A Three-Stage Learning Framework for Low-Resource Knowledge-Grounded Dialogue Generation

Shilei Liu, Xiaofeng Zhao, Bochao Li, Feiliang Ren*, Longhui Zhang, Shujuan Yin
School of Computer Science and Engineering
Key Laboratory of Medical Image Computing of Ministry of Education
Northeastern University, Shenyang, 110169, China
liusl@live.cn, renfeiliang@cse.neu.edu.cn

Abstract

Neural conversation models have shown great potentials towards generating fluent and informative responses by introducing external background knowledge. Nevertheless, it is laborious to construct such knowledge-grounded dialogues, and existing models usually perform poorly when transfer to new domains with limited training samples. Therefore, building a knowledge-grounded dialogue system under the low-resource setting is a still crucial issue. In this paper, we propose a novel three-stage learning framework based on weakly supervised learning which benefits from large scale ungrounded dialogues and unstructured knowledge base. To better cooperate with this framework, we devise a variant of Transformer with decoupled decoder which facilitates the disentangled learning of response generation and knowledge incorporation. Evaluation results on two benchmarks indicate that our approach can outperform other state-of-the-art methods with less training data, and even in zero-resource scenario, our approach still performs well.

1 Introduction

Neural dialogue systems have made rapid progress in recent years thanks to the advances in sequence generation technology (Vinyals and Le, 2015; Vaswani et al., 2017). Though such models in neural architectures are able to reply with plausible responses regarding to dialogue history, people can still feel a clear gap when they converse with the chatbots, compared with the conversation with humans. To bridge the gap and generate fluent and informative responses, a number of approaches have been proposed by leveraging external knowledge. Knowledge-grounded dialogue is a task of generating an informative response based on both dialogue history and a collection of external knowledge (Dinan et al., 2019). The forms of knowledge are diverse, and in this work, we only focus on knowledge in the form of unstructured documents.

Generally, it is difficult to construct large scale conversations that are naturally grounded on the documents for learning of a response generation model (Zhao et al., 2020a), and most of the previous methods (Lian et al., 2019; Li et al., 2019; Kim et al., 2020; Dinan et al., 2019) perform poorly when transfer into a new domain with limited training samples. So there are growing appeals for low-resource dialogue response generation, which aims to leverage past experience to improve the performance with limited labeled training examples of target corpus.

To address this issue, we envisage to absorb useful information from other easily accessible heterogeneous datasets to enhance the performance of the knowledge-based dialogue model under low-resource setting. Based on this assumption, we propose a novel **Three-Stage Learning Framework (TSLF)**. TSLF attempts to divide the parameters of a model into dialogue-related and knowledge integration-related. In the first stage, we use supervised learning to pre-train dialogue-related parameters on general dialogues (e.g., online forum comments), and perform domain-adaptive pre-training (Gururangan et al., 2020) to initialize knowledge-related parameters on unlabeled knowledge base (e.g., items in Wikipedia). In the second stage, inspired by the distant supervision in the relation extraction (Mintz et al., 2009), we match a set of pseudo-knowledge for each ungrounded dialogue to construct a lower quality knowledge-grounded dialogue dataset, and further co-pretrain the above two groups of parameters on this dataset. In the third stage, the trained model will be fine-tuned on the target low-resource dataset. The flow of TSLF is shown in Figure 1.

In order to better cooperate with the disentangled learning mechanism in TSLF, we devise **Knowledge-Aware Transformer (KAT)**, a vari-

* Corresponding author.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2262–2272
November 7–11, 2021.

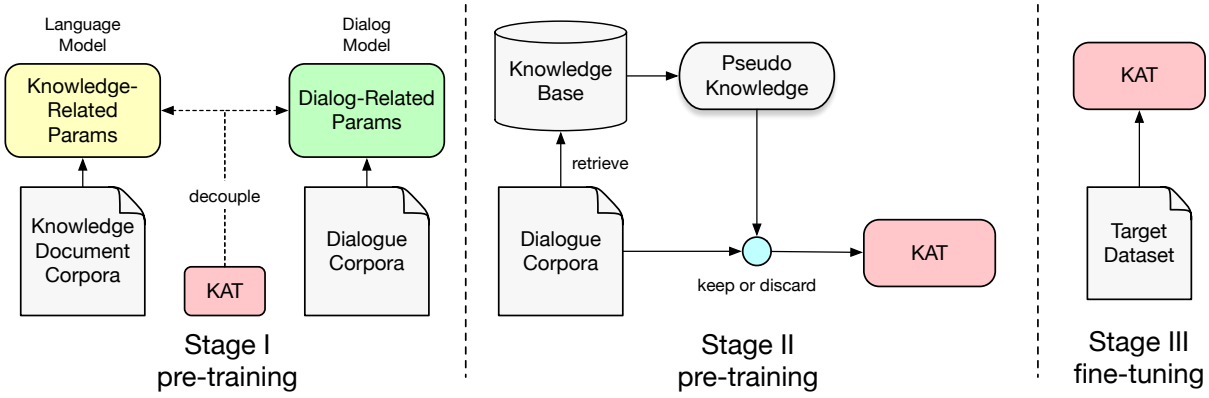

Figure 1: Our three-stage learning framework (TSLF).

ant of vanilla Transformer (Vaswani et al., 2017) whose parameters are decoupled that facilitates the separate learning of dialogue generation and knowledge incorporation. As shown in Figure 2, besides dialogue history, KAT also accepts a set of knowledge as additional input. KAT has a knowledge-aware decoder which could obtains information from the dialogue context and background documents through cross-attention and integrates them through a controller.

We conduct experiments on two knowledge-grounded dialogue generation benchmarks including Wizard-of-Wikipedia (Dinan et al., 2019) and CMU_DoG (Zhou et al., 2018). Evaluation results in terms of both automatic metrics and human judgment indicate that using only about 1/4 of the training data on Wizard (1/16 on CMU_DoG), the performance of our approach outperforms the competitive baselines which are learned from full crowd-sourced training corpora. Even without using any training data of the target dataset, our method still performs well.

The contributions in this work are summarized as follows: (1) We propose a novel three-stage learning framework that leverages weakly supervised learning to help build a low-resource knowledge-grounded dialogue generation model; (2) We devise knowledge-aware Transformer, a knowledge-grounded neural conversation model with a novel dynamic knowledge selection mechanism, which can fully exploits the external knowledge to generate fluent and informative dialogue responses; (3) Our KAT-TSLF achieves surprising performance under the scenarios of full data, low-resource and even zero-resource.

The source code is available at https://github.com/neukg/KAT-TSLF.

2 Approach

Low-resource knowledge-grounded dialogue generation is task that requires a method to learn from experience E, which consists of direct experience E_d containing limited monolingual context-knowledge-response triples and indirect experience E_i, to improve the performance in response generation measured by the evaluation metric P. The direct experience E_d refers to the training samples of target corpus $\mathcal{D}_l = \{(U_i, \mathcal{K}_i, Y_i)\}_{i=1}^{m_1}$ (U_i is dialog history, Y_i is response, and $\mathcal{K}_i = \{K_j\}_{j=1}^{s}$ is a set of external knowledge documents of i-th sample) which are under low-resource settings. In this work, we consider E_i as a large scale ungrounded dialogue dataset $\mathcal{D}_d = \{(U_i, Y_i)\}_{i=1}^{m_2}$, a knowledge base $\mathcal{D}_k = \{K_i\}_{i=1}^{m_3}$ ($m_2, m_3 \gg m_1$) and a pre-trained language model which are easy to obtain. In the following, we first introduce our KAT, and then show how to train it from coarse to fine under our TSLF.

2.1 Knowledge-Aware Transformer

KAT accepts U and $\mathcal{K} = \{K_i\}_{i=1}^{s}$ as inputs, and generates a response $\hat{Y}$. It consists of three components: a dialogue context encoder (DE) to encode U, a knowledge encoder (KE) to encode $\mathcal{K}$, and a decoder to incorporate dialog history, dynamically select knowledge and generate response. The architecture of KAT is shown in Figure 2.

2.1.1 Encoder

We define DE as a Transformer encoder, and the output is represented as $\mathbf{U} \in \mathbb{R}^{n \times d}$, where n is the sequence length, and d is the hidden state dimension. Similarly, KE is defined as another Transformer encoder, and it encode each document individually. Following KE is a concatenation opera-

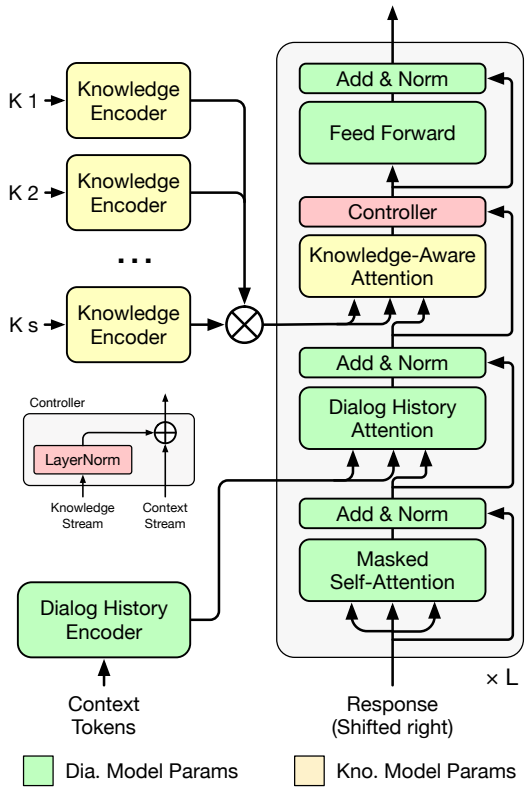

Figure 2: The architecture of our KAT.

tion that concatenates all document representations: $\mathbf{K} = [\mathbf{K}_1; ...; \mathbf{K}_s] \in \mathbb{R}^{sz \times d}$, where $\mathbf{K}_i \in \mathbb{R}^{z \times d}$ is output of i-th KE, and z is the sequence length of each document. $\mathbf{K}$ and $\mathbf{U}$ will be used for the input of the decoder.

2.1.2 Knowledge-Aware Decoder

Generally, not all knowledge in the $\mathcal{K}$ contributes to the generation of the response, so the model should have the ability to select knowledge. Different from (Dinan et al., 2019; Lian et al., 2019; Kim et al., 2020) who perform knowledge selection in the encoding phase (or in a pipeline), we leaves it to the decoding phase. Based on the Transformer decoder, we propose a cross attention based decoder which can select knowledge dynamically and generate informative response.

Knowledge Integration Block (KIB) As shown in the right part of Figure 2, we add a new block after the dialogue history attention block in Transformer decoder layer. It takes the output from last block as *query*, and the memory from $\mathbf{K}$ as *key* and *value*. The output of this block can be obtained by multi-head attention mechanism (Vaswani et al., 2017). During decoding, KIB can dynamically select different knowledge according to dialogue context and the tokens that have been generated at current time step.

Controller To control the knowledge and context contributions in each layer, we add a gate after the knowledge selection block. Denote $\mathbf{h}_k$ as output of KIB and $\mathbf{h}_c$ as the residual from the previous block, the output of controller can be expressed by

$$\begin{aligned} \mathrm{CT}(\mathbf{h}_k, \mathbf{h}_c) &= \beta \cdot \mathrm{LN}(\mathbf{h}_k) + (1 - \beta) \cdot \mathbf{h}_c \\ \beta &= \sigma\left(\mathbf{w} \cdot [\mathbf{h}_k; \mathbf{h}_c]\right) \end{aligned} \quad (1)$$

where $\mathbf{w} \in \mathbb{R}^{2d}$ is a learnable parameter and σ denotes sigmoid function.

2.2 Three-Stage Learning Framework

For further discussion, we denote θ_d, θ_k, and θ_a as the learnable parameters of the green, yellow and pink parts in Figure 2 respectively. We can observe that θ_d is related to context encoding and response generation, θ_k is related to knowledge representation and integration, and these two parts are disentangled. In order to benefit from a wealth of heterogeneous corpora, we propose a three-stage learning framework. In TSLF, we first initialize θ_d and θ_k in a decoupled scheme by training in ungrounded dialogues and unstructured knowledge documents respectively, and then co-optimize them with θ_a by weakly supervised learning and finally transfer KAT to target low-resource dataset. The illustration of TSLF is shown in Figure 1.

2.2.1 Stage I

We choose the state-of-the-art Transformer based encoder-decoder model BART (Lewis et al., 2020) as the the backbone, pre-training it on $\mathcal{D}_d$ with dialogue response generation task:

$$\mathcal{L}_d(\theta_d) = -\sum_{(U,Y)\in\mathcal{D}_d} \sum_t \log p(y_t | y_{<t}, U) \quad (2)$$

Besides, inspired by Gururangan et al. (2020), we conduct domain-adaptive pre-training on unlabeled knowledge documents to improve knowledge representation ability. Specifically, 15% of tokens in a text K are replaced with `<mask>` or noise words, and another Transformer tries to rebuild it:

$$\mathcal{L}_k(\theta_k^+) = -\sum_{K\in\mathcal{D}_k} \sum_t \log p(k_t | k_{<t}, \hat{K}) \quad (3)$$

where $\hat{K}$ is the corrupt K. We disentangle the encoder and the cross-attention block in each decoder layer from this Transformer (θ_k^+) and initialize θ_k with them.

Algorithm 1 Construction of $\mathcal{D}_p$

Input: Ungrounded dialogues $\mathcal{D}_d$, documents $\mathcal{D}_k$, threshold γ and number of negative samples o;
Output: $\mathcal{D}_p$;

```
Initialize D_p = φ;
for (U, Y) in D_d do
    K, score = I(Y, D_k);
    if score > γ then
        K = {K};
        for i in {1, ..., o} do
            Sample K' from D_k − K randomly;
            K ← K ∩ {K'};
        end for
        D_p ← D_p ∩ {(U, K, Y)};
    end if
end for
return D_p;
```

2.2.2 Stage II

In stage I, θ_d and θ_k are trained separately, and the connection between knowledge and dialogue has not yet been established. If KAT is fine-tuned directly on low-resource dataset $\mathcal{D}_k$, it may cause inconsistency problems, so we add a warm-up process to it.

Intuitively, responses from humans carry clues to relevance of the knowledge candidates (Zhao et al., 2020b), so the knowledge document that promotes the flow of dialogue usually has a high textual similarity with the response. Based on this assumption, we construct a set of pseudo-knowledge for some dialogues in $\mathcal{D}_d$ to form a new weak supervision dataset $\mathcal{D}_p$ according to Algorithm 1.

$\mathcal{I}(query, documents)$ means retrieve the document with the highest similarity (e.g., TF-IDF and BM25). Context-response pairs with low quality will be removed. In the knowledge-grounded dialogue corpora, only less documents in knowledge pool are valuable, and others are noise. The design of negative samples is to simulate this situation and make the distribution of knowledge in $\mathcal{D}_p$ closer to the target data set.

We perform weakly supervised learning on $\mathcal{D}_p$ to warmup KAT:

$$\mathcal{L}(\theta_d, \theta_k, \theta_a) = -\sum_{(U,\mathcal{K},Y)\in\mathcal{D}_p} \log p(Y|\mathcal{K}, U) \quad (4)$$

2.2.3 Stage III

After warming up on $\mathcal{D}_p$, KAT will be fine-tuned on the target low-resource dataset:

$$\mathcal{L}(\theta_d, \theta_k, \theta_a) = -\sum_{(U,\mathcal{K},Y)\in\mathcal{D}_l} \log p(Y|\mathcal{K}, U) \quad (5)$$

If not fine-tuned, KAT can also be directly applied to zero-resource response generation.

3 Experiments

3.1 Datasets and Evaluation Methods

We conduct extensive experiments on two public English knowledge-grounded datasets: Wizard-of-Wikipedia (Dinan et al., 2019) and CMU_DoG (Zhou et al., 2018). Wizard-of-Wikipedia is a chit-chatting dataset between two agents, and the two participants are not quite symmetric: one will play the role of a knowledgeable expert (which we refer to as the wizard) while the other is a curious learner (the apprentice). Each wizard turn is associated with ∼60 sentences retrieved from the Wikipedia and each sentence contains ∼30 words, and most of them are noise. The test set is split into two subsets, test seen and test unseen. The difference between the two is that the former contains some topics that overlap with the training set. CMU_DoG also contains conversations between two workers who know the background documents and try to discuss the content in depth. Different from Wizard-of-Wikipedia which spans multiple topics, CMU_DoG mainly focuses on film reviews.

Reddit Conversation Corpus is a large scale open domain dialogue corpus cleaned by Dziri et al. (2018) which consists of ∼15M samples for training and ∼0.8M samples for validation. Following Zhao et al. (2020a); Li et al. (2020), we merge the training and validation data of RedditCC as $\mathcal{D}_d$. Besides, we split ∼0.5M Wikipedia articles provided by ParlAI(Miller et al., 2017) into ∼6.6M sentences as $\mathcal{D}_k$. Information retrieval function $\mathcal{I}$ mentioned in Sec. 2.2.2 is implemented by Apache Lucene with BM25 algorithm and the size of $\mathcal{D}_p$ is ∼0.1M. γ and o are set to 16.4 and 39 respectively.

Following the common practice in evaluating open domain dialogue generation, we choose perplexity (PPL), corpus-level BLEU (Papineni et al., 2002), sentence-level ROUGE (Lin, 2004) and corpus-level DISTINCT (Li et al., 2016) as metrics. Response with higher BLEU and ROUGE is closer to the ground-truth, and response with higher DIST

Models	PPL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	R-1	R-2	DIST-1	DIST-2
ITDD (Li et al., 2019)	17.8	15.8	7.1	4.0	2.5	16.2	-	-	-
BART$_{cat}$	19.7	23.1	11.4	6.7	4.3	19.3	5.1	7.1	29.9
BART$_{skt}$ (Kim et al., 2020)	20.3	23.2	11.9	7.6	4.4	19.4	5.4	6.8	30.3
DRD (Zhao et al., 2020a)	23.0	21.8	11.5	7.5	5.5	18.0	-	-	-
ZRKGC† (Li et al., 2020)	40.4	22.2	7.3	2.8	1.8	18.6	2.4	5.4	22.5
KAT Full Data	14.5	25.5	13.9	9.0	6.6	21.6	7.5	9.3	37.0
KAT-TSLF Full Data	14.4	25.5	13.9	9.1	6.7	21.7	7.6	9.5	38.3
KAT-TSLF 1/4 Data	17.6	23.3	12.2	7.7	5.5	20.3	6.8	9.9	39.1
KAT-TSLF 1/8 Data	18.8	22.5	11.5	7.1	4.9	19.8	6.3	9.9	39.5
KAT-TSLF Zero Data	100+	19.5	8.1	4.0	2.2	14.7	3.0	7.5	33.9

Table 1: Evaluation results on Wizard test seen. † marks zero-resource setting. The results of ITDD and DRD are copied from (Zhao et al., 2020a) and DRD is under full-data. The performance of *KAT-TSLF 1/4 Data* outperforms BART$_{cat}$ and BART$_{skt}$ significantly except BLEU-1 (t-test with p-value < 0.01, the same table below).

Models	PPL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	DIST-1	DIST-2
ITDD	44.8	13.4	4.7	2.1	1.1	11.4	-	-	-
BART$_{cat}$	24.5	23.2	11.0	6.3	4.1	18.9	4.5	5.3	22.2
BART$_{skt}$	22.3	23.4	10.9	6.8	4.6	19.0	4.7	5.2	24.5
DRD	25.6	20.7	10.1	6.2	4.3	16.5	-	-	-
ZRKGC†	41.5	21.8	7.1	2.7	1.1	18.5	2.4	3.4	15.6
KAT	15.8	24.4	12.5	7.8	6.6	20.5	6.4	10.1	39.1
Full Data	15.8	24.1	12.9	8.3	6.0	20.7	7.2	6.7	26.0
1/4 Data	18.4	23.1	11.9	7.5	5.2	19.9	6.4	6.6	25.1
1/8 Data	20.1	22.3	11.3	7.0	4.8	19.0	5.9	6.6	25.3
Zero Data	100+	19.6	8.6	4.7	2.7	14.9	3.0	5.7	26.4

Table 2: Evaluation results on Wizard-of-Wikipedia test unseen.

has a larger vocabulary that could express more information. BLEU is computed with NLTK library (Bird, 2006) and ROUGE is calculated with the code published with Kim et al. (2020).

Besides quantitative evaluation, we also recruit three human annotators to do qualitative analysis on response quality. For each dataset, we randomly sample 100 samples, and each sample contains the conversation history, response, and external knowledge set (for Wizard-of-Wikipedia, we only provide ground-truth knowledge). The annotators then judge the quality of the responses from three aspects, including context coherence, language fluency and knowledge relevance, and assign a score in {0, 1, 2} to each response for each aspect. Each response receives 3 scores per aspect, and the agreement among the annotators is measured via Fleiss' kappa (Fleiss, 1971).

3.2 Baselines

We compare our approach with the following baselines: (1) **ITDD**: an Transformer-based architecture which incrementally represents multi-turn dialogues and knowledge, and conducts response decoding in two passes (Li et al., 2019); (2) **BART**$_{cat}$: A simple BART-based model that take the concatenation of dialogue context and all knowledge as the input of BART for response generation. BART sets constraint on the maximum number of tokens it can handle, and we directly truncate the text that exceeds the length limit; (2) **BART**$_{skt}$: SKT is variational model that introduced BERT on the basis of Lian et al. (2019) and considered the knowledge selection history in multi-turn dialogue (Kim et al., 2020). We feed the knowledge candidate selected by SKT to BART for response generation. It should be noted that training SKT requires human labels that indicate ground-truth knowledge which are crucial to the performance of the model. For fair comparison, we use $\mathcal{I}$ to reselect the knowledge label; (3) **DRD**: Another low-resource dialogue model which devise a disentangled response decoder with copy mechanism (See et al., 2017) and use a two-stage framework to learn it (Zhao et al., 2020a). DRD is not open source, so we can't make a very detailed comparison with it; (4) **ZRKGC**: A double latent variable model that achieves the state-of-the-art performance in zero-resource knowledge-grounded dialogue generation (Li et al., 2020). ZRKGC is based on UNILM

Models	PPL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	DIST-1	DIST-2
ITDD	26.0	9.5	3.6	1.7	0.9	10.4	-	-	-
BART$_{cat}$	36.4	17.0	8.6	5.3	3.4	13.6	3.1	1.5	7.3
BART$_{skt}$	40.1	16.2	8.3	5.1	3.1	12.7	2.6	1.2	7.3
DRD	54.4	15.0	5.7	2.5	1.2	10.7	-	-	-
ZRKGC†	53.5	15.1	4.2	1.2	0.4	12.5	0.7	1.2	8.1
KAT	22.2	19.4	10.5	6.9	4.7	14.4	3.3	1.8	8.9
Full Data	21.7	20.4	10.6	6.7	4.4	15.1	3.7	2.0	11.1
1/8 Data	25.7	19.1	10.1	6.5	4.4	13.9	3.2	1.9	10.5
1/16 Data	28.1	18.5	9.8	6.3	4.2	13.4	2.9	1.8	9.9
Zero Data	100+	12.8	4.7	2.4	1.4	7.9	1.0	2.6	15.7

Table 3: Evaluation results on CMU_DoG. The performance of *KAT-TSLF 1/16 Data* outperforms BART$_{cat}$ and BART$_{skt}$ significantly except ROUGE-1 and ROUGE-2 (t-test with p-value < 0.01).

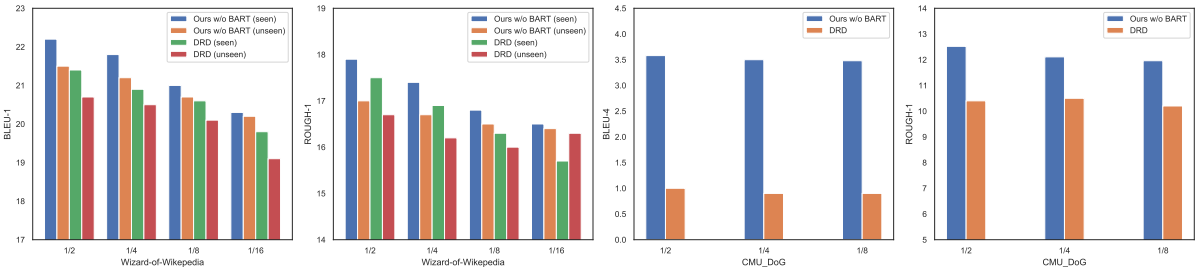

Figure 3: Comparison with DRD in low-resource setting. DRD does not provide results when the training data is less than 1/16 (1/8 in CMU_DoG). In order to save space, we merge the Wizard seen and unseen into one subfigure.

(Dong et al., 2019) with 110M parameters whose performance is close to BART, so we will not replace the backbone of ZRKGC.

3.3 Implementation Details

The knowledge pool of target dataset is usually very large (e.g. ∼60 sentences in Wizard), in order to reduce the time overhead, following (Kim et al., 2020), we only keep the first 40 sentences. We use the base version of BART with 139M parameters in our work, and the number of parameters of KAT is 196M. The batch size in stage I, II and III is 2048, 128 and 16 respectively. The max sequence length in source and target is 256 and 64 respectively. All models are optimized with AdamW (Loshchilov and Hutter, 2017) with learning rate $5e-5$ in 3 epochs. We employ beam search in response decoding (the number of beams from 1 to 3) implemented by Wolf et al. (2020).

3.4 Evaluation Results

Table 1, 2 and 3 reports the evaluation results on automatic metrics, and we have the following observations: (1) In the full-data scenario, KAT achieves state-of-the-art performance without using any additional corpora, which means that KAT itself is an excellent dialogue model. Besides, additional resources are unnecessary when there are enriched training datas, so TSLF has little effect in this setting; (2) KAT-TSLF achieves the comparable performance with BART$_{cat/skt}$ even though the baselines have leveraged all training data, while our model is only learned with 1/4 training data on Wizard (1/16 on CMU_DoG). We compare the low-resource performance with DRD, and the results are shown in Figure 3. For a fair comparison, we removed the pre-training language model and reduce the number of model parameters. We can see that KAT-TSLF outperforms DRD (especially in CMU_DoG). The comparison with BART$_{cat}$ is supplemented in Figure 4; (3) Although our TSLF is mainly for low-resource scenarios, under the setting of zero resources (i.e., without stage III), the performance of KAT-TSLF also surpasses ZRKGC in most evaluation metrics; (4) Responses generated by KAT have higher DIST-n, which means that our KAT can better obtain information from multiple knowledge and generate more diverse texts.

Table 4 reports the human evaluation results. We observe that responses from our KAT-TSLF are more fluent and more contextually coherent than those from BART$_{skt}$ and ZRKGC. Compared with our low-resource model, SKT has stronger knowledge relevance in the case of full data, thanks to its

Models	Wizard Test Seen				Wizard Test Unseen				CMU_DoG		
	CC	LF	KR	Kappa	CC	LF	KR	Kappa	CC	LF	Kappa
BART$_{skt}$	1.78	1.80	1.34	0.61	1.72	1.74	1.36	0.64	1.70	1.72	0.65
ZRKGC	1.72	1.75	1.12	0.63	1.69	1.70	1.16	0.63	1.67	1.69	0.63
Ours 1/8 Data	1.81	1.82	1.35	0.63	1.79	1.78	1.35	0.66	1.74	1.75	0.69
Ours Zero Data	1.76	1.78	1.14	0.64	1.70	1.72	1.24	0.64	1.69	1.71	0.66

Table 4: Human evaluation results on Wizard-of-Wikipedia and CMU_DoG. CC, LF and KR marks *context coherence*, *language fluency* and *knowledge relevance* respectively. In zero-resource setting, our KAT-TSLF outperforms ZRKGC. Besides, our model surpasses BART$_{skt}$ (full data) in most metrics with only only 1/8 of the training data.

well-designed knowledge selection module.

3.5 Ablation Study

We conduct ablation experiments on Wizard and CMU_DoG, and the results are shown in Figure 4.

So as to verify the effect of TSLF, we first removed stage I, stage II, and stage I II respectively. Inserting a new module into an already well-trained large-scale pre-trained language model will cause inconsistency problems, which require a lot of data to reconcile, so after removing stage II or stage I II, the performance of our KAT in low-resource dropped sharply. Although the quality of the automatically constructed warm-up dataset $\mathcal{D}_p$ is lower than the target dataset $\mathcal{D}_l$, it also helps to establish the connection between the knowledge representation component and the dialogue component. Besides, we tried not to pre-train θ_k on unlabeled documents, and the result has dropped slightly, which demonstrates that is still helpful to tailor a pre-trained model to the domain of a target task. In addition, replacing negative sampling with top-k retrieval will increase the inconsistency with the knowledge distribution of target dataset, leading to performance degradation. Moreover, the controller also has an effect on the generalization of the model. It can help KAT quickly adapt to new domains by adjusting the proportion of knowledge and context in the response. In order to improve the generalization performance with limited training data, some works (Chen and Shuai, 2021; Zhao et al., 2020a) fix most of the parameters during fine-tuning. We also tried to frozen knowledge encoder and context encoder in stage III or stage II III, and the results show that the performance has not improved, indicating that with the help of stage II, our model can hardly fall into overfitting.

In order to verify the effect of our TSLF on other models, we try to combine BART$_{cat}$ with TSLF. Since the parameters of BART are tightly coupled, we can only apply stage II to it. Experimental results show that the performance is improved significantly under low-resource setting.

3.6 Discussions

Case Study Table 5 shows a case from Wizard, from which we can see that the response from our model with zero data not only smoothly catches the ground-truth knowledge (highlighted in blue), but also expands the topic with proper pieces of other knowledge (highlighted in yellow). ZRKGC generated sentences that were inconsistent with the facts. Although BART$_{skt}$ chose the correct knowledge, the narrative was too straightforward, and there is a repetition phenomenon. We showed some other cases in the supplementary material.

Comparison with DRD If we ignore the details, DRD is actually a special case of our method, which skips stage II. During pre-training, DRD completely separates dialogue-related components and knowledge representation-related components, which makes it difficult to effectively promote the integration of dialogue and knowledge with only a small number of samples during fine-tuning. So when the training data is extremely small, DRD can hardly work. Besides, in order to prevent overfitting, DRD has to limit the number of parameters of the knowledge integration component and use fix other parameters when fine-tuning, which leads to limited performance of the model. In addition, the complex model structure makes it difficult for DRD to use pre-trained language models.

KAT v.s. BART$_{cat}$ BART (as well as most other pre-training language models) has a limit on the maximum tokens of the input, so useful knowledge is likely to be truncated. For example, there are about 60 external documents per sample in Wizard, and about 40 documents will be truncated. In theory, KAT can accept an unlimited number of knowledge, so this should be one of the reasons why KAT's performance is better than BAER$_{cat}$.

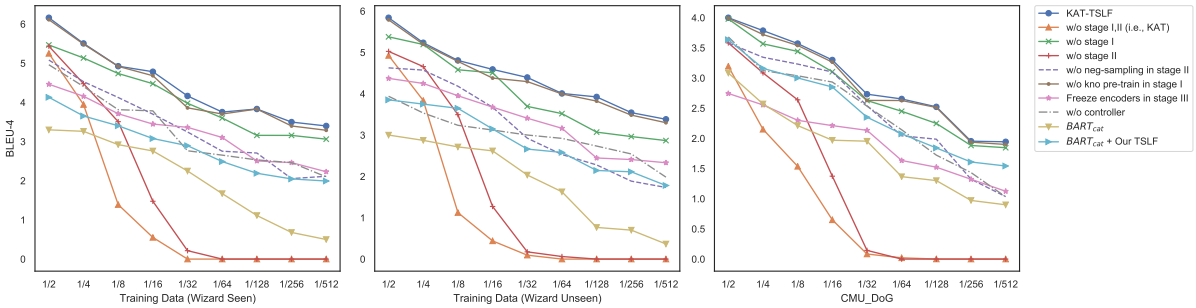

Figure 4: Ablation experiments on Wizard-of-Wikipedia and CMU_DoG. The number of beams are set to 1 for all models. It is recommended to view the picture after zooming in, and the more the curve is to the upper right, the better the result.

When we reduce the maximum number of knowledge that KAT can handle (a hyperparameter) to 15, the performance is close to BART$_{cat}$.

Dial. Hist.	A: Yea it was a great movie. The Last of the Mohicans was released in 1992. B: I didn't realize it's been out that long! What is it about?
GT Kno.	The Last of the Mohicans is a 1992 American epic historical drama, set in 1757 during the French and Indian War.
Ref.	Well The Last of the Mohicans is an epic historical drama. It was set in 1757 during the Indian and French war.
(BART$_{skt}$) It's about the French and Indian War. It's about the French and Indian War. **(ZRKGC)** It 's a classic movie. The Last of My Mohicans was released in 2016, and is still out on Netflix. **(Ours Zero Data)** It's a series of short stories set in 1757 during the French and Indian War in the Adirondack mountains of Virginia . **(Ours 1/16 Data)** It is about a group of people who fight to keep their independence from the French and Indian War.	

Table 5: A case from test seen of Wizard-of-Wikipedia. This dialogue contains a total of 40 external knowledge, one of which is marked as ground-truth (GT).

4 Related Work

Open domain end-to-end dialogue response generation is inspired by the success of applying neural sequence to sequence models on machine translation (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). Very recently, in order to generate fluent, coherent and informative response, many approaches have been proposed by introducing external background documents (Ghazvininejad et al., 2018; Yavuz et al., 2019; Li et al., 2019; Lin et al., 2020). Besides documents (Dinan et al., 2019; Zhou et al., 2018), the are many forms of knowledge such as images (Huber et al., 2018) and triples in knowledge graph (Wu et al., 2019; Tuan et al., 2019).

Dinan et al. (2019) presents to divide knowledge-grounded dialogue into two steps: knowledge selection and dialogue generation. PostKS (Lian et al., 2019), SKT (Kim et al., 2020), PIPM (Chen et al., 2020) and SKT-KG (Zhan et al., 2021) use the prior and posterior distribution of knowledge to improve the accuracy of knowledge selection. Zhao et al. (2020b) devise a reinforcement learning method to train a knowledge selector without ground-truth knowledge label. DeepCopy (Yavuz et al., 2019), ITDD (Li et al., 2019) and KIC (Lin et al., 2020) have improved the structure of the decoder so that it can better integrate knowledge. Since knowledge-guided dialogue corpora need to be constructed through crowdsourcing, the size of datasets such as Wizard-of-Wikipedia (Dinan et al., 2019) are relatively small. Zhao et al. (2020a) and Li et al. (2020) proposed to conduct the knowledge-grounded conversation under the low-resource and zero-resource settings respectively. We do not compare with Lin et al. (2020); Zhao et al. (2020b) since they did not release their entire source codes.

Our three-stage learning framework is inspired by Zhao et al. (2020a), which uses ungrounded dialogues and unstructured documents to train a knowledge-grounded dialogue model that can work in low-resource situations. In addition, the design of stage II is inspired by distant supervision technology in relation extraction task (Mintz et al., 2009). The idea of KAT is also encouraged by disentangled decoder (Raghu et al., 2019) and the recent breakthrough in variants of Transformer (Li et al., 2019; Hashemi et al., 2020; Izacard and Grave, 2020).

5 Conclusion

We study knowledge-grounded dialogue generation under a low-resource setting by proposing a three-stage learning framework and a knowledge-aware Transformer. Evaluation results on two benchmarks indicate that our model achieves the state-of-the-art performance with less training data. Besides, KAT-TSLF exhibits a good generalization ability on zero-resource scenario.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U1708261 and No. 61572120), Shenyang Medical Imaging Processing Engineering Technology Research Center (17-134-8-00), the Fundamental Research Funds for the Central Universities (No. N181602013 and No.N2016006), Ten Thousand Talent Program (No.ZX20200035), and Liaoning Distinguished Professor (No.XLYC1902057).

Broader Impact

Incorporating knowledge into dialogue systems has been the pursuit of researchers in this field for many years. This kind of system will make AI dialogue more natural definitely. It will be more favored by people when the technology does not require a large amount of artificially annotated data. More importantly, the knowledge-based dialogue system can fundamentally change the experience of human-machine dialogue, because system can develop with the update of external knowledge base. One day it will be true that people can obtain effective information through simple conversations. However, coins always have two sides. In addition to the well-known problems caused by large pre-trained datasets for end-to-end dialogue models, special knowledge bases which may be deliberately tailored can also be used to make the generated dialogues biased, just as search engines inadvertently spread biased content created by someone. In order to prevent this technology from being abused, we look forward to more research effort for detecting fake/biased/offensive content. At the same time, we recommend that developers choose content carefully to build a knowledge base for the dialogue system. Good external knowledge can adjust the behavior of the dialogue model in the response process and help the model overcome the biases hidden in large-scale social media datasets.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Steven Bird. 2006. NLTK: the natural language toolkit. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the Gap between Prior and Posterior Knowledge Selection for Knowledge-Grounded Dialogue Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3426–3437.

Yi-Syuan Chen and Hong-Han Shuai. 2021. Meta-transfer learning for low-resource abstractive summarization. *CoRR*, abs/2102.09397.

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*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

Nouha Dziri, Ehsan Kamalloo, Kory Wallace Mathewson, and Osmar R. Zaïane. 2018. Augmenting neural response generation with context-aware topical attention. *CoRR*, abs/1811.01063.

Joseph Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–.

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 *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5110–5117.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2020. Guided transformer: Leveraging multiple external sources for representation learning in conversational search. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1131–1140. ACM.

Bernd Huber, Daniel J. McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, page 277. ACM.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *CoRR*, abs/2007.01282.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-Resource Knowledge-Grounded Dialogue Generation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 12–21. Association for Computational Linguistics.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to Select Knowledge for Response Generation in Dialog Systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5081–5087.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. page 10.

Xiexiong Lin, Weiyu Jian, Jianshan He, Taifeng Wang, and Wei Chu. 2020. Generating Informative Conversational Response using Recurrent Knowledge-Interaction and Knowledge-Copy. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 41–52.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Alexander H. Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017 - System Demonstrations*, pages 79–84. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pages 1003–1011. The Association for Computer Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Dinesh Raghu, Nikhil Gupta, and Mausam. 2019. Disentangling language and knowledge in task-oriented dialogs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN,*

USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 1239–1255. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1855–1865. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3794–3804. Association for Computational Linguistics.

Semih Yavuz, Abhinav Rastogi, Guan-Lin Chao, and Dilek Hakkani-Tür. 2019. DeepCopy: Grounded Response Generation with Hierarchical Pointer Networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pages 122–132.

Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021. Augmenting Knowledge-grounded Conversations with Sequential Knowledge Transition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5621–5630, Online. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. Low-Resource Knowledge-Grounded Dialogue Generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020b. Knowledge-Grounded Dialogue Generation with Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 708–713. Association for Computational Linguistics.