(r^+|C; \theta, \phi)} \\ &\propto p(r^+|C, \bar{k}_t; \theta) p(\bar{k}_t|C; \phi) \end{aligned} \tag{9}$$

Here we choose not to normalize with denominator $p(r^+|C; \theta, \phi)$ because computing this quantity would necessitate summing over all n_k knowledge sentences[2]. The response ranker now computes the probability of ground-truth response $p(r^+|C, \bar{k}_t; \theta)$ conditioned on only one knowledge sentence $\bar{k}_t$ with a current value of θ.

In fact, in knowledge-based dialogues, lots of samples may be able to match ground-truth response only based on the dialogue context which contains enough retrieval clues (would be illustrated in the ablation study in Section 4.5). In this case, the contribution from knowledge would be very small and the training time may be increased. Meanwhile, we also consider introducing heuristic similarity unigram F1 (denoted as $\rho(\cdot, \cdot)$) between a retrieved knowledge $\bar{k}_t$ and ground-truth response r^+ as supplementary posterior information for supervising knowledge retriever, as we intuitively believe human responses have a strong correlation with the selected knowledge sentence. To better measure the contribution of $\bar{k}_t$, while reducing distractions of the dialogue context C, we estimate $p(r^+|C, \bar{k}_t; \theta)$ as:

$$p(r^+|\bar{k}_t; \theta) + \rho(\bar{k}_t, r^+) \tag{10}$$

where the probability $p(r^+|\bar{k}_t; \theta)$ can be obtained similar to Eq. 5 but the context C is removed from the input sequence of Eq. 3. In our preliminary experiments, we observe that the introduction of the dialogue context C makes the training of knowledge retriever unstable and degrades the performance.

Algorithm 1: The proposed reciprocal learning approach

```
Input: Training set $\mathcal{D}$, knowledge retriever $\phi$, response ranker $\theta$, learning rate
       $\eta_\phi, \eta_\theta$, number of epochs $N_{ep}$, number of iterations $N_{it}$;
Initialize knowledge retriever $\phi$ and response ranker $\theta$ with BERT-small and
  BERT-base respectively;
for $e = 1, 2, ..., N_{ep}$ do
    Shuffle training set $\mathcal{D}$;
    for $t = 1, 2, ..., N_{it}$ do
        Fetch a batch of training data $\mathcal{B}$;
        Obtain the top-$m$ knowledge sentences $\bar{K}$ with current value of $\phi$ by Eq 2;
        Compute $\mathcal{L}_\theta$ with $\bar{K}$ by Eq. 3, 4, 5 and 8;
        Compute the gradients and update $\theta$:
            $\theta \leftarrow \theta + \eta_\theta \frac{\partial \mathcal{L}_\theta(\mathcal{B})}{\partial \theta}$
        Compute $\mathcal{L}_\phi$ with $\bar{K}$ and current value of $\theta$ by Eq. 9, 10, 11 and 12 ;
        Compute the gradients and update $\phi$:
            $\phi \leftarrow \phi + \eta_\phi \frac{\partial \mathcal{L}_\phi(\mathcal{B})}{\partial \phi}$
Output: $\phi$, $\theta$.
```

Then, we compute $p(\bar{k}_t|C; \phi)$ by

$$p(\bar{k}_i|C; \phi) = \frac{exp(s(C, \bar{k}_i; \phi)/\tau)}{\sum_{t=1}^{m} exp(s(C, \bar{k}_t; \phi)/\tau)} \tag{11}$$

where $\bar{k}_t \in \bar{K}$. Note that there is a slight difference in form between Eq. 11 and Eq. 2 where we do not sum over all knowledge sentences K in the denominator which may be massive in practice. As an alternative, we introduce τ as a temperature hyperparameter assuming that knowledge sentences beyond the top-m contribute very small scores to the approximation. The training objective of knowl-

[2] Nevertheless, we observe that our training method still behaves well in practice.

Statistics	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. # turns	9.0	9.1	9.0	9.1	22.2	21.8	22.0
Avg. # words per turn	16.4	16.4	16.4	16.1	18.6	20.1	18.1
Avg. # knowledge entries	61.2	61.5	60.8	61.0	31.3	30.4	31.8
Avg. # words per knowledge	37.2	37.6	36.9	37.0	27.2	28.2	27.0

Table 1: The statistics of two benchmarks.

edge retriever is to minimize the following loss:

$$\mathcal{L}_\phi = -\log \sum_{t=1}^{m} p(\bar{k}_t | C, r^+; \theta, \phi) \qquad (12)$$

Joint Optimization of Overall Model In the overall model, the knowledge retriever and response ranker are jointly optimized in an end-to-end differentiable way. The training objective is to minimize:

$$\mathcal{L} = \mathcal{L}_\phi + \mathcal{L}_\theta \qquad (13)$$

Intuitively, we train the response ranker using prior knowledge distribution in Eq. 2 since we do not introduce the information from ground-truth response r^+. Consequently, there is no mismatch between training and inference, which is useful when the ground-truth response is not known during inference. While for the knowledge retriever, we introduce additional information from r^+ for posterior estimate to learn from richer training signals rather than relying solely on the prior. Algorithm 1 demonstrates the pseudo code of our proposed reciprocal learning approach.

4 Experiments

To demonstrate the effects of the proposed models, we conduct experiments on two public data sets.

4.1 Benchmarks and Evaluation Metrics

We evaluate the proposed method on two public benchmarks including Wizard of Wikipedia (WoW) (Dinan et al., 2019) and CMU Document Grounded Conversations (CMU_DoG) (Zhou et al., 2018a). The statistics of the two benchmarks are shown in Table 1.

The first benchmark we employ is the Wizard of Wikipedia (WoW) (Dinan et al., 2019). During the conversation collection, one of the paired speakers is asked to play the role of a knowledgeable expert with access to the given knowledge collection, while the other one acts as a curious learner. The test set is divided into two subsets by Dinan et al. (2019): 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. In the validation set or test set, the ratio between positive and negative responses is 1:99. Since the training data set do not contain negative responses, we adopt in-batch negatives consistent with Dinan et al. (2019), where the ground-truth responses of the other batch elements are treated as negative training responses.

The second benchmark we use is CMU_DoG data set published in Zhou et al. (2018a). Amazon Mechanical Turk is used to collect conversations based on certain knowledge documents in this data set. The knowledge topics are all about movies, which provide interlocutors with common topics to discuss in a natural way. Two situations are investigated to compel two paired workers to talk about the given documents. In the first one, only one interlocutor has access to the document, while the other does not. The interlocutor with access to the given knowledge document is instructed to introduce the movie to the other. In the second one, both interlocutors can see the given document and are required to talk about its content. Consistent with previous works, we follow Zhao et al. (2019) and merge data in the two scenarios to form a larger data set considering the small number of conversations in each scenario. The ratio between positive and negative responses is 1:19. For a fair comparison, we use the version of data released by DGMN (Zhao et al., 2019).

Consistent with the widely adopted settings on these two benchmarks, we employ recall n at k (i.e., R@k, where $n = 100$ for WoW and $n = 20$ for CMU_DoG and $k = \{1, 2, 5\}$) as the evaluation metrics of response ranking in Table 2 and Table 3, measuring if the ground-truth response can be ranked in top k positions when there are n response candidates. For evaluating the performance of knowledge retrieving in Table 4, we also use recall n_k at k to measure if the ground-truth knowledge can be ranked in top k positions when there are n_k knowledge sentences on WoW data.

4.2 Baselines

As the characteristics of the two benchmarks are different (e.g. only WoW data provide the ground-truth knowledge labels), we compare the proposed model with the baselines on both data individually.

Baselines on WoW. 1) *IR Baseline* (Dinan et al., 2019) uses word overlap for response selection; 2) *BoW MemNet* (Dinan et al., 2019) is a memory network where knowledge sentences are embedded with bag-of-words representation, and the model jointly learns the knowledge selection and response matching; 3) *Two-stage Transformer* (Dinan et al., 2019) trains two individual Transformers for knowledge selection and response retrieval respectively. The best-performing model on knowledge selection is selected for dialogue retrieval; 4) *Transformer MemNet* (Dinan et al., 2019) is an extension of BoW MemNet, and the dialogue context, knowledge sentences and candidate responses are encoded with Transformer encoder that pre-trained on a large-scale corpus; 5) *PTKGC* (Tao et al., 2021) conduct knowledge-grounded response matching in a zero-resource setting, which decomposes the training of response selection into three tasks and jointly trains all tasks in a unified model. It should be noted that we do not compare with the MNDB model (Zhang et al., 2021), because the authors reconstruct the dataset and only retain 32 knowledge candidates for each dialogue, which make this task easier.

Baselines on CMU_DoG 1) *Starspace* (Wu et al., 2018) match the response using the cosine similarity between a concatenated sequence of dialogue context and knowledge, and the response candidate represented by StarSpace; 2) *BoW MemNet* (Zhang et al., 2018) is a memory network with the BOW representation of knowledge as memory entries; 3) *KV Profile Memory* (Zhang et al., 2018) is a key-value memory network grounded on knowledge profiles; 4) *Transformer* (Mazare et al., 2018) encode all utterances with a pre-trained Transformer similar to BoW MemNet; 5) *DGMN* (Zhao et al., 2019) lets the dialogue context and all knowledge sentences interact with the candidate response respectively through cross-attention; 6) *DIM* (Gu et al., 2019) is similar to DGMN and all utterance are encoded with BiLSTMs; 7) *RSM-DCK* (Hua et al., 2020) obtains query-aware knowledge representation and query-aware context representation for response matching; 8) *FIRE* (Gu et al., 2020) filters the context and knowledge first and then use the filtered context and knowledge to iteratively conduct response matching.

4.3 Technical Details

The response ranker and knowledge retriever are implemented with *transformers* library provided by huggingface[3]. Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is the optimizer and the initial learning rate of knowledge retriever and response ranker are 1e-5 and 3e-5. We choose 32 as the size of mini-batches for training on WoW data and 8 on CMU_DoG. During the training, the maximum lengths of the knowledge sentence, dialogue context, and response candidate are set to 40, 80 and 60 on WoW data respectively, while on CMU_DoG we set the the maximum length of the dialogue context as 200. m is set as 5 on both data. τ is set to 0.2 on Eq. 11. We avoid computing the gradients of response ranker parameters θ during estimating $p(r^+|\bar{k}_t;\theta)$ on Eq. 10. Early stopping on the validation set is adopted as a regularization strategy. The best model is selected based on the validation performance.

4.4 Evaluation Results

Performance of Response Ranking. Table 2 and 3 provide the evaluation results of response selection on WoW and CMU_DoG respectively. Numbers in bold mean that improvement over the best baseline is statistically significant (t-test with p < 0.05). Our proposed RecKGC can significantly outperform state-of-the-art models across all metrics on both data. Besides, it is interesting to find that our model achieves more improvement gain on Test Unseen set than Test Seen compared with baselines. The results may be attributed to our model's superior generalization abilities on dialogues with new topics as compared to the previous work, demonstrating the advantages of our proposed reciprocal learning approach.

[3] https://github.com/huggingface/transformers

Models	Test Seen			Test Unseen		
	R@1	R@2	R@5	R@1	R@2	R@5
IR Baseline (Dinan et al., 2019)	17.8	-	-	14.2	-	-
BoW MemNet (Dinan et al., 2019)	71.3	-	-	33.1	-	-
Two-stage Transformer (Dinan et al., 2019)	84.2	-	-	63.1	-	-
Transformer MemNet (Dinan et al., 2019)	87.4	-	-	69.8	-	-
PTKGC (Tao et al., 2021)	89.5	96.7	98.9	69.6	85.8	96.3
DIM (Gu et al., 2019)	83.1	91.1	95.7	60.3	77.8	92.3
FIRE (Gu et al., 2020)	88.3	95.3	97.7	68.3	84.5	95.1
RecKGC	**92.6**	**97.2**	**99.2**	**76.7**	**88.7**	**96.6**

Table 2: Evaluation results of response selection on the test sets of the Wizard of Wikipedia data. Numbers in bold mean that improvement over the best baseline is statistically significant (t-test, p-value < 0.05).

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 (Mazare et al., 2018)	60.3	74.4	87.4
PTKGC (Tao et al., 2021)	66.1	77.8	88.7
DGMN (Zhao et al., 2019)	65.6	78.3	91.2
DIM (Gu et al., 2019)	78.7	89.0	97.1
RSM-DCK (Hua et al., 2020)	79.3	88.8	96.7
FIRE (Gu et al., 2020)	81.8	90.8	97.4
RecKGC	**84.0**	**92.9**	**98.2**

Table 3: Evaluation results of response selection on the test set of the CMU_DoG data. Numbers in bold mean that improvement over the best baseline is statistically significant (t-test, p-value < 0.05).

Performance of Knowledge Retrieving. Since the WoW data contain the ground-truth knowledge labels, we also assess the performance of knowledge retriever with Recall-based metrics in Table 4. Besides, we design two baselines where knowledge retriever (a dual-encoder) is merely trained with supervised or weakly supervised labels. First, we train it with ground-truth knowledge labels (denoted as *"Dual-Enc (supervised)"*). Then in the weakly supervised scenario, we consider $k_i \in K$ that has the highest $\rho(k_i, r^+)$ in each sample as pseudo ground-truth knowledge (denoted as *"Dual-Enc (weakly supervised)"*). We can find that training with ground-truth knowledge labels brings more improvement to the dual encoder than training with pseudo knowledge labels, indicating that pseudo labels are just a sub-optimal supervised learning signal. Notably, the knowledge retriever trained with our proposed reciprocal learning approach outperforms several supervised or weakly supervised baselines and obtains comparable results with *"Dual-Enc (supervised)"*, which proves the effectiveness of our approach.

4.5 Discussions

Ablation Study. We conduct a comprehensive ablation study to investigate the impact of different inputs, posterior information and learning strategies. Table 5 also provides the ablation results. Firstly, we remove the knowledge from the model, which is denoted as "RecKGC *(w/o. knowledge)"*. This model is degraded into a traditional context-response matching paradigm. We can find that removing the knowledge will lead to a dramatic performance drop, which indicates that knowledge is important in response retrieval. However, this model can still outperform some baselines such as BoW MemNet, which proves that some ground-truth responses can be inferred only from the context. Then, we remove the $p(r^+|\bar{k}_t;\theta)$ and $\rho(\bar{k}_t, r^+)$ on Eq. 10 and denote them as "RecKGC *(w/o. kr)"* and "RecKGC *(w/o. f1)"* respectively. We can easily conclude that both posterior information is useful, as removing either information leads to a certain degree of performance degradation. Finally, to prove the advantages of joint learning, we also propose a two-stage training baseline (denoted as *"Two-stage training"*) where we first train the knowledge retriever with pseudo ground-truth knowledge, and then freeze the parameters of knowledge retriever and train the response ranker conditioned on top-m knowledge sentences provided by knowledge retriever. Our model can consistently outperform the model with two-stage training, which confirms the rationality of our reciprocal learning approach.

The Impact of the Number of Retrieved Knowledge Sentences. Furthermore, we investigate how the number of retrieved knowledge sentences

Models	Test Seen			Test 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	-	-
PTKGC	22.0	31.2	48.8	23.1	32.1	50.7
Dual-Enc (weakly supervised)	22.3	33.1	54.3	21.5	31.6	53.1
Dual-Enc (supervised)	23.1	34.0	55.8	22.4	33.4	53.2
RecKGC	22.8	33.6	55.7	23.2	32.9	53.7

Table 4: The performance of knowledge retriever on the test sets of WoW data.

Models	Test Seen			Test Unseen		
	R@1	R@2	R@5	R@1	R@2	R@5
RecKGC	92.6	97.2	99.2	76.7	88.7	96.6
RecKGC (w/o. knowledge)	88.0	94.5	97.6	70.8	84.8	94.5
RecKGC (w/o. f1)	91.7	96.5	99.0	75.3	89.1	96.3
RecKGC (w/o. kr)	92.0	96.8	99.1	74.7	88.6	96.2
Two-stage training	89.2	95.5	98.5	71.6	86.9	95.9

Table 5: Ablation results on two test sets of WoW data.

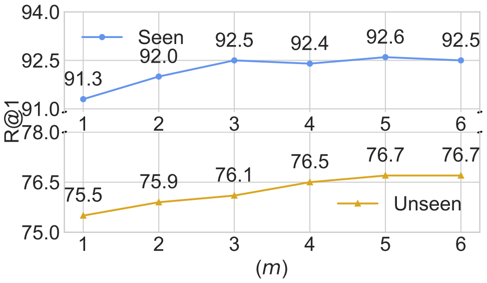

Figure 2: Performance of RecKGC across different number of retrieved knowledge sentences (i.e. m) on Test Seen set and Test Unseen set of WoW data.

(i.e. m) influences the model performance. Figure 2 shows the performance of response selection on test sets of WoW with respect to different m. The curves first monotonically increase until m reaches 5, and then stabilize when m keeps increasing. The reason could be that when only a few knowledge sentences are provided for dialogue, the model cannot capture enough information for response matching, but when the retrieved knowledge becomes sufficient, noise would be introduced into matching because redundant knowledge may be irrelevant to the current dialogue context.

5 Conclusion

In this paper, we study the retrieval-based knowledge-grounded dialogues. To effectively optimize the knowledge retriever and response ranker without ground-truth knowledge labels, we propose a reciprocal learning approach to jointly optimize the two components in an end-to-end way. Concretely, the knowledge retriever takes the feedback from the response ranker as pseudo supervised signals of knowledge retrieval, while the response ranker receives the top-ranked knowledge sentences from the knowledge retriever to optimize itself. By this means, our model can be trained without ground-truth knowledge labels. Evaluation results on two benchmarks indicate that our model can significantly outperform state-of-the-art methods.

Acknowledgement

We would like to thank the anonymous reviewers for their constructive comments and suggestions. This work was supported by National R&D project of China under contract No. 2021YFC3340303. Dongyan Zhao is the corresponding author.

References

Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688.

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

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

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

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang 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. In *INTERSPEECH*, pages 1891–1895.

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.

Jia-Chen Gu, Zhenhua Ling, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2020. Filtering before iteratively referring for knowledge-grounded response selection in retrieval-based chatbots. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1412–1422.

Chulaka Gunasekara, Jonathan K. Kummerfeld, Lazaros Polymenakos, and Walter Lasecki. 2019. DSTC7 task 1: Noetic end-to-end response selection. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 60–67, Florence, Italy. Association for Computational Linguistics.

Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1549–1558, Online. Association for Computational Linguistics.

Kai Hua, Zhiyuan Feng, Chongyang Tao, Rui Yan, and Lu Zhang. 2020. Learning to detect relevant contexts and knowledge for response selection in retrieval-based dialogue systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 525–534.

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B 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.

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*.

Ryan Lowe, Nissan Pow, Iulian Vlad 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.

Pierre-Emmanuel Mazare, 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.

Devendra Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L. Hamilton, and Bryan Catanzaro. 2021. End-to-end training of neural retrievers for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6648–6662, Online. Association for Computational Linguistics.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron 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*, pages 3776–3783.

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

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, Florence, Italy. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

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.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2015. Syntax-based deep matching of short texts. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 1354–1361. AAAI Press.

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 *Proc. Interspeech 2020*.

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*.

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 (Volume 1: Long Papers)*, pages 496–505.

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 (EMNLP-IJCNLP)*, pages 111–120.

Chen Zhang, Hao Wang, Feijun Jiang, and Hongzhi Yin. 2021. Adapting to context-aware knowledge in natural conversation for multi-turn response selection. In *Proceedings of the Web Conference 2021*, pages 1990–2001.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B 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.

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 28th International Joint Conference on Artificial Intelligence*, pages 5443–5449. AAAI Press.

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.

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 (Volume 1: Long Papers)*, pages 1118–1127.