Dial-MAE: ConTextual Masked Auto-Encoder for Retrieval-based Dialogue Systems

Zhenpeng Su[1,2]*, **Xing Wu**[1,2]*, **Wei Zhou**[1,2]†, **Guangyuan Ma**[1,2], **Songlin Hu**[1,2]†
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{suzhenpeng,wuxing,zhouwei,maguangyuan,husonglin}@iie.ac.cn

Abstract

Dialogue response selection aims to select an appropriate response from several candidates based on a given user and system utterance history. Most existing works primarily focus on post-training and fine-tuning tailored for cross-encoders. However, there are no post-training methods tailored for dense encoders in dialogue response selection. We argue that when the current language model, based on dense dialogue systems (such as BERT), is employed as a dense encoder, it separately encodes dialogue context and response, leading to a struggle to achieve the alignment of both representations. Thus, we propose Dial-MAE (Dialogue Contextual Masking Auto-Encoder), a straightforward yet effective post-training technique tailored for dense encoders in dialogue response selection. Dial-MAE uses an asymmetric encoder-decoder architecture to compress the dialogue semantics into dense vectors, which achieves better alignment between the features of the dialogue context and response. Our experiments have demonstrated that Dial-MAE is highly effective, achieving state-of-the-art performance on two commonly evaluated benchmarks.

1 Introduction

The retrieval-based dialogue system is a popular research topic. Pre-trained language models (PLMs), especially deep bidirectional Transformer Language Models (LMs) like BERT encoder (Vaswani et al., 2017; Devlin et al., 2019), have been widely used in dialogue response. Common uses of deep LM are cross-encoder and bi-encoder (Gao and Callan, 2021). Recent works (Gu et al., 2020; Whang et al., 2021; Xu et al., 2021; Han et al., 2021; Zhang et al., 2022) on dialogue response retrieval systems are mostly based on cross-encoders, which feed both the dialogue context and response directly into LM and use attention over all tokens to output a relevance score. Although cross-encoders have relatively stronger performances, they need to compute the matches for every possible combination of context-response pairs, which is time-consuming (Lan et al., 2021). In practice, cross-encoders are often used for re-ranking after dialogue retrieval. In contrast, another common use of deep LM is the dense encoder, i.e. bi-encoder, which encodes dialogue context and response into the context vector and response vector respectively. The correlations between context and responses are computed using cosine similarity or dot product functions in vector space (Lan et al., 2021; Gao et al., 2022). The bi-encoders have a faster computational speed but usually perform worse than the cross-encoder.

Bi-encoders generally underperform compared to cross-encoders due to two main reasons below (Han et al., 2021; Gao and Callan, 2021; Lan et al., 2021). Firstly, bi-encoders encode dialogue context and responses separately, which lacks deep interaction like the cross-encoder (Han et al., 2021). We consider this as a potential information barrier that hinders the performance of bi-encoders, resulting in significant differences between the dense vector representations of the dialogue context and response vectors. Secondly, language models like BERT (Devlin et al., 2019) have not been trained to aggregate complex information into a single dense representation (Gao and Callan, 2021). Although using contrastive learning during the fine-tuning can alleviate the above two issues (Lan et al., 2021), the discussion regarding their mitigation with post-training remains absent in dialogue response selection. We argue that post-training a PLM specifically tailored for the dense dialogue retrieval is essential for achieving optimal performance.

In this paper, we focus on the above two issues and propose **Dial-MAE** (**Dial**ogue Contextual **M**asking **A**uto-**E**ncoder), a simple and effective post-training method tailored for the bi-encoder

*These authors contributed equally to this work.
†Corresponding authors.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 820–830

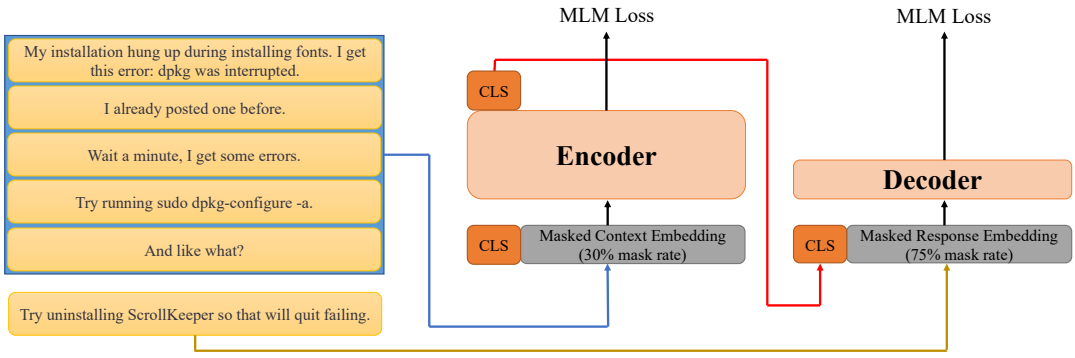

Figure 1: The model design for Dial-MAE. The input of the encoder is the dialogue context, and its next response and dialogue context embedding output by the encoder is used as the input to the decoder.

to compress dialogue semantic information and enhance the representation of dialogue-dense vectors. Our method provides a stronger foundation for model fine-tuning. Specifically, during the modeling process, we consider both the semantics of the dialogue context and the semantic relevance of the response.

As shown in Figure 1, we introduce an asymmetric encoder-decoder architecture. With the help of the dialogue context embedding [**CLS**] output by the encoder, the auxiliary task utilizes a weak decoder to reconstruct the masked response text. In other words, we employ the embedding of the dialogue context to directly generate responses. Therefore, even if the encoder side only receives the inputs of dialogue contexts, the output dialogue context embedding still needs to consider the correct response. This enables the dialogue context embedding [**CLS**] to incorporate contextual information. In addition, the encoder is required to directly predict the correct response when encoding the dialogue context, which breaks the information barrier between the context and the response. Therefore, the context and response features output by Dial-MAE are more similar, and our ablation experiments also prove this.

Furthermore, it is noteworthy that, similar to (Xiao et al., 2022; Gao and Callan, 2021), we apply asymmetric mask rates to the encoder and decoder. The decoder side has a higher mask rate than the encoder side. Such a design has the following advantage. Since the decoder has limited modeling capacity and high mask rate, the reconstruction on the decoder side is difficult to accomplish only by relying on masked response and rely more on the dialogue embedding output by the encoder, this forces the encoder to sufficiently aggregate the semantics of the dialogue context to aid the decoder in its MLM task (Xiao et al., 2022; Gao and Callan, 2021).

Our contributions are as follows:

1. We introduce Dial-MAE, a novel post-training method designed for bi-encoders, which utilizes dialogue context embeddings to generate responses, aiming to achieve feature alignment.

2. We design a novel asymmetric encoder-decoder architecture to enhance the representational power of dialogue embedding.

3. Experimental results show that in dialogue response retrieval, our method achieves state-of-the-art on two benchmarks with faster response speed.

2 Related Work

In this section, we first discuss traditional retrieval dialogue systems based on neural networks, and then we discuss current dialogue systems based on pre-trained language models.

2.1 Neural Dialogue Response Retrieval

Dialogue response selection aims to select the most appropriate response from a range of candidates. Earlier studies (Kadlec et al., 2015; Lowe et al., 2015) focused on single-turn response selection. Later, more and more studies paid attention to multi-turn dialogue response selection. Lowe et al. (2015) introduce a method that calculates the matching degree between dialogue and response based on Recurrent Neural Networks (RNNs). They also contributed a benchmark dataset named Ubuntu V1. In a similar vein, Kadlec

et al. (2015) advocate for the use of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) as encoders to represent both the context and response. However, these methods do not explicitly treat each utterance as a unit, making it difficult to capture utterance-level discourse information. Zhou et al. (2016) propose a multi-view model that encodes both word-level and utterance-level representations. Meanwhile, to fully reflect the relationship between dialogue and response, Wu et al. (2017) suggest utilizing word embeddings and their sequential representations, encoded by Gated Recurrent Units (GRU), to construct a matching matrix between the dialogue context and response. With the popularity of attention mechanisms(Luong et al., 2015; Vaswani et al., 2017). Zhou et al. (2018) propose a deep attention-matching network that applies the attention mechanism to the response selection dialogue system. Furthermore, Tao et al. (2019) advocate for context and response matching by stacking multiple interaction blocks, providing a nuanced perspective. In a similar vein, Yuan et al. (2019) introduce a multi-hop selector network designed to identify relevant utterances in the context of response matching. However, most traditional retrieval models are lightweight networks, and their performance is difficult to compare with PLMs.

2.2 PLM-based Dialogue Response Retrieval

Since PLMs show impressive performances in various downstream NLP tasks(Devlin et al., 2019; Brown et al., 2020; Touvron et al., 2023; Su et al., 2023; Wu et al., 2023). More and more studies apply PLMs to response selection. BERT-VFT (Whang et al., 2020) first applies the pre-trained language model BERT to dialogue response selection, and achieves state-of-the-art results. SA-BERT (Gu et al., 2020) adds speaker embedding to the model, in order to make the model aware of the speaker change information. Multi-Task Learning is also an effective way, UMS$_{BERT+}$ (Whang et al., 2021) proposes a set of strategies, which aids the response selection model towards maintaining dialogue coherence. Alternatively, Xu et al. (2021) propose learning a context-response matching model with multiple auxiliary self-supervised tasks. However, these methods have the problem of not fully considering the relationship between each utterance in the context. BERT-FP (Han et al., 2021) proposes to classify the relationship between a given utterance and a target utterance into more fine-grained labels, which makes the model learn the semantic relevance and coherence between the utterances. Zhang et al. (2022) propose two-level supervised contrastive learning so that the learned dialogue representations can be further separated in the embedding space. In addition, DR-BERT(Lan et al., 2021) explores the transfer of techniques from dense passage retrieval community to dialogue response selection. Although DR-BERT (Lan et al., 2021) propose fine-tuning PLMs through contrastive learning to enhance the representation capability of dialogue-dense vectors, there has been no research on tailoring post-training tasks to enhance the representation ability of dialogue-dense vectors.

3 Methodology

This section first introduces masked language model pre-training as preliminary knowledge. Then we introduce detailed post-training, including the construction of data and the auxiliary task. Finally, we introduce the details of fine-tuning.

3.1 Masked Language Model Pre-training

MLM is an unsupervised method that masks parts of the input tokens and requires the Transformers-based LM to predict them based on the unmasked tokens. Formally, given an input sentence $X = [x_1, x_2, ..., x_n]$. We select a certain percentage of tokens from X and replace them with a special token $[\mathbf{MASK}]$ to get corrupted $\widetilde{X}$. We denote these tokens replaced by $[\mathbf{MASK}]$ as $m(X)$. Then, LM is used to transform the corrupted input into the hidden states:

$$[\mathbf{h}^l_{cls}, \mathbf{h}^l] = LM([CLS], \tilde{X}) \tag{1}$$

Here, $[\mathbf{CLS}]$ is a special token that is prepended at the beginning of the text. $\mathbf{h}^l_{cls}$ and $\mathbf{h}^l$ respectively represent the hidden states of the final layer output after the $[\mathbf{CLS}]$ and $\widetilde{X}$ pass through the LM, i.e., $\mathbf{h}^l = [\mathbf{h}^l_1, \mathbf{h}^l_2, ..., \mathbf{h}^l_n]$. For masked token, its corresponding hidden feature is used to predict the actual label. We formulate this process as:

$$\mathcal{L}_{mlm} = -\sum_{x_i \in m(X)} \log p(x_i | LM([CLS], \tilde{X})) \tag{2}$$

3.2 Dial-MAE: Dialogue Contextual Masking Auto-Encoder

Dial-MAE learns dialogue context information, which jointly models the semantics of the tokens inside a dialogue context and its response. We first describe how to build training data from all utterances of the dialogue session and then introduce the Dial-MAE post-training method. We randomly sample multiple consecutive utterances as context and the next utterance as its response. Multiple utterances of the context are connected using $[\mathbf{SEG}]$. For each dialogue scene, we sample multiple sets of such context and response pairs. The sampled context and response will serve as input to the encoder and decoder, respectively.

Then, we introduce the post-training design for Dial-MAE, as shown in Figure 1, we use an asymmetric encoder-decoder: A deep encoder to generate dialogue context embedding, and a shallow transformer-based decoder (e.g. one or two layers) for response reconstruction. We apply a BERT encoder $Enc(.)$ with 12 layers, which receives masked dialogue context as inputs. The deep encoder has enough parameters to learn good dialogue representations, following the common practice, we select the final hidden state from the $[\mathbf{CLS}]$ token as the dialogue context embedding. The decoder is designed to assist the encoder in learning a better semantic representation of the dialogue. The input of the decoder $Dec(.)$ is the masked response as well as the dialogue context embedding, and it reconstructs the masked response tokens with the help of the context embedding.

Through our design, the encoder $Enc(.)$ needs to predict the features of the correct response when encoding the dialogue context. This makes the dense encoder with behavior similar to that of a cross-encoder: simultaneously considering both the dialogue context and the response. The advantage of doing this is to achieve feature alignment between the dialogue context and response during the post-training. Meanwhile, since the auxiliary MLM task breaks down the information barrier between separately encoding the dialogue context and response, the encoded output's $[\mathbf{CLS}]$ hidden state encompasses information from both. Furthermore, it is worth noting that we employ an asymmetric masking operation(eg., 30% for encoder, 75% for decoder). On the decoder side, an aggressive mask rate and fewer model parameters will force its MLM task to rely more on the encoder's context embedding, which helps the encoder side aggregate complex information about the dialogue context into a dense vector.

Formally, we denote the dialogue context as c and the response as r. We apply random mask operation to context to get $\tilde{c}$, denoting these tokens replaced by $[\mathbf{MASK}]$ in context as $m_{enc}(c)$. Similarly, we apply a random masking operation with a higher masking ratio for response to get $\tilde{r}$, denoting these tokens replaced by $[\mathbf{MASK}]$ in response as $m_{dec}(r)$. The encoding process can be expressed as:

$$[\mathbf{h}^c_{cls}, \mathbf{h}^c] = Enc([CLS], \tilde{c}) \tag{3}$$

$$[\mathbf{h}^r_{cls}, \mathbf{h}^r] = Dec(\mathbf{h}^c_{cls}, \tilde{r}) \tag{4}$$

On the encoder side, the original context is learned to be reconstructed by optimizing the cross-entropy loss:

$$\mathcal{L}_{enc} = -\sum_{c_i \in m_{enc}(c)} \log p(c_i | Enc([CLS], \tilde{c})) \tag{5}$$

Differently, on the decoder side, the decoder reconstructs the original response with the help of the context embedding h^c_{cls}. We formulate this process as:

$$\mathcal{L}_{dec} = -\sum_{r_i \in m_{dec}(r)} \log p(r_i | Dec(h^c_{cls}, \tilde{r})) \tag{6}$$

Then, we add the encoder and decoder losses to obtain a summed loss:

$$\mathcal{L} = \mathcal{L}_{enc} + \mathcal{L}_{dec} \tag{7}$$

3.3 Fine-tuning for dialogue response selection

At the end of Dial-MAE post-training, fine-tuning is conducted on the downstream dialogue response selection to verify the effectiveness of post-training. As shown in Figure 2, in the fine-tuning stage, we only keep the encoder and discard the decoder. The encoder weights are used to initialize a dialogue context encoder f_c and a response encoder f_r, respectively.

The dialogue consists of a context c that includes multiple utterances and a response r with one utterance. After the dialogue context and response pass through the encoder, the context vector and response vector are respectively output. We train a dialogue response selection model using a contrastive learning loss function.

$$\mathcal{L}_{ft} = \frac{exp(d(c, r^+))}{exp(d(c, r^+)) + \sum_j exp(d(c, r^-_j))} \tag{8}$$

r^+ is the correct response corresponding to the dialogue context c. r^- represents negative samples within a mini-batch. At inference time, we use the dot product $d(c, r)$ to measure the similarity between the context vector and the response vector:

$$d(c, r) = f_c(c) \cdot f_r(r) \tag{9}$$

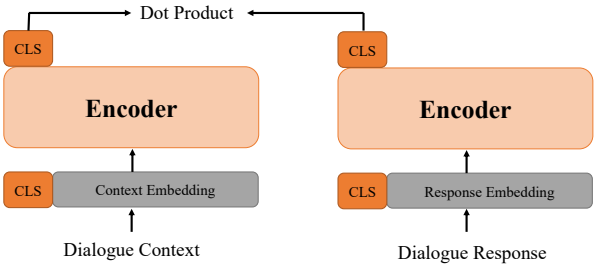

Figure 2: We discard the decoder, initialize the context encoder and response encoder using the encoder part of Dial-MAE, and fine-tune using contrastive learning. At inference time, We use a dot product to measure similarity.

4 Experiment

In this section, we first introduce our experimental details, including datasets, evaluation metrics, post-training, and fine-tuning. Then we introduce the experimental results.

4.1 Datasets

We tested our model on widely used benchmarks that include Ubuntu Corpus and E-commerce Corpus. The statistics for the two datasets are presented in Table 1.

1. **Ubuntu Corpus.** Ubuntu IRC Corpus V1 (Lowe et al., 2015) is a publicly available domain-specific dialogue dataset. Each set of conversations has two participants discussing how to troubleshoot Ubuntu systems.

2. **E-commerce Corpus.** E-commerce Corpus (Zhang et al., 2018) comprises genuine conversations in Chinese between customers and customer service personnel, collected from Taobao, a Chinese e-commerce platform.

4.2 Evaluation Metric

We evaluated our model using $R_{10}@k$, following previous studies (Han et al., 2021; Zhang et al., 2022), we evaluate our model using $R_{10}@k$. The notation $R_{10}@k$ represents Recall, indicating that among ten possible responses, the correct answer is included within the top k options.

Dataset	Ubuntu			E-commerce		
	train	val	test	train	val	test
context-response pairs	1M	500k	500k	1M	10k	10k
pos : neg	1:1	1:9	1:9	1:1	1:1	1:9
avg turns	10.13	10.11	10.11	5.11	5.48	5.64

Table 1: Statistics related to data for the Ubuntu and E-commerce Corpus.

4.3 Implementation Details

We first introduce the experimental setup for post-training, followed by the experimental setup for contrastive learning.

Post-training. Dial-MAE's encoder is initialized with a pre-trained 12-layer BERT-base model, while the decoder is initialized from scratch. Specifically, following the previous works, for the E-commerce dataset, we employ bert-base-chinese[1]. For the Ubuntu dataset, we utilize the bert-base-uncased[2]. We pre-train the model using the AdamW optimizer for a maximum of 15k steps, a global batch size of 1k, and a linear schedule with a warmup ratio of 0.1 on all two datasets. We set the input sequence lengths to 256 and 64 for the encoder and decoder, respectively. In fact, for the Chinese datasets E-commerce, we followed the parameter settings from Cot-MAE(Wu et al., 2023): The masking ratio of the encoder is 30%, the masking rate of the decoder is 45%, the learning ratio is 1e-4, and the decoder has two layers. Differently, for the English dataset Ubuntu, the masking ratio of the encoder is 30%, the masking ratio of the decoder is 75%, and the decoder is one layer. We also adjust the learning rate to 3e-4 to ensure the loss function converges. We set a widely used random seed as 42 for reproducibility. After post-training, we discard the decoder, only leaving the encoder for fine-tuning.

Fine-tuning. We fine-tune using contrastive learning on each dataset. During training, we follow (Lan et al., 2021) regarding every utterance in the dialogue sense as a response and its previous utterances as a context. Our model is optimized by AdamW optimizer, and the linear learning ratio scheduler is used. We tuned the hypermeters of individual tasks on their development sets. For

[1]https://huggingface.co/bert-base-chinese
[2]https://huggingface.co/bert-base-uncased

Models	Ubuntu			E-commerce		
	R_{10}@1	R_{10}@2	R_{10}@5	R_{10}@1	R_{10}@2	R_{10}@5
TF-IDF (Lowe et al., 2015)	0.410	0.545	0.708	0.159	0.256	0.477
RNN (Lowe et al., 2015)	0.403	0.547	0.819	0.118	0.223	0.589
CNN (Kadlec et al., 2015)	0.549	0.684	0.896	0.328	0.515	0.792
LSTM (Kadlec et al., 2015)	0.638	0.784	0.949	0.365	0.536	0.828
SMN (Wu et al., 2017)	0.726	0.847	0.961	0.453	0.654	0.886
DUA (Zhang et al., 2018)	0.752	0.868	0.962	0.501	0.700	0.921
DAM (Zhou et al., 2018)	0.767	0.874	0.969	0.526	0.727	0.933
IOI (Tao et al., 2019)	0.796	0.894	0.974	0.563	0.768	0.950
ESIM (Chen and Wang, 2019)	0.796	0.894	0.975	0.570	0.767	0.948
MSN (Yuan et al., 2019)	0.800	0.899	0.978	0.606	0.770	0.937
RoBERTa-SS-DA (Lu et al., 2020)	0.826	0.909	0.978	0.627	0.835	0.980
BERT-VFT (Whang et al., 2020)	0.855	0.928	0.985	-	-	-
SA-BERT (Gu et al., 2020)	0.855	0.928	0.983	0.704	0.879	0.985
UMS$_{BERT+}$ **(Whang et al., 2021)**	0.875	0.942	0.988	0.764	0.905	0.986
BERT-SL (Xu et al., 2021)	0.884	0.946	0.990	0.776	0.919	0.991
DR-BERT (Lan et al., 2021) ♣	0.910	0.962	0.993	-	-	-
BERT-FP (Han et al., 2021)	<u>0.911</u>	0.962	**0.994**	0.870	0.956	0.993
BERT-TL (Zhang et al., 2022)	0.910	<u>0.962</u>	0.993	<u>0.927</u>	<u>0.974</u>	<u>0.997</u>
BERT$_{+CL}$	0.887	0.948	0.989	0.849	0.937	0.991
Dial-MAE	**0.918***	**0.964***	<u>0.993</u>	**0.930***	**0.977***	**0.997**
diff. %p	+3.1%	+2.4%	+0.4%	+8.1%	+4%	+0.6%

Table 2: Main experiment results on E-commerce Corpus and Ubuntu Corpus. **BERT**$_{+CL}$ means fine-tuning BERT using contrastive learning. The best score on a given dataset is marked in **bold**, and the second best is <u>underlined</u>. ♣ : According to the published code, for E-commerce, they adjusted the hyperparameters on the test set without cross-validation, we think the results are misleading, and this part has been removed. Two-tailed t-tests demonstrate statistically significant improvements of Dial-MAE over baselines ($* \leq 0.01$).

Ubuntu, we fine-tune for 5 epochs, the learning rate is set to 5e-5, and the batch size is set to 64. For E-commerce, we fine-tune for 2 epochs, the learning rate is set to 1e-4, and the batch size is set to 128. We set a widely used random seed as 42 for reproducibility.

4.4 Results and Discussions

We show the main results in Table 2, which shows that Dial-MAE achieves new state-of-the-art on the Ubuntu dataset and E-commerce dataset. We are able to achieve comparable performance to the state-of-the-art cross-encoders using a bi-encoder, and we have lower computational requirements compared to cross-encoders. Compared to BERT-FP, our model achieved an absolute improvement of 0.7%p in R_{10}@1 on the Ubuntu Corpus and 6%p in R_{10}@1 on the E-commerce. Compared to BERT-TL, our model achieves an absolute improvement of 0.8%p in R_{10}@1 on the Ubuntu Corpus and a slight improvement of 0.3%p in E-commerce. This suggests that our carefully tailored post-training method for the bi-encoder can achieve comparable performance to the complex-designed cross-encoder.

BERT$_{+CL}$ means fine-tuning BERT using contrastive learning. In comparison to BERT$_{+CL}$, Dial-MAE achieve an absolute improvement in R_{10}@1 by 3.1%p, 8.1%p on Ubuntu Corpus and E-commerce Corpus, respectively. This suggests that our custom post-training approach for dialogue retrieval models is effective. Aligning the features of the dialogue context and response during post-training enables improvements in contrastive fine-tuning. We believe the improvement comes from two aspects. On the one hand, the post-training method considers both the semantics of the tokens inside the context and its response. On the other hand, the asymmetric encoder-decoder structure with an asymmetric masking strategy facilitates post-training, which forces the encoder to learn better dialogue embeddings.

Models	Ubuntu			E-commerce		
	R_{10}@1	R_{10}@2	R_{10}@5	R_{10}@1	R_{10}@2	R_{10}@5
BERT$_{+CL}$	0.887	0.948	0.989	0.849	0.937	0.991
w/o Contrastive loss	0.205	0.341	0.647	0.141	0.242	0.466
Dial-MAE	**0.918**	**0.964**	**0.993**	**0.930**	**0.977**	**0.997**
w/o Contrastive loss	0.783	0.867	0.950	0.483	0.639	0.853

Table 3: Ablation results on the test sets of the two benchmarks.

4.5 Ablation Study

In this section, we analyze the experimental results to demonstrate the effectiveness of the proposed Dial-MAE method. In the following experimental analysis, due to high computing budgets, most experiments use Ubuntu Corpus.

The Impact of Auxilary Network. We remove the contrastive loss in BERT$_{+CL}$ and Dial-MAE, then evaluate their performance changes. As shown in Table 3, Dial-MAE achieved an absolute improvement in R_{10}@1 by 57.7%p, and 34.2%p on Ubuntu Corpus and E-commerce Corpus, respectively.

This suggests that our proposed post-training method effectively achieves the alignment of contextual representations, making the dialogue context more similar to the features of the response. We believe the gain comes from our auxiliary network helping the encoder aggregate dialogue contextual information. First, the encoder achieves feature alignment in the dialogue's contextual information by predicting the features of the correct response during the encoding of the context. Secondly, due to the small number of parameters of the decoder and the high mask rate on the decoder side, this will force the MLM task of the decoder to rely more on the dialogue context embedding output by the encoder. This enables the decoder to aggregate complex information about the dialogue context into a dense vector.

We then use contrastive learning to fine-tune the post-training models, and the performance of the models can be further improved. We also give the fine-tuning schedule on Ubuntu Corpus as shown in Figure 3, with the accuracy steadily improving as the training time increases, and Dial-MAE consistently outperforms BERT$_{+CL}$. This result shows that both the contrastive loss and the auxiliary MLM loss are crucial in our method. Both contrastive learning and our post-training method are effective in achieving dialogue context and response feature alignment, and their effects can be additive.

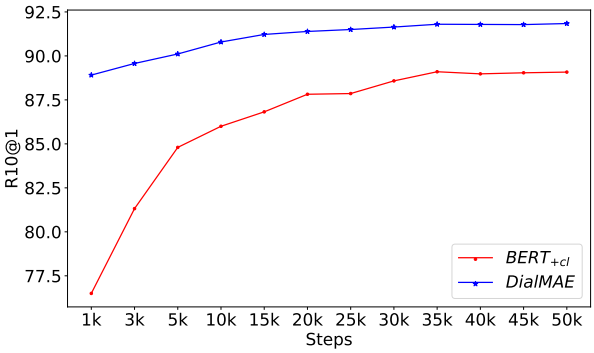

Figure 3: Fine-tuning schedules on the dev set of Ubuntu Corpus. A longer fine-tuning schedule gives a noticeable improvement. The performance of Dial-MAE is always better than BERT$_{+CL}$.

Enc	Dec	R_{10}@1	R_{10}@2	R_{10}@5
0.15	0	91.0	96.0	99.2
0.15	0.15	91.3	96.1	99.2
0.15	0.45	91.5	96.2	99.3
0.15	0.75	91.5	96.3	99.3
0.30	0.45	91.7	96.5	99.3
0.30	0.75	**91.9**	**96.5**	99.3
0.30	0.90	91.6	96.4	99.3
0.45	0.75	91.8	96.4	**99.4**

Table 4: Impact of mask rate on the dev set of Ubuntu Corpus. "Enc" denotes encoder, "Dec" denotes decoder. "Enc=0.15 Dec=0" means only using BERT's native MLM task without the decoder part.

Impact of Mask Rate. Wu et al. (2023) find that using a larger mask rate in both the encoder and decoder can enhance the performance of the contextual masking Auto-Encoder. As shown in Table 4, in our experiments, we find that an aggressive mask rate helps the learning of Dial-MAE. when the encoder mask rate equals 30%, and the decoder mask rate equals 75%, Dial-MAE achieves

the best performance. When the encoder mask rate stays below 30%, the performance of Dial-MAE improves as the decoder mask rate increases. When the encoder mask rate rises to 45%, Dial-MAE's performance declines slightly. We believe this is due to the encoder doesn't provide enough dialogue context semantic information when its mask rate is too high. In addition, from the experimental results, no matter what set of mask rates, Dial-MAE obviously exceeds the result of post-training for MLM tasks alone, which proves the robustness of Dial-MAE.

Impact of Decoder Layer Number. As shown in Figure 4, we further explore the impact of different decoder layer numbers on Dial-MAE performance. we find that using only one layer of the decoder yields the best results. Fewer decoder parameters can force the auxiliary MLM task to rely more on dialogue context embeddings output by the encoder. We believe that the more layers of the decoder, the stronger the decoding ability, and the decoder's dependence on context embedding will decrease, leading to insufficient constraints on encoder training. In general, no matter what set of layers, R@1 obviously exceeds the result of post-training for MLM tasks alone (Enc=0.15 Dec=0), as shown in Table 4, which proves the robustness of Dial-MAE.

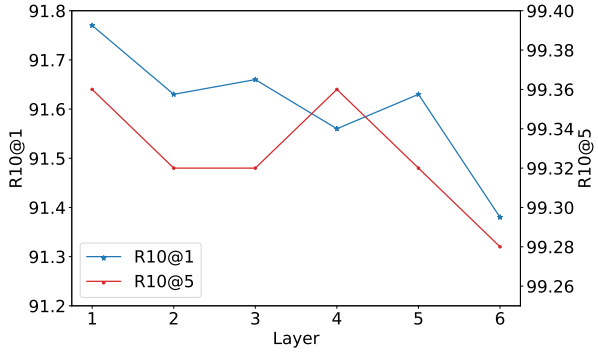

Figure 4: Impact of layer number on Ubuntu Corpus.

Compared with Dense Models. To further illustrate the effectiveness of our custom approach for bi-encoders in dialogue response selection, we compared it with state-of-the-art dense models in the Information Retrieval(IR) community. On the Ubuntu dataset, we fine-tune the dense models proposed by the IR community using contrastive learning, and the experimental results are shown in the table 5. During pre-training, the corpus of CoT-MAE(Wu et al., 2023) and RetroMAE(Xiao et al., 2022) contains an additional 3.2M documents dataset MS-MARCO(Nguyen et al., 2016) in addition to BooksCorpus and Wikipedia. However, our experimental results show that although the results of the three dense models have improved compared with BERT$_{+CL}$, they are still not as good as our proposed Dial-MAE. This shows that our proposed method is better suited for encoding dense vectors of dialogue than other dense models.

Models	R_{10}@1	R_{10}@2	R_{10}@5
BERT$_{+CL}$	89.2	95.1	99.2
Condenser(Gao and Callan, 2021)	89.4	95.4	99.1
RetroMAE(Xiao et al., 2022)	89.3	95.3	99.1
Cot-MAE(Wu et al., 2023)	89.8	95.9	99.2
Dial-MAE	**91.9**	**96.5**	**99.3**

Table 5: Comparison results of Dial-MAE and dense retrieval models on the Ubuntu dev set.

Qualitative Analysis. To qualitatively analyze our post-training method, as shown in Table 6, we provide the example. Dial-MAE can sort out the most appropriate response more accurately than BERT$_{+CL}$. The response sorted by BERT$_{+CL}$ have some token overlap with the dialogue context but are not semantically related. Compared with BERT$_{+CL}$, Dial-MAE can better understand dialogue semantics due to the joint modeling of context and response through post-training. This further demonstrates the effectiveness of our method.

Relevant	Model	Rank 1st response
		USER_A:I already have these disks in my system just want to migrate my current homefolder to the new drive. USER_B:Mount your new disk to a temporarely mount point move or copy your home folder to the new disc after that delete you old homedir than mount the new disk to _path_.
✗	BERT$_{+CL}$	USER_A: Do you use the ubuntu desktop or server i am using the desktop on my laptop i have only started with the server a little before _number_ came out maybe before april.
✓	Dial-MAE	USER_A: I presume i use gparted to get the mountpoints correct or am i wrong

Table 6: Examples of rank 1st response recalled by different models on the the Ubuntu Corpus.

5 Conclusion

In this paper, we propose a post-training method tailored for dialogue response, considering the semantics of dialogue context and its corresponding responses. Precisely, we leverage a shallow decoder to force the encoder output dialogue embeddings to be more expressive. Experimental results show that our post-training method leads to considerable improvements, achieving state-of-the-art on two benchmark datasets. We also demonstrate the effectiveness of Dial-MAE through ablation experiments. Specifically, both contrastive learning and our post-training method are effective in achieving dialogue context and response feature alignment, and their effects can be additive.

6 Limitations

Recently, generative conversational models based on large language models (LLMs) have demonstrated powerful performance. Despite the advantages of retrieval-based dialogue models in terms of computational cost and answer controllability, generative conversational systems based on LLMs surpass retrieval-based models in terms of answer diversity and flexibility. Furthermore, there has been much recent work exploring retrieval-augmented generation (RAG). In the future, we will further expand Dial-MAE to explore the effective integration with LLMs, using a dialogue response selection approach to attempt to address issues such as large model hallucinations and challenges related to knowledge updates. We hope that our work can also bring benefits to large language models.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Qian Chen and Wen Wang. 2019. Sequential attention-based network for noetic end-to-end response selection. *arXiv preprint arXiv:1901.02609*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 981–993. Association for Computational Linguistics.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Tevatron: An efficient and flexible toolkit for dense retrieval. *CoRR*, abs/2203.05765.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 2041–2044. ACM.

Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1549–1558. Association for Computational Linguistics.

Rudolf Kadlec, Martin Schmid, and Jan Kleindienst. 2015. Improved deep learning baselines for ubuntu corpus dialogs. *CoRR*, abs/1510.03753.

Tian Lan, Deng Cai, Yan Wang, Yixuan Su, Xian-Ling Mao, and Heyan Huang. 2021. Exploring dense retrieval for dialogue response selection. *CoRR*, abs/2110.06612.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 285–294. The Association for Computer Linguistics.

Junyu Lu, Xiancong Ren, Yazhou Ren, Ao Liu, and Zenglin Xu. 2020. Improving contextual language models for response retrieval in multi-turn conversation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1805–1808.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Zhenpeng Su, Xing Wu, Xue Bai, Zijia Lin, Hui Chen, Guiguang Ding, Wei Zhou, and Songlin Hu. 2023. Infoentropy loss to mitigate bias of learning difficulties for generative language models. *arXiv preprint arXiv:2310.19531*.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1–11, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and Heuiseok Lim. 2020. An effective domain adaptive post-training method for BERT in response selection. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 1585–1589. ISCA.

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14041–14049. AAAI Press.

Xing Wu, Guangyuan Ma, Meng Lin, Zijia Lin, Zhongyuan Wang, and Songlin Hu. 2023. Contextual masked auto-encoder for dense passage retrieval. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 4738–4746. AAAI Press.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 496–505. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. Retromae: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 538–548. Association for Computational Linguistics.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2021. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14158–14166. AAAI Press.

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

Wentao Zhang, Shuang Xu, and Haoran Huang. 2022. Two-level supervised contrastive learning for response selection in multi-turn dialogue. *CoRR*, abs/2203.00793.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on*

Computational Linguistics, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 372–381. The Association for Computational Linguistics.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia. Association for Computational Linguistics.