Contextual Knowledge Learning for Dialogue Generation

Wen Zheng[1], Natasa Milic-Frayling[1,2], Ke Zhou[1,3]

[1]University of Nottingham, [2]Qatar Computing Research Institute, [3]Nokia Bell Labs

{wen.zheng, natasa.milic-frayling, ke.zhou}@nottingham.ac.uk

Abstract

Incorporating conversational context and knowledge into dialogue generation models has been essential for improving the quality of the generated responses. The context, comprising utterances from previous dialogue exchanges, is used as a source of content for response generation and as a means of selecting external knowledge. However, to avoid introducing irrelevant content, it is key to enable fine-grained scoring of context and knowledge. In this paper, we present a novel approach to context and knowledge weighting as an integral part of model training. We guide the model training through a Contextual Knowledge Learning (CKL) process which involves Latent Vectors for context and knowledge, respectively. CKL Latent Vectors capture the relationship between context, knowledge, and responses through weak supervision and enable differential weighting of context utterances and knowledge sentences during the training process. Experiments with two standard datasets and human evaluation demonstrate that CKL leads to a significant improvement compared with the performance of six strong baseline models and shows robustness with regard to reduced sizes of training sets.

1 Introduction

Dialogue generation is concerned with conversational settings where participants take turns and the task is to generate a response to previous utterances. In order to generate relevant responses, prior research explored the use of the conversational history, i.e., the utterances already exchanged between the participants, as a context for the new response. It also investigated the use of external knowledge that may be a good source of content and not necessarily present in the conversational history (Ghazvininejad et al., 2018; Zheng and Zhou, 2019; Liu et al., 2021a,b; Prabhumoye et al., 2021a). However, it has been shown that adding knowledge indiscriminately, can hurt performance. Thus, the context has been used to select the best knowledge for the response generation.

Figure 1: Example showing utterances of participants A and B, the scored knowledge, and the target response. The knowledge sentence (1) is deemed the best knowledge for response generation by our proposed CKL model. The best context segments for retrieving the best knowledge are colored Brown; the best context segments for generating response are colored Blue and the best knowledge segments for the response are Purple.

Since the context itself consists of multiple utterances, the same concern applies: not all the prior utterances are equally useful for generating the response. Therefore, the context needs to be evaluated for its importance in relation to generating the response and identifying the relevant knowledge, separately. In Figure 1, we show an example from the Wizard of Wikipedia (test seen) dataset, illustrating that the best context segments for response generation may not necessarily be the best context segments for retrieving the best knowledge. Furthermore, both the context and the contextual knowledge contribute to the coverage of the target response (blue and purple words). Thus, it is important to devise effective learning methods to identify the best context for response generation

*Qatar Computing Research Institute, Doha, Qatar

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7822–7839
July 9-14, 2023

and for knowledge selection. Once the knowledge is selected, there is still a question of whether and how to refine its selection for optimal use.

Recent studies differentiated between the two context roles by adopting a pipeline approach and training different models for each of them. Zheng et al. (2020) proposed a knowledge retrieval model TPPA that re-orders retrieved knowledge guided by its relevance to the response and investigated the effects of the resulting knowledge sets in combination with generative models such as TED (Zheng and Zhou, 2019) and WSeq (Tian et al., 2017). Paranjape et al. (2021) introduced a posterior-guided training to guide the retrieval of the relevant knowledge. A BART-based generative model (a generator) is used to generate responses but the retriever and the generator are trained independently. Similarly, Glass et al. (2022) developed a Re^2G model which comprises a retriever, a re-ranker, and a generator. The re-ranker can take as input the outputs of multiple retrieval systems, e.g., ANN-based retrieval and BM25 method and the content retrieval training is an integral part of the content generation. This approach can differentiate the context roles from knowledge selection and response generation tasks. However, it requires additional training stages, which may incur and accumulate additional errors, and cannot separate the context information used for knowledge selection and response generation within the unified model.

In our work we hypothesize that the integrated approach to model training and selection of context and knowledge can be improved through a parallel learning architecture where specific content selection roles (context and knowledge) are clearly differentiated and each learning facet is supervised, controlling for model training. Guided by the hypothesis, we propose a Contextual Knowledge Learning (**CKL**) model in which we introduce *Latent Vectors* to capture context roles and knowledge characteristics: the *Context Latent Vector* for the relationship of context to the responses and to the 'best' knowledge, and the *Knowledge Latent Vector* for the knowledge to capture the importance of knowledge to the responses. *Latent Weights* are then derived from the Latent Vectors to indicate the importance of context utterances and knowledge sentences.

We also extend the notion of the Attention operation, where tokens' attention scores are entirely decided by the scaled dot product between two representations, and devise a *Latent Weight Enhanced Attention*. The attention operation is augmented with the multiplication by the tokens' attention scores and the *Latent Weights* (i.e., the context utterance's weight and knowledge sentence weight). By adopting the weak supervision technique, the *Latent Weights* for context and knowledge are supervised by the (noisy) pseudo ground truth, removing the need for human annotations. Combined with the Negative Log Likelihood loss, the CKL is trained in a unified way, differentiating the context utterances for the knowledge selection and response generation tasks.

The performance that our CKL model is superior to six strong baseline approaches, including Transformer-based and pre-trained model-based methods, on two publicly available datasets Wizard of Wikipedia and CM-DoG. By experimenting with a 50% smaller training set, our approach still outperforms the baseline methods. Figure 1 shows the effectiveness of recognizing relevant context utterances for both the knowledge selection task and dialogue generation task using the CKL model.

In summary, the key contributions of our research are through the novel design and clear advantages of the CKL method*:

- Differentiated functionality of the context utterances for the knowledge selection task and response generation task, achieved through the technique of training latent vector;
- Latent Weight Enhanced Attention module that incorporates the latent weights into the generation process;
- Effective weak supervision of latent weights training by defining the pseudo ground truths for the context latent weights and knowledge latent weights;
- Robustness of CKL, retaining its effectiveness with reduced amounts of data.

2 Related Work

Knowledge-Grounded Dialogue Generation. Research on knowledge injection into dialogue generation can be traced to Ghazvininejad et al. (2018) who demonstrated that injecting knowledge into the generative model benefits the performance due to additional information available in knowledge

*https://github.com/tonywenuon/acl2023-ckl

sentences. This led to a range of methods for knowledge injection (Zheng and Zhou, 2019; Zhao et al., 2020b; Li et al., 2019). Recently, pre-trained models were also adopted. Liu et al. (2021b) use BART (Lewis et al., 2020a) as the backbone to fine-tune the model with few-resource datasets. Prabhumoye et al. (2021a) also chose BART as the basic pre-trained framework to project context and knowledge in a unified model. However, it has been shown that not all the knowledge is useful, and therefore knowledge has to be carefully selected (Kim et al., 2020; Lian et al., 2019). Lewis et al. (2020b) propose a RAG model by leveraging retrieval techniques to obtain relevant knowledge for enhancing dialogue generation. Built on top of the RAG, Shuster et al. (2021) study various types of architectures with multiple components, including retrievers, rankers, and encoder-decoders. All previously mentioned approaches take the same context information for both the knowledge selection task and the dialogue generation task. To differentiate the functionality of the context for the two tasks, Glass et al. (2022); Paranjape et al. (2021); Zheng et al. (2020) introduce pipelines to achieve multi-stage generative models. Commonly, they devise a retriever that takes responsibility of retrieving knowledge units given the context information and follow by a generator that generates the final responses. The limitation is that this multi-stage training may cause error accumulation. Motivated by this, we train our CKL model in an end-to-end way by differentiating the context inside the model.

Weak Supervision for Dialogue Generation. For the dialogue generation task, even though the responses are naturally the ground truth for the model training, previous studies showed that several auxiliary tasks with weak supervision can help to improve the generation performance (Chang et al., 2021). Zhao et al. (2020b) uses BERT encoders (Kenton and Toutanova, 2019) and a GPT-2 decoder (Radford et al., 2019) for knowledge projection and prediction. They built pseudo ground truth documents by leveraging the similarity between each document and the response to weakly supervise document selection. Zheng et al. (2021) propose a de-noising mechanism for the knowledge tokens to be injected in a weak supervision manner. Wang et al. (2019) design a weak supervision-based discriminator to capture the relations between the answer and the corresponding passage and eventually generate the questions. Informed by these studies, in our work we introduce a weak supervision to complement the generation model training.

Our proposed CKL model combines context differentiation and weak supervision for response generation purposes. The CKL model differs from prior studies by (1) devising latent vectors for both context and knowledge to derive the context latent weights and knowledge latent weights for knowledge selection and response generation tasks, and (2) designing a latent weight enhanced attention operation, combined with a weak supervision technique to provide more effective use of context and knowledge for response generation.

3 Method

Problem and Definitions. Considering a conversational history that comprises context $C = \{c_1, c_2, \ldots, c_m\}$, our goal is to generate a response $R = \{r_1, r_2, \ldots, r_L\}$ by leveraging knowledge $K = \{k_1, k_2, \ldots, k_l\}$ that is relevant to the context C. Among the notations, r_i is each word of the response, c_i means the context utterance, and k_i denotes the knowledge sentence. L is the maximum token number of the response; m is the number of context utterances and l is the number of background knowledge sentences.

We aim to (1) calculate latent weights of context utterances and knowledge sentences (Sec. 3.2 and 3.3); (2) generate the final response given context, knowledge, and their latent weights (Sec. 3.4). In our approach, we do not integrate latent vectors into the content representation. Instead, we transform vectors into scalar values, referred to as *latent weights*. The Context Latent Weights for Response and Knowledge, (*CLWR*) and (*CLWK*), respectively, are used in the loss function and by the decoder to score content utterances. The Contextual Knowledge Latent Weights (*KLW*) are similarly used in the loss function and the decoder to score knowledge sentences.

Contextual Knowledge Learning (CKL) Architecture. Our proposed CKL method consists of four components: an encoder, a Context Latent Weight generator (*CLW* Generator in Figure 2), a Knowledge Latent Weight generator (*KLW* Generator), and a decoder. We use the state-of-the-art Transformer-based encoder-decoder model BART (Lewis et al., 2020a) as the backbone of our Encoder and Decoder. The *CLW* generator takes responsibility for producing two sets of context latent weights, one set for response generation (*CLWR*)

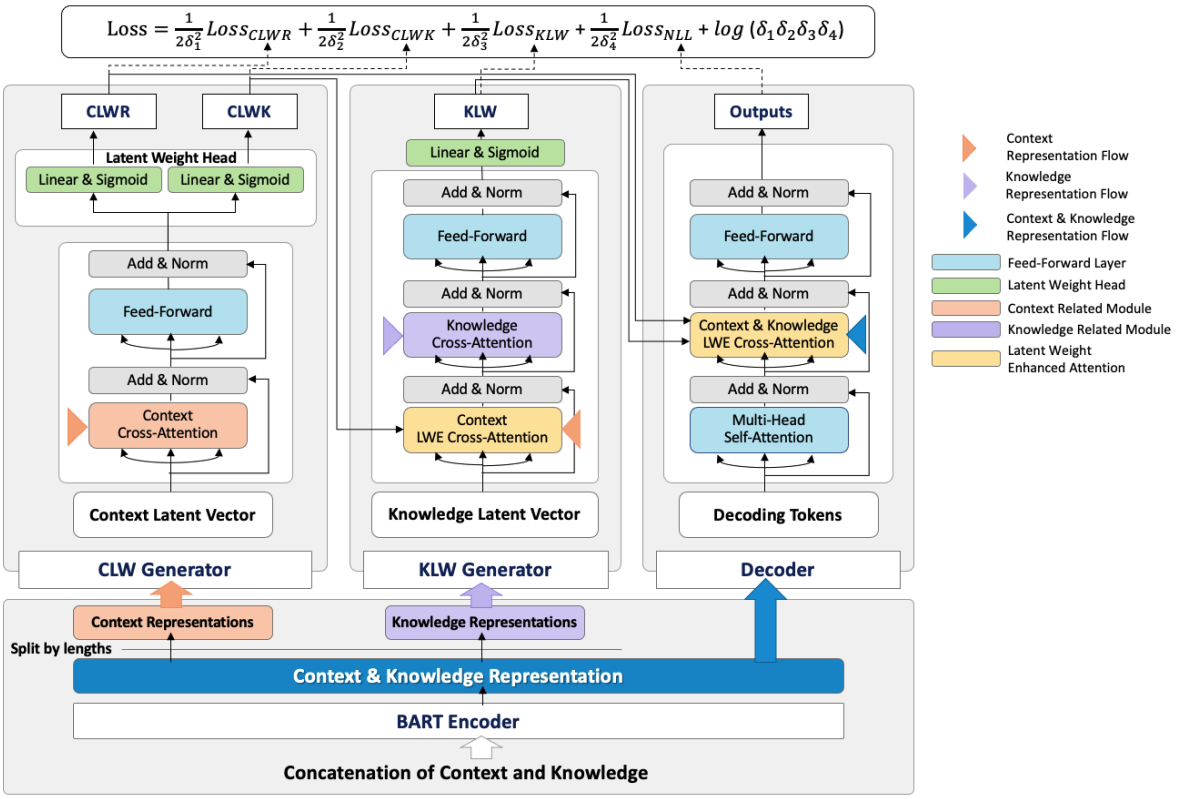

Figure 2: Contextual Knowledge Learning (CKL) model architecture comprising the BART-based encoder, two generators for latent vectors training and latent weights generation, and decoder with context and knowledge scoring. Trainable parameters δ_i balance multi-losses.

and another set for knowledge latent weight generation (*CLWK*). Similar to the *CLW* generator, the *KLW* generator is used to generate knowledge latent weights (*KLW*) which are conditioned on the context and knowledge. Finally, the decoder is a normal BART decoder but equipped with the latent weight enhanced attention mechanism.

3.1 Encoder

Leveraging the pre-trained model BART, we directly use the BART encoder to get the context and knowledge representations. The proposed CKL model needs context utterances' and knowledge sentences' representations, so they are expected to be passed through the BART encoder sequence by sequence. However, that would destroy the inner dependency between words from sequences, i.e., this means discarding the long dependency between context and knowledge. To tackle this, we first inject the concatenation of the context and knowledge to get a whole sequence representation, i.e., 'Context & Knowledge Representation' in Figure 2. Then by recognizing the context utterances' lengths and knowledge sentences' lengths, we split the whole representation into several sub-sequences, obtaining representations that take word long-dependency into account.

3.2 Context Latent Weight Generator

As shown in Figure 2, *CLW* Generator is designed to generate two sets of context latent weights: Context Latent Weight for Response (*CLWR*) and Context Latent Weight for Knowledge (*CLWK*).

Context Latent Vector. The *CLW* generator starts from a Context Latent Vector which is a trainable vector. Practically, it is a word embedding indexed by a fixed word index of 1.

Context Latent Vector Interaction with Context Representations. Here in the *CLW* generator, like the Transformer architecture, a standard cross-attention, feed-forward, and residual network are used. We introduce the cross-attention operation in detail because it will be used in the next sections. For the rest of the Transformer modules, e.g., feed-forward layer and residual network, please refer to Vaswani et al. (2017). Formally, the attention is calculated as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (1)$$

in which Q, K, and V are matrices and d is the representation dimension. Through the softmax function, the attention weights, i.e., QK^T, are normalized. Multiplying with V, the Q's representation is updated by K and V. In the *CLW* generator, Q is the context latent vector, while K and V are the context representation.

Latent Weight Head. The Latent Weight Head module contains a linear layer and a Sigmoid function. The purpose is to transfer the d dimensional context latent vector to scalar values. By doing so, each context utterance will have a latent score. To be specific, we define the Latent Weight Head as follows:

$$CLWR = Sigmoid(x_{CLV}W_1 + b_1) \tag{2}$$

$$CLWK = Sigmoid(x_{CLV}W_2 + b_2) \tag{3}$$

where $CLWR \in \mathcal{R}^{1\times m}$ and $CLWK \in \mathcal{R}^{1\times m}$ are the context latent weight scores for the response and knowledge respectively. $W_1, W_2 \in \mathcal{R}^{d\times 1}$ and $b_1, b_2 \in \mathcal{R}^1$ are trainable parameters. $x_{CLV} \in \mathcal{R}^{1\times d}$ denotes the vector converted from the Context Latent Vector. It is important to note that *CLWR* and *CLWK* have the same Latent Weight Head architecture, but do not share parameters. *CLWR* is used to identify the importance of a context utterance when generating responses and *CLWK* is used when producing knowledge latent weights, i.e., knowledge sentences' importance.

Weak Supervision on *CLWR* and *CLWK*. As illustrated, when predicting the response and producing the knowledge latent weight, the role of the context utterances should not be treated as the same: *CLWR* and *CLWK* reflect the difference. We devise two loss functions to weakly supervise them. The latent weight scores are expected to be a continuous value thus we consider this task as a regression task rather than a classification task. Mean Squared Error (MSE) is adopted as the loss function. To obtain the pseudo ground truth, we use the F1 score to measure the closeness between a context utterance and the response on the word level. As for its values, for *CLWR*, the context utterance with the maximum F1 score is tagged as 1 and the rest of the utterances to be 0. It is worth noting that the last utterance, i.e., the post, is always 1 because it has been proven crucial for response generation (Sankar et al., 2019).

For training *CLWK*, we use the same method for constructing the pseudo ground truth. The only difference is that *CLWK* is built for the knowledge latent weight, so we produce the most relevant knowledge for the F1 score calculation. First of all, we use the TF-IDF approach to retrieve from the knowledge sentences by taking the response as the query, i.e., ranking the knowledge sentence by *TF-IDF*(knowledge sentence, response).[†] The top-1 ranked sentence based on the TF-IDF is treated as the most important knowledge sentence, being tagged as *Top1-RK*. Secondly, similar to *CLWR*, the context utterance with the maximum F1(Context Utterance, *Top1-RK*) is used to supervise *CLWK*. Formally,

$$GT_{CLWR(i)} = \begin{cases} 1, \text{ if } c_i = \text{argmax(F1}(c_i, \text{R})) \\ 1, \text{ if } c_i = \text{post} \\ 0, \text{ otherwise} \end{cases} \tag{4}$$

$$GT_{CLWK(i)} = \begin{cases} 1, \text{ if } c_i = \text{argmax(F1}(c_i, \textit{Top1-RK})) \\ 1, \text{ if } c_i = \text{post} \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

in which c_i means each context utterance. Then, we define the loss function to be:

$$Loss_{CLWR} = MSE(CLWR, GT_{CLWR}) \tag{6}$$

$$Loss_{CLWK} = MSE(CLWK, GT_{CLWK}) \tag{7}$$

where GT_{CLWR} and GT_{CLWK} are the pseudo ground-truth context utterance scores for response generation and knowledge selection tasks respectively.

3.3 Knowledge Latent Weight Generator

The knowledge Latent Weight generator is designed to generate a knowledge latent weight (KLW). It begins with a knowledge latent vector, which is a word embedding indexed by a fixed index of 1. Note that the knowledge latent word embedding is different from the context latent embedding.

Latent Weight Enhanced Attention. Latent Weight Enhanced Attention (LWE Attention) is built on top of the standard attention by considering the latent weights. Originally, the attention is calculated between two sequence representations from the word level (shown in Eq. 1). The LWE Attention takes sentence-level scores, i.e., the latent weights, into consideration. By this, the Eq. 1 is then changed to be:

$$LWE\ Attention(Q, K, V) = LW \times softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{8}$$

[†]TF-IDF(·) is the TF-IDF (term frequency-inverse document frequency) function. IDF is obtained by the individual dataset.

where *LW* stands for latent weights. *LW* will be different when predicting responses and generating knowledge latent weight. Namely, in the *KLW* generator, the *LW* is replaced with *CLWK*. In the Decoder, it is changed to *CLWR* and *KLW*, which will be introduced in Sec. 3.4.

Context & Knowledge Dependency (CK-Dep for short). Prior studies (Prabhumoye et al., 2021b; Liu et al., 2021c) consider the context and knowledge dependency by stacking a context cross-attention and a knowledge cross-attention from word level. We also leverage the stacked architecture and consider the context sentence-level weights (through LWE Attention), i.e., the context LWE cross-attention module and the knowledge cross-attention module in *KLW* generator.

Weak Supervision on *KLW*. After going through the CK-Dep operation, the Latent Vector is processed by the Latent Weight Head module to get the *KLW*. The knowledge generally contains richer information than the context (Zheng and Zhou, 2019; Kim et al., 2020). For context we take the top-1 ranked utterance as the pseudo ground truth GT_{CLWK}. However, for knowledge, we set a hyper-parameter N to get the pseudo ground truth knowledge sentences GT_{KLW}. Namely, the top N ranked knowledge sentences are considered to be the ground truth for supervising *KLW*.

$$GT_{KLW(i)} = \begin{cases} 1, \text{k}_i\text{=Top N argmax(TF-IDF(k, R))} \\ 0, \text{otherwise} \end{cases} \quad (9)$$

where, k_i is each knowledge sentence. With different Extrema and Greedy scores of the ZRKGC model are lower than the CKL. This means although the generated re- sponses of the ZRKGC model are closer to the ground truth response on average, it can not seman- tically capture the most important words. Third, in terms of the diversity scores, the pro- posed CKL does not improve over other models but we expect that the method can be improved by re- fining the use of latent weights which are currently normalized between 0 and 1 and multiplied by the word attention scores. Despite of CKL model's gen- erated responses being not the most diverse among all compared models, our human evaluation results reveal that CKL is preferred by the 5 annotators with moderate agreement, in terms of relevance, coherence, informativeness, and overall preference (Appendix B.2). 3https://github.com/ellenmellon/DIALKI 4https://github.com/neukg/KAT-TSLF 5https://github.com/shrimai/ Focused-Attention-Improves-Document-Grounded-GenerationN for the individual dataset, the performance could vary. We discuss the effect of top-N in Appendix B.4. Similar to the *CLWR* and *CLWK*, we also consider the *KLW* generation as a regression task, and the loss function is:

$$Loss_{KLW} = MSE(KLW, GT_{KLW}) \quad (10)$$

3.4 Decoder and Training

The Decoder is a BART decoder but equipped with LWE Attention. In the 'Context & Knowledge LWE Cross-Attention' module in Figure 2, the context and knowledge representations are multiplied by the corresponding latent weights. Namely, the *LW* in Eq. 8 will be replaced by *CLWR* and *KLW* when dealing with context and knowledge in the Decoder. Formally, the Eq. 8 is instantiated to be:

$$\begin{aligned} &PE\ Attention(Q, K, V) = \\ &\sum_{i=1}^{m} CLWR_i \times softmax\left(\frac{QK_i^T}{\sqrt{d}}\right) V_i + \\ &\sum_{j=1}^{l} KLW_j \times softmax\left(\frac{QK_j^T}{\sqrt{d}}\right) V_j \end{aligned} \quad (11)$$

where, K_i and V_i means i-th context utterance. K_j and V_j denote j-th knowledge sentence. $CLWR_i \in \mathcal{R}^1$ and $KLW_j \in \mathcal{R}^1$ stand for the corresponding context and knowledge latent weights. The loss function for response generation is a Negative Log Likelihood loss (NLL).

$$Loss_{NLL} = -\sum_{i=1}^{L} log\ p(R_t | R_{<t}, C, K). \quad (12)$$

in which, L is the maximum length of the response, t is the t-th token to be generated and $R_{<t}$ denotes the generation steps prior to t.

Aggregation of Loss Functions. In this paper, we have four different loss functions, including Eq. 6, Eq. 7, Eq. 10 and Eq. 12. Previous studies simply aggregate different loss functions by either an addition operation (Li et al., 2019; Zheng et al., 2021) or setting hyper-parameters to do a weighted sum (Wu et al., 2021), which are sub-optimal. Kendall et al. (2018) propose a principled approach to multi-task learning which weighs multiple loss functions by considering the homoscedastic uncertainty of each task. This Automatic Weighted Loss (AWL) allows the model to simultaneously learn various quantities with different units or scales in various

settings. Adopting this strategy, we define our final loss as follows:

$$Loss = \frac{1}{2\delta_1^2} Loss_{CLWR} + \frac{1}{2\delta_2^2} Loss_{CLWK} + \frac{1}{2\delta_3^2} Loss_{KLW} + \frac{1}{2\delta_4^2} Loss_{NLL} + log(\delta_1\delta_2\delta_3\delta_4) \quad (13)$$

The final goal is to minimize the objective with respect to $\delta_1, \delta_2, \delta_3$ and δ_4 as learning the relative weight of the four different losses.

4 Experiment

4.1 Datasets, Settings, and Metrics

Datasets. Following previous research practices (Prabhumoye et al., 2021a; Liu et al., 2021b; Li et al., 2019; Zhao et al., 2020b; Liu et al., 2021c), we use two public datasets: Wizard of Wikipedia (WoW for short. Dinan et al. (2019)) and CMU-DoG (Zhou et al., 2018) to conduct our experiments. We introduce them in detail in Appendix A.1.

Experimental Settings. All of the experiments are conducted with the same hyper-parameter settings as described in Appendix A.2.

Metrics. As used in previous works (Ghazvininejad et al., 2018; Zhao et al., 2020a; Li et al., 2019; Zheng et al., 2021), we employ BLEU (Papineni et al., 2002), Rouge (Lin, 2004), Diversity (Li et al., 2015) and embedding-based metric, BOW Embedding (Liu et al., 2016a) as the metrics. Among them, BLEU calculates N-grams co-occurrence between two sequences. Rouge measures the number of overlapping units such as word sequences, and word pairs between the generated sequence and the ground truth sequence. Diversity score counts distinct N-grams number divided by the total number of the generated corpus. BOW Embedding metric leverages the pre-trained word vector to calculate the similarity between sequences from the semantic space. Meanwhile, Liu et al. (2016b); Banerjee and Lavie (2005) suggest that compared with the other metrics, BLEU2, and embedding-based metrics have a better correlation with human assessment, and thus in this paper, we take BLEU2 and embedding-based measurements as the main metrics for discussion.

To have a better understanding of the proposed model, we also conducted a human evaluation (reported in Appendix B.2).

4.2 Baseline Approaches

We compare our CKL model with six baselines.

ITDD (Li et al., 2019) proposes an incremental Transformer architecture to improve context coherence and knowledge correctness.

DRD (Zhao et al., 2020a) proposes a disentangled response decoder to isolate parameters that depend on knowledge-grounded dialogues from the entire generation model.

ZRKGC (Li et al., 2020) treats the knowledge as latent variables so that the model can estimate the knowledge representation distribution from the latent space.[‡]

DIALKI (Wu et al., 2021) proposes a knowledge identification model to provide dialogue-contextualized passage encodings and locate knowledge that is relevant to the conversation.[§]

KAT (Liu et al., 2021b) devises a three-stage architecture to get better context inner-relationship, knowledge representation, and interaction between context and knowledge.[¶]

DoHA (Prabhumoye et al., 2021a) focuses on building a context-driven representation of the document and enabling specific attention to the information in the document.[‖]

4.3 Experiment Results

Main Results. The experimental results are shown in Table 1. Because of the page limitation, we report the results of the WoW test unseen set (with similar trends) in Appendix B.1. First, based on BLEU and Rouge-L scores, the proposed CKL models perform consistently better than the baseline approaches. This reflects that the results from the CKL share more consecutive tokens with the ground truth responses. Looking closely at the BLEU-2 scores, the CKL's results are improved by large margins compared to the best results of the baseline approaches (DIALKI); they are around 15% better for the WoW test seen (improving from 13.72% to 15.80%) and around 14% better for the CMU-DoG dataset.

Second, for the embedding-based metrics, the CKL is better than most of the baseline models except for the ZRKGC model on the Embedding Average measurement. However, the Extrema and Greedy scores of the ZRKGC model are lower than the CKL. This means although the generated responses of the ZRKGC model are closer to the

[‡] https://github.com/nlpxucan/ZRKGC
[§] https://github.com/ellenmellon/DIALKI
[¶] https://github.com/neukg/KAT-TSLF
[‖] https://github.com/shrimai/Focused-Attention-Improves-Document-Grounded-Generation

Table 1: Automatic evaluation results on Wizard of Wikipedia (WoW) test seen and CMU-DoG datasets. * means significant test value with $p < 0.05$, compared to the CKL. Note that the results of ITDD and DRD are copied from the papers, so they do not have significant test results. 'w/o' means without a certain module for the ablation study while 'red.' means reduced to a subset of the whole dataset for training. All values are expressed as percentages (%).

Models	Wizard of Wikipedia test seen									
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-L	Div-1	Div-2	Average	Extrema	Greedy
ITDD (Li et al., 2019)	15.80	7.10	4.00	2.50	-	-	-	-	-	-
DRD (Zhao et al., 2020a)	21.80	11.50	7.50	5.50	-	-	-	-	-	-
ZRKGC (Li et al., 2020)	23.80*	8.80*	4.04*	1.95*	16.86*	5.44*	22.66*	**72.32***	40.40*	41.67*
KAT (Liu et al., 2021b)	16.92*	9.28*	6.04*	4.37*	16.41*	**14.66***	**45.99**	67.86*	39.06*	39.37*
DoHA (Prabhumoye et al., 2021a)	23.19*	11.70*	6.99*	4.65*	21.32*	7.44	31.47	69.91*	40.91*	41.64*
DIALKI (Wu et al., 2021)	25.00*	13.72*	9.09*	6.68*	22.10*	9.33	41.71	70.43*	42.31*	41.73*
CKL	**27.29**	**15.80**	**10.99**	**8.41**	**23.96**	9.03	36.36	71.11	42.95	42.54
w/o $Loss_{KLW}$	26.01	14.80	10.16	7.68	23.25	9.37	37.27	70.69*	42.56	42.02*
w/o $Loss_{CLWR}$	26.57	15.43	10.76	8.22	23.75	9.31	37.38	71.05	**43.50***	42.39
w/o $Loss_{CLWK}$	26.67	15.49	10.75	8.21	23.81	9.37	37.11*	70.99	43.37*	42.50
w/o *CK-Dep*	26.68	15.51	10.76	8.21	23.82	9.25	36.99*	71.18	43.35*	**42.66**
red. *1/2 training data*	25.01*	13.81*	9.23*	6.85*	22.37*	9.04	35.79*	70.94	42.30*	42.41
red. *1/4 training data*	23.48*	12.48*	8.03*	5.77*	21.17*	8.56*	34.45*	70.76	42.16*	42.63
Models	CMU-DoG test set									
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-L	Div-1	Div-2	Average	Extrema	Greedy
ITDD	9.50	3.60	1.70	0.90	-	-	-	-	-	-
DRD	15.00	5.70	2.50	1.20	-	-	-	-	-	-
ZRKGC	17.35*	5.68*	2.31*	1.00*	13.05*	1.34*	8.26*	**66.26***	31.42*	37.91*
KAT	13.53*	5.81*	2.81*	1.49*	11.98*	**4.19***	17.60*	63.72*	35.18*	37.72*
DoHA	17.02*	6.95*	3.30*	1.72*	14.41*	2.39*	12.14*	65.69	35.51*	39.26*
DIALKI	15.83*	6.41*	3.10	1.69	14.64*	3.43*	**20.43***	63.57*	34.89*	37.60*
CKL	**17.74**	**7.91**	4.11	2.29	15.87	2.30	11.10	65.63	35.81	39.46
w/o $Loss_{KLW}$	17.05	7.66	4.06	2.34	15.75	2.47	11.69	65.57*	35.74	**39.48**
w/o $Loss_{CLWR}$	17.02	7.62	3.98	2.23	**15.93**	2.34	11.32	65.19	35.71*	39.20*
w/o $Loss_{CLWK}$	17.33	7.80	**4.14**	**2.38**	15.91	2.37*	11.30	65.65	**35.83**	39.47
w/o *CK-Dep*	17.16	7.64	4.01	2.27	15.86	2.37*	11.67*	65.30*	35.63	39.10*
red. *1/2 training data*	16.99*	7.25*	3.67*	2.02*	14.84*	2.32*	11.51*	65.07*	34.68*	38.30*
red. *1/4 training data*	17.22*	7.27*	3.61*	1.94*	14.20*	2.07*	10.35*	65.11*	34.90*	39.09*

ground truth response on average, it can not semantically capture the most important words.

Third, in terms of the diversity scores, the proposed CKL does not improve over other models but we expect that the method can be improved by refining the use of latent weights which are currently normalized between 0 and 1 and multiplied by the word attention scores. Despite of CKL model's generated responses being not the most diverse among all compared models, our human evaluation results reveal that CKL is preferred by the 5 annotators with moderate agreement, in terms of relevance, coherence, informativeness, and overall preference (Appendix B.2).

Ablation Study. Four downgraded versions of CKL are provided. (1) w/o $Loss_{KLW}$, i.e., removing the knowledge latent weight loss function; (2) w/o $Loss_{CLWR}$ (deleting the context latent weight supervision for response generation); (3) w/o $Loss_{CLWK}$ (deleting the context latent weight supervision for knowledge prediction); and (4) w/o *CK-Dep* removes the context-knowledge dependency when generating *KLW*, i.e., removing the Context LWE Cross-Attention module from the *KLW* Generator in Figure 2. From Table 1 we can see that all of the BLEU-2 scores decrease. For the WoW dataset, when removing $Loss_{KLW}$ and $Loss_{CLWK}$, the BLEU-2 gets the lowest score among all of the ablation experiments, indicating the importance of knowledge latent weights generation. In terms of the CMU-DoG dataset, which has patterns different from the WoW test seen, w/o $Loss_{CLWR}$ decreases the most. Thus, correctly identifying context seems more crucial than knowledge selection for the CMU-DoG dataset. We presume that CMU-DoG's knowledge sentences are complementary, i.e., different knowledge sentences contain similar information, resulting in the context recognition showing more importance. We further verify this assumption in Appendix B.4. Other metrics decreased to varying degrees but most remain better than the baseline approaches. On the whole, the full version of the CKL performs the best.

Low-Resource Experiments. In order to test the CKL's robustness, we also conduct experiments on low-resources scenarios. From Table 1, the BLEU-2 scores of the CKL model with half of the training data are respectively 13.81% and 7.25% on WoW test seen and CMU-DoG datasets, outperforming the best baseline models (DIALKI with 13.72% on WoW, and DoHA with 6.95% on CMU-DoG). Appendix B.3 further shows although as the scale

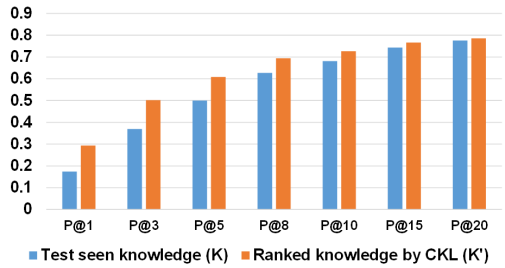

Figure 3: P@N scores for WoW test seen original knowledge set and the re-ranked knowledge set by CKL.

of the training set decreases, the performance drops gradually, our proposed CKL model still performs reasonably well with limited training data.

Latent Weight Analysis. To illustrate the effectiveness of the proposed CKL model, we demonstrate (1) knowledge re-ranking by the knowledge latent weights and (2) Spearman's Correlation between the knowledge latent weights and the pseudo ground truth scores. We use the WoW test seen set for illustration. The same patterns are found for WoW test unseen and CMU-DoG datasets.

WoW and CMU-DoG datasets provide a set of initial knowledge, designated by K. The predicted knowledge latent weights by CKL are scores for each knowledge sentence that can be used to rank knowledge and obtain re-ranked knowledge set K'. We construct pseudo ground truth knowledge order by using the response as the query to retrieve from the knowledge named K_{GT}. At this point, the top 1 ranked knowledge sentence in K_{GT} is the most relevant to the response, *Top1Klg*. We use *P@N* as the metric to evaluate the precision. For a sample, we calculate the percentage of *Top1Klg* included within the top N-ranked knowledge sentences.

Figure 3 shows the results: for the original knowledge order *K*, *P@1* is about 17.5%; for *K'* *P@1* score is around 30%. For each N, the *P@N* for *K'* is higher than for *K*. That confirms that the latent weight modules can improve the relevance scoring of knowledge sentences. In Appendix C we also provide a qualitative case study of the rankings.

To further analyze the effectiveness of the latent weights *CLWR*, *CLWK* and *KLW*, we calculate Spearman's Correlation between each weight group and the corresponding pseudo ground truths which have been elaborated on in Sec. 3 (e.g., for *CLWR* we calculate the Spearman's Correlation between *CLWR* and GT_{CLWR}). We use the ZRKGC model to provide weights for knowledge sentences without context weight (ZRKGC does not provide it) and obtain Spearman's Correlation with the pseudo-ground truth GT_{KLW}. We compare the resulting Spearman's Correlation coefficients with those of the CKL and CKL's ablated models. The results are shown in Table 2. For *KLW*, the coefficients are higher than for ZRKGC by a large margin. For CKL's ablated models, the coefficients are lower than for the full CKL model. For instance, *KLW* correlation score 0.0966 for 'w/o $Loss_{KLW}$' is much lower than the score 0.37 for CKL. This further demonstrates all supervised modules are helpful to the entire model.

Table 2: Spearman's Correlation between latent weights and the pseudo ground truths. * means t-test value with $p < 0.05$, in comparison with the proposed CKL.

Models	KLW	CLWR	CLWK
ZRKGC	0.1001*	-	-
CKL	0.3700	0.6697	0.6455
w/o $Loss_{KLW}$	0.0966*	0.6699	0.6455
w/o $Loss_{CLWR}$	0.3585	0.4658*	0.6455
w/o $Loss_{CLWK}$	0.3694	0.6696	0.5769*
w/o *CK-Dep*	0.3732	0.6696	0.4816*

5 Conclusion

Past studies on capturing context and knowledge relationships to boost the quality of response generation models were restricted to coarse-grain characterization through context-knowledge cross-attention. In this paper, we describe the Contextual Knowledge Learning (CKL) method for response generation. We propose our CKL model by using two latent vectors which are trained to capture the relationship between context, responses, and the 'best knowledge' (identified through a pre-defined default retrieval process by taking the response as the query) as well as the relationship between contextual knowledge and responses. The trained latent vectors are used to generate latent weights that enhance the traditional attention operation by multiplying them with the token-level attention scores. Furthermore, we leverage the weak supervision technique to jointly train the latent weights production as well as the response generation. With these two mechanisms, the CKL has the flexibility of influencing the learning process and has demonstrated superior performance against six strong baselines. Future work will explore more diverse use of latent vectors and latent weights as part of the learning process.

6 Limitations

The proposed CKL can automatically produce the scores for each utterance in the context and each sentence in the knowledge. However, it is constrained by the total length of the input sequence. CKL takes BART as its foundation, thus the bottleneck of BART limits the upper ability of CKL. BART requests the input length to be 1,024, which means the CKL can receive at most 1,024 tokens at a time. For some samples, the concatenation of the context and knowledge contains far more than 1,024 but is truncated to fit with the length requirement. In those cases, the CKL cannot get enough information, resulting in the sub-optimal performance of the CKL.

7 Ethical Statement

We are aware of how personal identities, expectations, norms, and values may influence our study. Our datasets were released by previous studies and are publicly available sources. Since researchers had no interactions with human subjects in this case, our ethics review board did not consider this part of our research as human subjects research. The only part of the study that involved human subjects was a crowd-sourcing human evaluation study (Appendix B.2), where participants were paid at least the minimum wage. All of our analysis were based on aggregated data without tracking down to individuals.

Acknowledgments

We would like to thank Benjamin Towle for his generous assistance in providing valuable suggestions and proofreading support. This work is partly supported by Engineering and Physical Sciences Research Council (EPSRC Grant No. EP/S515528/1, 2102871). The Titan V used for this research was donated by the NVIDIA Corporation. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Ernie Chang, Vera Demberg, and Alex Marin. 2021. Jointly improving language understanding and generation with quality-weighted weak supervision of automatic labeling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 818–829.

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 (ICLR)*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

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 *Thirty-Second AAAI Conference on Artificial Intelligence*.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue. In *ICLR*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. 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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. *Advances in Neural Information Processing Systems*, 33:8475–8485.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 12–21.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *IJCAI International Joint Conference on Artificial Intelligence*, page 5081.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yanxiang Ling, Fei Cai, Xuejun Hu, Jun Liu, Wanyu Chen, and Honghui Chen. 2021. Context-controlled topic-aware neural response generation for open-domain dialog systems. *Info. Processing & Management*, 58(1):102392.

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

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

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Shilei Liu, Xiaofeng Zhao, Bochao Li, Feiliang Ren, Longhui Zhang, and Shujuan Yin. 2021b. A three-stage learning framework for low-resource knowledge-grounded dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2262–2272.

Shilei Liu, Xiaofeng Zhao, Bochao Li, Feiliang Ren, Longhui Zhang, and Shujuan Yin. 2021c. A three-stage learning framework for low-resource knowledge-grounded dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2262–2272.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ashwin Paranjape, Omar Khattab, Christopher Potts, Matei Zaharia, and Christopher D Manning. 2021. Hindsight: Posterior-guided training of retrievers for improved open-ended generation. *arXiv preprint arXiv:2110.07752*.

Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W Black, and Ruslan Salakhutdinov. 2021a. Focused attention improves document-grounded generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4274–4287.

Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W Black, and Ruslan Salakhutdinov. 2021b. Focused attention improves document-grounded generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4274–4287.

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

Chinnadhurai Sankar, Sandeep Subramanian, Christopher Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to make context more useful? an empirical study on context-aware neural conversational models. In *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–236.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yutong Wang, Jiyuan Zheng, Qijiong Liu, Zhou Zhao, Jun Xiao, and Yueting Zhuang. 2019. Weak supervision enhanced generative network for question generation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3806–3812.

Zeqiu Wu, Bo-Ru Lu, Hannaneh Hajishirzi, and Mari Ostendorf. 2021. Dialki: Knowledge identification in conversational systems through dialogue-document contextualization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1852–1863.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. Low-resource knowledge-grounded dialogue generation. In *International Conference on Learning Representations (ICLR)*.

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.

Wen Zheng, Natasa Milic-Frayling, and Ke Zhou. 2020. Approximation of response knowledge retrieval in knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3581–3591.

Wen Zheng, Nataša Milić-Frayling, and Ke Zhou. 2021. Knowledge-grounded dialogue generation with term-level de-noising. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2972–2983.

Wen Zheng and Ke Zhou. 2019. Enhancing conversational dialogue models with grounded knowledge. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 709–718.

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*, pages 708–713.

A Experimental Setup

To enable the reproduction of results, we make our code publicly available at https://github.com/tonywenuon/acl2023-ckl.

A.1 Datasets

Wizard of Wikipedia Dataset. The dataset (Dinan et al., 2019) includes knowledge annotated by workers from the Amazon Mechanical Turk (AMT) platform. Each worker was given a set of background knowledge retrieved based on the dialogue history from Wikipedia articles to assess. The dataset is split into train/validation/test sets. For the validation and test sets, based on the topic existence in the train set, there are two versions: seen set (the topics exist in the train set) and the unseen set (some new topics are not contained in the train set). The original dataset can be obtained at http://parl.ai. The train/seen validation/seen test sets' sizes are 74,092/3,939/3,865. For the seen validation/seen test sets, the sizes are 3,927/3,924.

CMU-DoG Dataset. This dataset is proposed by Zhou et al. (2018). It also employs workers from the AMT and its conversations are mainly about movies. They are required to exchange ideas about movies. The original dataset can be obtained from the published paper[**]. Li et al. (2019) also released a CMU-DoG dataset[††], which has tokenization to all of the source and target sequences. In our paper, we use the ITDD version. The train/validation/test sets consist of 66,332/3,269/10,502 samples.

A.2 Experiment Settings

In our experiment, the BART-base model[‡‡] is used. The maximum source length is 1024 tokens and 64 tokens for the target length. For the number of context utterances, we use the latest 10 utterances. In terms of knowledge, we use the maximum source length rather than the number of knowledge sentences to determine how many to incorporate. The learning rate is set to be 5e-5. All of the experiments are trained for 10 epochs on a single TITAN V GPU. The proposed CKL model needs about 20 hours for training on the Wizard of Wikipedia dataset and about 8 hours on the CMU-DoG dataset.

B Complementary Experimental Results

B.1 Results on WoW Test Unseen Set

From Table 3, it is clear that the CKL performs best among all of the experiments in terms of the BLEU and Rouge scores. Like the results of the

[**] https://github.com/festvox/datasets-CMU_DoG
[††] https://github.com/lizekang/ITDD
[‡‡] https://huggingface.co/facebook/bart-base

Table 3: Automatic evaluation results on Wizard of Wikipedia test unseen set. * means significant test value with $p < 0.05$, compared to the CKL. 'w/o' means without a certain module for the ablation study. All values are expressed as percentages (%).

Models	Wizard of Wikipedia test unseen									
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-L	Div-1	Div-2	Average	Extrema	Greedy
ITDD (Li et al., 2019)	13.40	4.70	2.10	1.10	-	-	-	-	-	-
DRD (Zhao et al., 2020a)	20.70	10.10	6.20	4.30	-	-	-	-	-	-
ZRKGC (Li et al., 2020)	23.30*	8.50*	3.94*	1.94*	16.81*	3.48*	15.78*	**71.93***	39.11*	41.43*
KAT (Liu et al., 2021b)	16.49*	8.49*	5.30*	3.67*	15.60*	**9.84***	**31.13***	67.02*	37.20*	38.55*
DoHA (Prabhumoye et al., 2021a)	22.28*	10.53*	6.16*	4.10*	20.04*	5.04*	21.97*	68.61*	37.96*	40.23*
DIALKI (Wu et al., 2021)	25.26*	13.96*	9.36*	6.96*	22.02*	5.28	25.67	69.67*	40.94*	41.39*
CKL	**27.68**	**16.05**	**11.22**	**8.62**	**23.93**	4.89	18.60	70.36	**41.93**	**41.86**
w/o $Loss_{KLW}$	26.57	15.13	10.38	7.86	23.39	5.31	20.25	69.99*	41.51	41.62
w/o $Loss_{CLWR}$	26.87	15.48	10.77	8.21	23.53	5.04	19.26	69.97*	41.68	41.73
w/o $Loss_{CLWK}$	26.49	15.13	10.47	7.97	23.33	5.16	19.67	69.85*	41.59	41.56*
w/o *CK-Dep*	27.01	15.55	10.85	8.34	23.67	5.20	19.93	70.19	41.74	41.83

Wizard of Wikipedia (WoW) test seen set, the diversity scores and the Embedding-Average score are slightly worse than KAT and ZRKGC respectively. Remarkably, the BLEU-2 score improves from 13.96% to 16.05%, which is even higher than that of the WoW test seen set, which indicates the effectiveness of the CKL model.

B.2 Human Evaluation

We also conducted human evaluation by deploying users through the crowd-sourcing Amazon MTurk platform. 5 AMT workers were employed to assess samples from 4 perspectives, i.e., Relevance, Coherence, Informativeness, and Overall Preference. Following Ling et al. (2021), the four criteria are referred to as: **Relevance** - whether the generated response is relevant to the given context. **Coherence** - whether the generated response is a coherent and meaningful continuation of the dialogue. **Informativeness** - how many new and diverse expressions do the generated responses introduce. **Overall Preference** - personal preference between two responses.

To do the human evaluation, we randomly select 100 samples from the outputs of the proposed CKL and DIALKI (the best-performing baseline model) for the Wizard of Wikipedia and CMU-DoG datasets respectively. Then we use the Amazon Mechanical Turk platform for assessment. The assessors are asked to select a response that they preferred from two models on different perspectives (relevance, coherence informativeness, and overall), and they are also allowed to consider both responses are equal to the given context, i.e., 'Tie' in Table 4.

The results are shown in Table 4, from which we can see that for all of the four criteria, the proposed CKL is better than the DIALKI. This indicates that

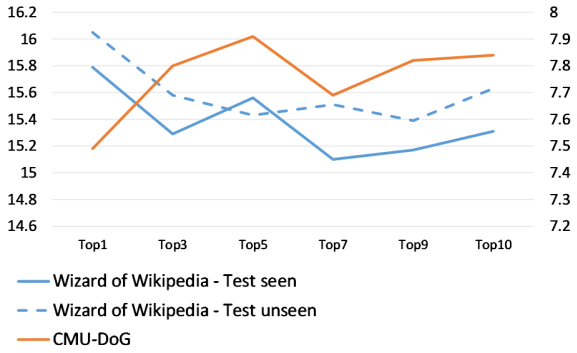

Figure 4: BLEU-2 scores (Y-axis) for WoW and CMU-DoG datasets with different numbers of top N retrieved knowledge sentences being the ground truth.

the CKL model improves in terms of relevance, coherence, and informativeness. We calculate Fleiss' Kappa (Fleiss and Cohen, 1973) for each criteria. The resulting Kappa scores are around 0.4, which indicates a moderate agreement among the assessors. From Table 4, we can observe that our proposed CKL model outperforms the best baseline DIALKI model from all perspectives.

B.3 Low-Resource Complementary Experiments

Table 5 shows the results of the complementary experiments for low-resource training. We can clearly see the effectiveness of the proposed CKL. As the scale of the training set decreases, the performance drops gradually. However, when the scale of the training data goes down to less than 1/4 of the original training data, the performance decreases more dramatically. Our proposed CKL model performs reasonably well with limited training data, but models with sufficient amount of training data are still preferred.

Table 4: Human evaluation on Wizard of Wikipedia test seen and CMU-DoG datasets. The values except for Kappa are in percentage (%).

CKL vs. DIALKI	Relevance				Coherence			
	Win	Loss	Tie	Kappa	Win	Loss	Tie	Kappa
WoW test seen	41.34	25.25	33.41	0.30	42.66	24.26	33.08	0.33
CMU-DoG	49.51	20.79	29.70	0.39	44.55	26.73	28.72	0.45
CKL vs. DIALKI	Informativeness				Overall Preference			
	Win	Loss	Tie	Kappa	Win	Loss	Tie	Kappa
WoW test seen	43.23	24.26	32.51	0.31	36.30	22.28	41.42	0.39
CMU-DoG	48.51	39.61	11.88	0.42	50.50	38.61	10.89	0.41

Table 5: Wizard of Wikipedia test seen & unseen and CMU-DoG evaluation results on low-resource scenarios. * means significant test value with $p < 0.05$, in comparison with the full version of CKL. All values are expressed as percentages (%).

Models	Wizard of Wikipedia test seen									
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-L	Div-1	Div-2	Average	Extrema	Greedy
Full training data	27.29	15.80	10.99	8.41	23.96	9.03	36.36	71.11	42.95	42.54
1/2 training data	25.01*	13.81*	9.23*	6.85*	22.37*	9.04	35.79*	70.94	42.30*	42.41
1/4 training data	23.48*	12.48*	8.03*	5.77*	21.17*	8.56*	34.45*	70.76	42.16*	42.63
1/8 training data	19.43*	10.38*	6.86*	5.10*	18.32*	9.58*	36.35*	67.84*	40.88*	41.08*
1/16 training data	20.02*	9.77*	6.04*	4.26*	18.84*	8.33*	31.82*	67.12*	40.03*	40.05*
Zero training data	9.01*	4.14*	2.34*	1.54*	11.44*	5.66*	25.63*	56.87*	34.90*	35.45*
Models	Wizard of Wikipedia test unseen									
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-L	Div-1	Div-2	Average	Extrema	Greedy
Full training data	27.68	16.05	11.22	8.62	23.93	4.89	18.60	70.36	41.93	41.86
1/2 training data	25.41*	14.09*	9.46*	7.06*	22.41*	5.14	20.17	69.88*	40.48*	41.75
1/4 training data	23.25*	12.25*	7.94*	5.76*	20.87*	5.31*	20.93*	69.69*	40.55*	41.70
1/8 training data	18.81*	10.09*	6.79*	5.11*	18.04*	5.48*	20.58*	65.88*	38.82*	39.92*
1/16 training data	19.96*	9.64*	6.04*	4.33*	18.52*	4.49*	17.04*	65.56*	38.12*	39.00*
Zero training data	8.76*	3.96*	2.24*	1.47*	11.32*	3.55*	17.10*	56.36*	33.54*	35.60*
Models	CMU-DoG									
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-L	Div-1	Div-2	Average	Extrema	Greedy
Full training data	17.74	7.91	4.11	2.29	15.87	2.30	11.10	65.63	35.81	39.46
1/2 training data	16.99*	7.25*	3.67*	2.02*	14.84*	2.32*	11.51*	65.07*	34.68*	38.30*
1/4 training data	17.22*	7.27*	3.61*	1.94*	14.20*	2.07*	10.35*	65.11*	34.90*	39.09*
1/8 training data	15.47*	6.11*	2.94*	1.59*	13.74*	2.59*	13.00*	62.11*	33.53*	37.16*
1/16 training data	15.01*	5.68*	2.61*	1.32*	13.97*	2.43*	12.22*	62.05*	34.06*	36.72*
Zero training data	7.21*	2.51*	0.96*	0.42*	8.47*	2.92*	18.99*	62.65*	30.26*	35.61*

B.4 Effect of Top-N Retrieved Knowledge

In Sec. 3.3, we set top N retrieved knowledge sentences as the ground truth for obtaining GT_{KLW}. We will discuss how different N affects the performance. As can be seen in Figure 4, we investigate from the top 1 to top 10 retrieved knowledge sentences. It is clear that for the WoW dataset, using the first retrieved knowledge gets the best results, while for the CMU-DoG dataset, the top-5 group peaks. That indicates that in the WoW dataset, the knowledge other than the top 1 retrieved sentence contains limited useful information. However, the knowledge in the CMU-DoG dataset complements each other. This also explains why when removing $Loss_{CLWR}$, the BLEU-2 decreases most for CMU-DoG (see ablation study in Sec 4.3).

C Case Study

In order to qualitatively demonstrate the results generated by the CKL and the baseline models, we report a good case and a bad case which are generated by the CKL model. The good case is selected by this criterion for a sample: if a knowledge weight is the highest among all of the knowledge sentences and it is also the ground truth knowledge (see latent weight analysis in Sec.4.3), it is viewed as good. On the contrary, if the knowledge sentence is the ground truth knowledge but predicted to have the lowest latent score, it is a bad case. From Table 6, we can see that in the good case, if the knowledge is predicted correctly, the CKL-generated response is very close to the ground truth target. The other responses produced by the baseline approaches tend to be generic. In terms of the bad case, due to the lack of important information used in the target, the CKL response does not share keywords with the ground truth target even though it is a proper answer for the given context. This comes to the one-to-many problem which describes that many responses are reasonable for a certain dialogue scenario by the natural language. This problem is out of this paper's scope.

	Context, Target and Generated Responses
Good Case	**Context:** Pop music. I sure am a fan of pop music these days. It is so much fun to listen to! **Ground Truth Target:** Pop music is a genre of popular music that originated in its modern form in the United States, and yes it is fun to listen to.
	ZRKGC: I'm a fan of pop music, but it is not the most popular genre. **KAT:** I love pop music too! it originated in the US and UK in the 1950s. **DoHA:** Pop music is a genre of music that originated in the US and Canada. **DIALKI:** Pop and rock are popular music. I like both but I prefer pop music. **CKL:** Pop music is a genre of popular music that originated in its modern form in the United States and the UK during the mid-1950s.
	Context, Target and Generated Responses
Bad Case	**Context:** **A:** Swimming is consistently among the top public recreational activities, and in some countries, swimming lessons are a compulsory part of the educational curriculum. **B:** Swimming has been consistently the most popular recreational activity. **A:** It is a fantastic exercise that people can do at almost any age. **Ground Truth Target:** Especially since it mainly relies on the body's natural buoyancy.
	ZRKGC: Swimming is consistently the most popular recreational activity. I don't know about it, but in some countries, they have to be a time. **KAT:** It is! Humans can hold their breath underwater and undertake rudimentary locomotive swimming within weeks of birth. **DoHA:** It is! And in some countries, swimming lessons are a compulsory part of the educational curriculum. **DIALKI:** It is amazing that people can do it almost every day, whether it be for recreation, sport, exercise, or survival. **CKL:** I think it is the most popular recreational activity.

Table 6: A good case and a bad case of the proposed CKL model. In the base case, 'A' and 'B' denote the two Amazon MTurk workers exchanging their opinions.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☑ A2. Did you discuss any potential risks of your work?
6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

4

☑ B1. Did you cite the creators of artifacts you used?
4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
4

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
7

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

C ☑ Did you run computational experiments?

4.1

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Appendix B.2

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We design an annotation task containing detailed instructions for the annotators, asking the annotators to assess which generated response is preferred. The design of the instruction is not this paper's focus, so we don't report it.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix B.2

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix B.2

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our human evaluation just collects the opinions from the annotators. For example, given two responses, they need to assess which is perferred. The whole process is fully anonymous.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
The annotators are anonymous on Amazon MTurk platform. We can't trace the demographic and geographic characteristics.