BERT-BC: A Unified Alignment and Interaction Model over Hierarchical BERT for Response Selection

Zhenfei Yang[1], **Beiming Yu**[1], **Yuan Cui**[2], **Shi Feng**[1*], **Daling Wang**[1], **Yifei Zhang**[1]
[1]Northeastern University, Shenyang, China
{zhenfyang, beiming403}@163.com, {fengshi, wangdaling, zhangyifei}@cse.neu.edu.cn
[2]Shenyang Polytechnic College, Shenyang, China
cyuan401@163.com

Abstract

Recently, we have witnessed a significant performance boosting for dialogue response selection task achieved by Cross-Encoder based models. However, such models directly feed the concatenation of context and response into the pre-trained model for interactive inference, ignoring the comprehensively independent representation modeling of context and response. Moreover, randomly sampling negative responses from other dialogue contexts is simplistic, and the learned models have poor generalization capability in realistic scenarios. In this paper, we propose a response selection model called BERT-BC that combines the representation-based Bi-Encoder and interaction-based Cross-Encoder. Three contrastive learning methods are devised for the Bi-Encoder to align context and response to obtain the better semantic representation. Meanwhile, according to the alignment difficulty of context and response semantics, the harder samples are dynamically selected from the same batch with negligible cost and sent to Cross-Encoder to enhance the model's interactive reasoning ability. Experimental results show that BERT-BC can achieve state-of-the-art performance on three benchmark datasets for multi-turn response selection.

Keywords: Response Selection, Bi-Encoder, Cross-Encoder, Contrastive Learning

1. Introduction

Dialogue response selection aims to find the best-matched response from a set of candidates given a dialogue context (Huang et al., 2020). In addition to the dialogue system, this technique can also be applied to in-context retrieval-augmented large language model to solve the problem of LLM hallucination (Borgeaud et al., 2022; Li et al., 2022; Ram et al., 2023). Therefore, response selection techniques have attracted widespread interest from industry and academia.

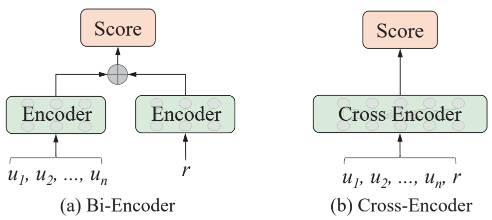


Figure 1: Matching paradigms of Bi-Encoder and Cross-Encoder. $u_1, u_2, ..., u_n$ denotes the context and r denotes the response.

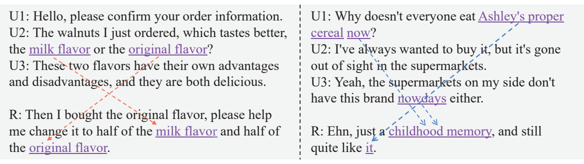
Pre-trained response selection model can be mainly divided into two approaches, namely representation-based Bi-Encoder model and interaction-based Cross-Encoder model (Thakur et al., 2021). Figure 1 illustrates the matching paradigms of Bi-Encoder and Cross-Encoder for response selection.

The Bi-Encoder model focuses on getting a better semantic representation of context and response and then employs a similarity function to obtain the matching score. The Bi-Encoder model is computationally fast with low cost and performs better for the **semantically related** samples (Figure 2 (a)) with high keyword co-occurrence and semantic approximation between context and response. However, the Bi-Encoder is restricted by the single vector representation, so it has to face the upper bound of representation capacity (Luan et al., 2021; Li et al., 2023). Researchers have studied post-interaction methods to exploit the potential of the Bi-Encoder. Poly-Encoder (Humeau et al., 2020) encodes the context into multiple potential vectors and uses a simple attention mechanism to post-interactively match the context with candidate responses. But this kind of work essentially designs a better similarity function for Bi-Encoder and cannot solve **conversationally related** samples that require conversational-level understanding and relational reasoning, such as Figure 2 (b). There is no explicit semantic similarity between the context and response of conversationally related samples, but they have a coherent relationship between dialogue contextual utterances. In Figure 2 (b), both U1 and U2 mention that they can't find a certain flavor of snack nowadays, and the response is implicitly related to the nostalgia for the taste of childhood.

In order to model conversational coherence relationship in multi-turn dialogue, recent research

[*]Corresponding author.

LREC-COLING 2024, pages 2253–2263



(a) semantically related sample (b) conversationally related sample

Figure 2: An illustration of the semantically related sample and conversationally related sample. The examples are drawn from Chinese Dataset E-Commerce (Zhang et al., 2018) and Douban (Wu et al., 2017), separately.

has focused on pre-trained Cross-Encoder models. Researchers devise auxiliary self-supervised tasks to learn the dependencies and coherence between utterances in multi-turn dialogue (Xu et al., 2021; Han et al., 2021). Although the above Cross-Encoder methods have achieved promising results, two shortcomings are retained. Firstly, the Cross-Encoder model concatenates all utterances in a dialogue using special tokens, leading to an incomprehensive representation of context and response as independent units. Secondly, most conventional methods leverage simple heuristics to construct negative samples by selecting responses from other conversations, which makes it challenging to distinguish stronger distractors in realistic scenarios (Li et al., 2019; Lin et al., 2020).

The evidence suggests that the adequate interaction between context and response is necessary for performance improvement; however, the performance can also benefit from comprehensively modeling context and response separately. Thus, in this paper we propose an end-to-end framework BERT-BC, which unifies the Bi-Encoder and Cross-Encoder models for response selection. In the proposed hierarchical BERT-like framework:

(i) The Bi-Encoder performs representation learning on context and response through multiple contrastive learning. By comparing context-response pair features within the same batch, it can enhance the representation ability of the encoder to model context and response separately. Furthermore, this comparison also assists the encoder extract critical features for distinguishing semantically related samples, thus reducing the learning difficulty of the high-layer Cross-Encoder and enabling the Cross-Encoder to focus on learning conversationally related samples that require conversation-level understanding and logical reasoning.

(ii) In order to improve the discriminative ability of the Cross-Encoder for conversationally related samples, we devise a negligible cost resampling strategy for hard negative samples, i.e., negative responses that are more difficult to be discriminated by Bi-Encoder in the same batch are selected as hard negative samples. Similar to curriculum learning (Bengio et al., 2009), the discriminative ability of Bi-Encoder is weaker in the early training stage, the model prefers to randomly select other responses within the same batch as negative samples. As the performance of the Bi-Encoder improves, the difficulty of the selected negative samples gradually increases. Notably, unlike previous curriculum learning approaches (Su et al., 2021), our model does not need training negative sample difficulty scoring function, which can significantly reduce the cost required for training.

Our main contributions are outlined below:

(i) We propose a pre-trained response selection model BERT-BC, which combines the representation-based Bi-Encoder and the interaction-based Cross-Encoder.

(ii) We devise a multiple contrastive learning method to enhance the Bi-Encoder's ability to learn semantically related samples and propose a hard negative resampling strategy to enhance the Cross-Encoder's interaction ability to learn conversationally related samples.

(iii) The empirical results show that our approach can achieve new state-of-the-art performance on three benchmark datasets. [1]

2. Related Work

2.1. Multi-turn Response Selection

In terms of modeling approaches, response selection models can be divided into representation-based and interaction-based methods, and in terms of encoder skeletons, they can be further divided into traditional encoding models and pre-trained models.

2.1.1. Traditional Models

The traditional response selection model mainly encodes the context and response by encoders such as RNN (Lowe et al., 2015), CNN (Pang et al., 2016), etc. Zhou et al. (2016) proposes a multi-view model which jointly models information from word view and utterance view. However, the representation-based approach ignores the conversational coherence between response and multi-turn context. With the emergence of attention mechanisms (Vaswani et al., 2017), researchers have proposed interaction-based approaches. Tao et al. (2019) performs shallow to deep matching between the context and the response through multiple interaction modules.

[1] https://github.com/thinkingmanyangyang/BERT-BC

2.1.2. Pre-trained Models

With the successful application of pre-trained models in many downstream tasks, BERT-based retrieval models have become the mainstream of research. Reimers and Gurevych (2019) employs BERT as the Bi-Encoder to represent the input text as a single vector for response selection. Khattab and Zaharia (2020) proposes a fine-grained alignment method that balances both performance and query speed. Although the above approaches optimize the Bi-Encoder to some extent, they still lack comprehensive contextual understanding and logical reasoning ability. Xu et al. (2021) devises four auxiliary tasks to endow the pre-trained Cross-Encoder models with coherence and consistency in dialogues.

Previous work mainly focuses on how to improve the performance of the Bi-Encoder or Cross-Encoder. We argue that Bi-Encoders and Cross-Encoders play distinct roles in the retrieval process, a facet that has seldom been explored in prior research. Diverging from previous methods, we achieve this by horizontally partitioning the BERT model into two separate components, working as the Bi-Encoder and the Cross-Encoder. We utilize contrastive learning and hard negative resampling to make Bi-Encoder and Cross-Encoder focus on different types of samples (semantically-related and conversationally-related), which serve complementary effects.

2.2. Contrastive Learning

Contrastive learning is a type of self-supervised learning that can effectively enhance the feature representation capability of the model. Li et al. (2021) proposes a multimodal pre-trained model to maximize the mutual information between text-image pairs by contrastive learning. Poddar et al. (2022) devises a ConMix method to construct positive and negative samples by mixing up the context token within the same batch, which enhances the robustness of dialogue representation.

The above work on contrastive learning focuses on the construction methods of positive and negative samples, however, cross-grained contrast has rarely been explored. In this paper, we introduce a multi-grained contrastive learning method to enhance the representation capability of the Bi-Encoder.

3. Method

3.1. Problem Formulation

Assume that given a conversation dataset consisting of a triplet $D = (c_i, r_i, y_i)_{(i=1)}^N$. $c_i = \{u_{i,1}, u_{i,2}, \ldots, u_{i,m}\}$ denotes dialogue history utterances; m indicates the number of utterances in the context; r_i denotes a candidate response; y_i is a label, when $y_i = 1$, r_i is a suitable response about c_i and $y_i = 0$ otherwise. The purpose of the response selection task is to learn a matching model $g(\cdot,\ \cdot)$. For a given context-response pair $(c_i,\ r_i)$, the matching scores of c_i and r_i are obtained by $g(c_i, r_i)$.

3.2. Model Architecture

The BERT-BC model follows the principle of "alignment first, interaction later", and the overall framework of the model is shown in Figure 3. The Bi-Encoder module is mainly used for alignment and the Cross-Encoder module is mainly used for interaction.

The Bi-Encoder module consists of a context encoder and a response encoder for feature extraction. The Cross-Encoder module conducts deep interactive reasoning on context and response features and computes the final match score. All the encoders are composed of Transformer blocks, where the weights of the context and response encoders are shared.

3.3. Context-Response Encoder

BERT-BC employs the low-layer BERT as the Bi-Encoder to encode the context and response. For the input context c_i, we concatenate all utterances into a sequence denoted as $[CLS]\,[BOC]\,u_{i,1}\,[EOU]\,u_{i,2}\,[EOU]\ldots u_{i,m}\,[EOU]$ $[SEP]$, $[CLS]$ is the classification token of BERT model, and $[SEP]$ is the segmentation token. $[BOC]$ represents the beginning of the context and $[EOU]$ represents the end of the utterance. For the input response, we add a $[BOR]$ token in front of the response to indicate the beginning of the response. We input the processed context and response to Bi-Encoder to learn the representation, $\{h_{cls}, h_{boc}, h_{c1} \ldots h_{sep}\}$ and $\{h_{bor}, h_{r1}, \ldots, h_{sep}\}$, where, h_{boc} and h_{bor} represents the global feature representation of context and response. $H_c = \{h_{c1} \ldots h_{sep}\}$ and $H_r = \{h_{r1}, \ldots, h_{sep}\}$ represents the token-level feature representation of context and response respectively.

3.4. Multiple Contrastive Learning

In order to effectively extract explicit semantic alignment information, we propose a multiple contrastive learning mechanism to optimize the context and response encoder. Specifically, our alignment method contains three contrastive learning objectives, i.e., CRA (Context Response Alignment), FGA (Fine-Grained Alignment), and CCL (Context Contrastive Learning).

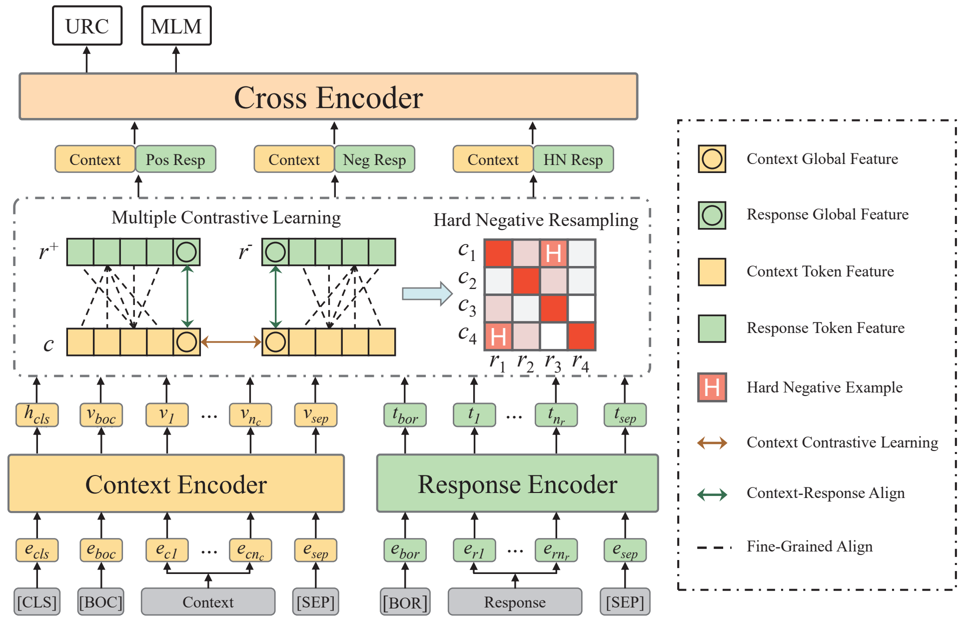


Figure 3: The overall framework of our BERT-BC model.

3.4.1. Context Response Alignment

CRA aims to pull the global representation of positive context-response pairs closer while pushing negative context-response pairs apart. In other words, CRA intends to maximize the lower bound of the global mutual information (MI) between the representation of context and response (Li et al., 2021). We use cosine similarity to compute the global alignment score S_g between the context and response.

$$S_g = \frac{g_c(h_{boc})^T g_r(h_{bor})}{||g_c(h_{boc})|| \cdot ||g_r(h_{bor})||} \tag{1}$$

where $S_g \in R^{B \times B}$, B denotes the batch size, g_c, g_r is a linear layer that map the global representation into a low-dimensional space representation.

3.4.2. Fine-Grained Alignment

Both context and response are composed of many tokens. For a token in context, not all tokens in response have a match, thus bringing in many noisy signals and unnecessary information for alignment (Yuan et al., 2019). We propose a fine-grained alignment mechanism (FGA) for learning the similarity at the token level.

First, we calculate the similarity matrix M^w between the context token feature H_c and response token feature H_r.

$$M^w = H_c^T H_r \tag{2}$$

where token similarity matrix $M^w \in R^{n_c \times n_r}$, n_c represents the token number in the context, n_r represents the token number in the response.

To adaptively adjust the importance of each token during the matching, we use the softmax function to measure the contribution of different tokens(α^c, α^r). Then, we use a weighted pooling function to obtain the final context-to-response FGA score sim_f^{c-r} and response-to-context FGA score sim_f^{r-c}:

$$sim_f^{c-r} = \sum_{i=1}^{n_c} \alpha_i^c max(M_{i,*}^w) \tag{3}$$

$$sim_f^{r-c} = \sum_{j=1}^{n_r} \alpha_j^r max(M_{*,j}^w) \tag{4}$$

$$sim_f = sim_f^{c-r} + sim_f^{r-c} \tag{5}$$

where fine-grained alignment score sim_f is a single real value, then we calculate the fine-grained matching matrix $S_f \in R^{B \times B}$ for each context and response within the same batch.

Finally, we sum the CRA score and the FGA score as the semantically related score S:

$$S = S_g + S_f \tag{6}$$

where semantically related score $S \in R^{B \times B}$.

We use InfoNCE loss to optimize the alignment objective:

$$L_{cr} = -E_{p(C,R)}\left[log \frac{\exp(S_{i,+}/\tau)}{\sum_{j=1}^{B} \exp(S_{i,j}/\tau)}\right] \tag{7}$$

$$L_{rc} = -E_{p(C,R)}\left[log \frac{\exp(S_{+,j}/\tau)}{\sum_{i=1}^{B} \exp(S_{i,j}/\tau)}\right] \tag{8}$$

$$L_a = L_{cr} + L_{rc} \tag{9}$$

where τ denotes temperature coefficient, L_{cr} denotes the context-to-response alignment loss, L_{rc} denotes the response-to-context alignment loss. L_a denotes the alignment loss.

3.4.3. Context Contrastive Learning

Since the context usually contains several utterances and the length is much longer than the response, we enhance the encoder's ability to represent the context by introducing CCL (Context Contrastive Learning). Given a set of a positive context-response pair (c_i, r_i^+) and a negative context-response pair (c_i, r_i^-) that have the same context and different responses, we follow Gao et al. (2021) by taking the two forward-propagated representations of the contexts c_i in the set as positive samples and using the other contexts within the same batch as negative samples. Unlike CRA and FGA, CCL mainly learns the differences between positive and negative samples among contexts. The CCL loss is denoted as L_c.

3.5. Hard Negative Resampling

In multi-turn dialogue response selection, if a negative context-response pair with high lexical overlap and semantic similarity, but without dialogue coherence and logic, this pair is easily misclassified as a positive sample by the Bi-Encoder. However, such matching clues can be learned more efficiently by the Cross-Encoder. In order to improve the ability of the Cross-Encoder to discriminate these conversationally unrelated samples, we propose a hard negative resampling strategy (HNR).

We use the semantically related score S of context and response in Equation 6 to find difficult samples within the same batch. For each context c_i within the same batch, we use the softmax value of the semantically related score S as the sampling probability $p_{i,j}$ of sampling to other negative responses r_j.

$$p_{i,j} = \begin{cases} \frac{exp(S_{i,j})}{\sum_{\hat{j} \in B} exp(S_{i,\hat{j}})}, \; j \neq i \\ 0, \; j = i \end{cases} \tag{10}$$

3.6. Context Response Matching

In order to fuse and interact context and response features to reason about the conversational coherence of dialogues, we adopt the high-layer BERT as the Cross-Encoder interaction model. After the context and response representation, Cross-Encoder takes concatenation of $\{h_{cls}, h_{boc}, h_{c1} \ldots h_{sep}\}$ and $\{h_{bor}, h_{r1}, \ldots, h_{sep}\}$ as input, using the hidden state of the encoded $[CLS]$ token as a joint representation of the input context-response pairs, and then feeds it into a fully connected classification layer to predict the matching probability $\phi(C, R)$. The ground-truth labels of the samples are $y^{(C,R)}$. The context-response matching loss L_{crm} is defined as:

$$L_{crm} = E_{p(C,R)} H\left(\phi(C,R),\; y^{(C,R)}\right) \tag{11}$$

where $H(\cdot,\cdot)$ is the cross-entropy loss function. We assume the samples constructed by hard negative resampling are labeled as 0. The loss L_{hm} of hard negative samples is calculated using the same method.

The overall training objective of our model is:

$$L = L_a + L_c + L_{crm} + L_{hm} \tag{12}$$

3.7. Dialogue Domain Pre-training

Previous studies on multi-turn response selection further reduce the adverse effects by designing self-supervised tasks related to dialogue features and post-training on a dialogue corpus (Xu et al., 2021). Following previous work (Han et al., 2021), we adopt a fine-grained dialogue domain pre-training approach (DDP). Specifically, the method splits all utterances in a dialogue into short context-response pairs to learn continuous relations and interactions at the utterance level. We train the three contrastive learning and the hard negative resampling together with the dialogue domain pre-training method and subsequently fine-tune them on the corresponding dataset.

4. Experiment

4.1. Datasets

We test our model on three widely used benchmark datasets, including Ubuntu Corpus V1 (Lowe et al., 2015), Douban Corpus (Wu et al., 2017), and the E-Commerce Corpus (Zhang et al., 2018). The statistics of three datasets are shown in Table 1.

Ubuntu Corpus: The Ubuntu Corpus V1 construct is based on log records of chats in Ubuntu forums, that focus on troubleshooting and technical support for the Ubuntu operating system.

Douban Corpus: The Douban corpus is an open-domain dataset crawled from the social networking service, Douban. It consists of conversations between two people that are longer than two turns.

E-Commerce corpus: The E-Commerce corpus is a multi-turn conversation in Chinese collected from Taobao. It contains real-world conversations between customers and customer service staff.

Dataset		Train	Valid	Test
Ubuntu	#pairs	1M	500K	500K
	pos:neg	1:1	1:9	1:9
Douban	#pairs	1M	50K	6670
	pos:neg	1:1	1:1	1.2:8.8
E-Commerce	#pairs	1M	10K	10K
	pos:neg	1:1	1:9	1:9

Table 1: Corpus statistics of datasets

Inspired by the previous work (Han et al., 2021), we reconstruct the pre-training data for three datasets using the same approach. Specifically, out of the one million triples in the training set of each benchmark, we use 500K triples with positive labels for construction.

4.2. Evaluation Metrics

Following previous studies (Yuan et al., 2019), we employ several retrieval metrics to evaluate our model. The recall ($R_{10}@k$) represents the probability that the correct response exists in the top k candidate responses out of the 10 candidate responses. Specifically, in the experiments, $R_{10}@1$, $R_{10}@2$, and $R_{10}@5$ are adopted. In addition to $R_{10}@k$, we also utilize MAP (mean average precision), MRR (mean reciprocal rank), and $P@1$ (precision at one) for the Douban corpus, since the Douban dataset may contain multiple positive responses from the same context.

4.3. Experiment setup

In this paper, we use AdamW optimizer to optimize the BERT-BC model, and the train batch size is set to 64, and the test batch size is set to 100. The maximum lengths of context and response are set to 190 and 70. The initial learning rates in the pre-training and fine-tuning stages are set to 2e-5, 5e-6, and gradually decays during the training process. The number of hard negative samples in HNR is set to 2. The BERT-BC model is trained on an A6000 GPU for 20 epochs. The layer of contrastive learning added to the model is 9, 6 and 9 on Ubuntu, Douban and E-Commerce, respectively.

4.4. Baseline Methods

We compare our proposed model BERT-BC with the following previous models.

Single-turn matching models: Lowe et al. (2015), Kadlec et al. (2015) proposed basic models based on RNN, CNN.

Multi-turn matching models: SMN (Wu et al., 2017) matches candidate responses and each utterance of the context interactively at multiple granularity. DAM (Zhou et al., 2018) computes matching between context and response by self-attention and cross-attention based on Transformer. MSN (Yuan et al., 2019) uses a multi-hop selector to filter out unnecessary information.

BERT-based models: BERT fine-tunes the response selection task on the pre-trained model. Poly-Encoder (Humeau et al., 2020) improves the accuracy of the Bi-Encoder by adding an attention post-interaction layer. UMS$_{BERT+}$ (Whang et al., 2021) devises three utterance manipulation strategies to learn the temporal dependencies between utterances. BERT-FP (Han et al., 2021) implements a post pre-training method including short context-response pair training and utterance relevance classification. BERT-TAP (Lin et al., 2022) highlights the significance of the NSP task for dialogue response selection pre-training. Uni-encoder (Song et al., 2023) concatenates all candidate responses to the context and jointly inputs them into the encoder.

4.5. Experimental Results

Table 2 shows the performance of baselines and the proposed BERT-BC model evaluated on three benchmark datasets. The Poly-Encoder, primarily composed of the Bi-Encoder component, performs better when applied to domain-specific E-commerce datasets characterized by a substantial correlation between context and response. In contrast, BERT-FP (Cross-Encoder) exhibits better results on more daily and open-domain dataset Douban. This observation partly indicates that the Bi-Encoder and Cross-Encoder excel in handling different types of response selection samples. In BERT-based models, our BERT-BC outperforms all other baseline models. Compared to the vanilla BERT model, BERT-BC achieves absolute improvements of 11.6%, 7.6% and 34.7% in $R_{10}@1$ on the Ubuntu Corpus V1, Douban Corpusand E-Commerce Corpus, respectively. Compared to the previous state-of-the-art model Uni-Encoder, BERT-BC consistently achieves notable performance improvements across all metrics. Compared to our BERT-BC model, although Uni-Encoder additionally introduces comparison information between responses by encoding all candidate responses simultaneously, it also leads to an increase in GPU memory and cannot handle scenarios where the number of candidate responses is not fixed.

In summary, the performance of response selection is significantly improved by the combination of Bi-Encoder and Cross-Encoder with multiple contrastive learning and hard negative resampling.

4.6. Ablation Study

To further investigate the role of modules in the BERT-BC model, we conduct extensive ablation studies on the E-Commerce dataset. Our Base model is a combination of the Bi-Encoder and Cross-Encoder models, initialized with the weights of BERT, without incorporating the contrastive learning and hard negative resampling strategy.

As presented in Table 3, the addition of CRA or FGA separately yields a marked improvement in the metrics to the Base model. This illustrates that the alignment of context and response improves the accuracy of the model in recognizing samples with high semantic relevance. When CRA and FGA are

Models	Ubuntu			Douban						E-commerce		
	R_{10}@1	R_{10}@2	R_{10}@5	MAP	MRR	P@1	R_{10}@1	R_{10}@2	R_{10}@5	R_{10}@1	R_{10}@2	R_{10}@5
TF-IDF	0.410	0.545	0.708	0.331	0.359	0.180	0.096	0.172	0.405	0.159	0.256	0.477
RNN	0.403	0.547	0.819	0.390	0.422	0.208	0.118	0.223	0.589	0.325	0.463	0.775
CNN	0.549	0.684	0.896	0.417	0.440	0.226	0.121	0.252	0.647	0.328	0.515	0.792
SMN	0.726	0.847	0.961	0.529	0.569	0.397	0.233	0.396	0.724	0.453	0.654	0.886
DAM	0.767	0.874	0.969	0.550	0.601	0.427	0.254	0.410	0.757	0.526	0.727	0.933
MSN	0.800	0.899	0.978	0.587	0.632	0.470	0.295	0.452	0.788	0.606	0.770	0.937
BERT	0.808	0.897	0.975	0.591	0.633	0.454	0.280	0.470	0.828	0.610	0.814	0.973
PolyEncoder+FP*	0.884	0.950	0.991	0.617	0.664	0.498	0.316	0.492	0.844	0.914	0.965	0.995
UMS$_{BERT+}$*	0.875	0.942	0.988	0.625	0.664	0.499	0.318	0.482	0.858	0.762	0.905	0.986
BERT-FP*	0.911	0.962	0.994	0.644	0.680	0.512	0.324	0.542	0.870	0.870	0.956	0.993
BERT-TAP*	0.912	0.966	0.994	0.644	0.684	0.511	0.323	0.548	0.853	0.926	0.980	0.998
Uni-Encoder*†	0.916	0.965	0.994	0.648	0.688	0.518	0.327	0.557	0.865	-	-	-
BERT-BC(ours)	**0.924**	**0.968**	**0.995**	**0.665**	**0.701**	**0.538**	**0.356**	**0.565**	**0.870**	**0.957**	**0.981**	**0.998**

Table 2: Evaluation results on Ubuntu, Douban, and E-Commerce datasets. * denotes pre-train on corresponding dialogue corpus. † denotes previous state-of-the-art model.

Methods						Metric	
Base	CRA	FGA	CCL	HNR	DDP	R_{10}@1/5/10	MAP
✓						0.641/0.824/0.970	0.777
✓	✓					0.826/0.934/0.989	0.898
✓		✓				0.837/0.935/0.985	0.903
✓	✓	✓				0.846/0.945/0.990	0.910
✓	✓	✓	✓			0.855/0.942/0.993	0.914
✓	✓	✓	✓	✓		0.905/0.960/0.998	0.943
✓	✓	✓	✓	✓	✓	**0.957/0.980/0.998**	**0.974**
✓	✓	✓	✓	✓	△	0.906/0.964/0.993	0.945

Table 3: Ablation study on E-Commerce, ✓ denotes the adoption of the different strategies and △ denotes the adoption of the LCCC dataset for DDP.

utilized in combination, they bring in better results than alone, demonstrating that the global and fine-grained alignment approaches are complementary. HNR effectively increases the recognition accuracy of the Cross-Encoder module for samples requiring conversational-level understanding and reasoning. By dialogue domain pre-training on the corresponding dialogue corpus, BERT-BC outperforms the current state-of-the-art model on the E-Commerce dataset. We also experimented with DDP on a different dataset, the LCCC corpus (Wang et al., 2020), and found that the advantage is not obvious. We argue that DDP primarily benefits from the domain knowledge from corresponding data.

5. Further Analysis

5.1. Impact of Contrastive Learning at Different Layer

We conduct experiments on contrastive learning at different BERT-BC layers to explore the impact of the proportion of Bi-Encoder and Cross-Encoder. The Base model initializes the BERT-BC with BERT as the checkpoint and uses only multiple contrastive learning. The results of the experiments on the three datasets are shown in Table 4. In this table, "3-layer" represents the model adding multiple contrastive learning at layer 3.

Dataset	Layer	R_{10}@1	R_{10}@2	R_{10}@5	MAP
Ubuntu	3	0.820	0.905	0.977	0.887
	6	0.827	0.911	0.980	0.891
	9	**0.835**	**0.917**	**0.981**	**0.897**
	11	0.778	0.887	0.973	0.861
Douban	3	0.296	0.480	0.822	0.604
	6	**0.299**	**0.490**	0.835	**0.610**
	9	0.283	0.484	**0.840**	0.601
	11	0.258	0.429	0.781	0.560
E-Commerce	3	0.752	0.891	0.983	0.850
	6	0.836	0.936	0.993	0.903
	9	**0.855**	**0.942**	**0.993**	**0.914**
	11	0.822	0.931	0.986	0.849

Table 4: Impact of contrastive learning at different layer

The results from the E-Commerce experiments in Table 4 demonstrate an increasing trend from 3 layers to 6 layers, and further to 9 layers. In contrast, the comparative results of 11-layer and 9-layer show a decrease, which suggests that both representation and interaction play an important role in the discrimination of response selection.

It is easy to notice that the Douban dataset achieves the best performance at 6-layer, while E-Commerce performs best at 9-layer, which suggests that the explicit alignment signal between context and response is less on the Douban dataset, and the model relies more on interaction and reasoning. This might be attributed to the fact that the Douban dataset primarily consists mainly of daily conversation posts on open-domain social networking sites, whereas Ubuntu and E-Commerce datasets have explicit themes. These evidences demonstrate that response selection tasks in different domains have different reliance on representation and interaction, and the model needs to adjust the proportion of representation and interaction according to the task characteristics in time.

5.2. Effectiveness of HNR Strategy

Table 5 compares the effects of the HNR strategy and different negative sampling strategies on the model performance. The Base model initializes BERT-BC with BERT as a checkpoint and without applying contrastive learning and other negative samples, while Base+MCL incorporates multiple contrastive learning on the Base model. Random indicates that negative responses are randomly selected as additional negative samples within the same batch. CUR employs curriculum learning to control the difficulty threshold of negative samples sampled during the training process. HNR directly samples the hardest samples within the same batch. HCL stands for Hierarchical Curriculum Learning proposed by Su et al. (2021).

Methods	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP
Base	0.641	0.824	0.970	0.777
MCL	0.855	0.942	**0.993**	0.914
MCL+Random	0.869	0.951	0.989	0.921
MCL+CUR	0.892	0.952	0.991	0.935
MCL+HNR	**0.905**	**0.960**	0.988	**0.943**
HCL	0.721	0.896	**0.993**	-

Table 5: Performance of different negative sampling strategies

As shown in Table 5, the results of Random improve 1.4% over $R_{10}@1$ of Base+MCL, which suggests that more negative samples can enhance the model's ability to discriminate wrong responses. Moreover, the HNR method achieves the best performance, exhibiting a 3.6% improvement in $R10@1$ compared to the ordinary Random strategy. This signifies that learning more challenging negative samples can effectively enhance the model's robustness. Meanwhile, the HNR does not need to train a difficult evaluator in advance as HCL does, which has a negligible computation cost.

5.3. Computational Cost Analysis

In addition to the analysis of model performance, we also compare the computational cost of BERT-BC with other paradigms (Cross-Encoder, Uni-Encoder, Poly-Encoder). We randomly select 1000 samples on Ubuntu V1 and vary the candidate size from 10, 20, 50, 100 to 200 for each context by randomly selecting additional responses from the corpus. The results are presented in Figure 4. BERT-BC demonstrates 2.4× faster inference speed compared to Cross-Encoder. As the candidate size increases, the advantages of BERT-BC become more pronounced.

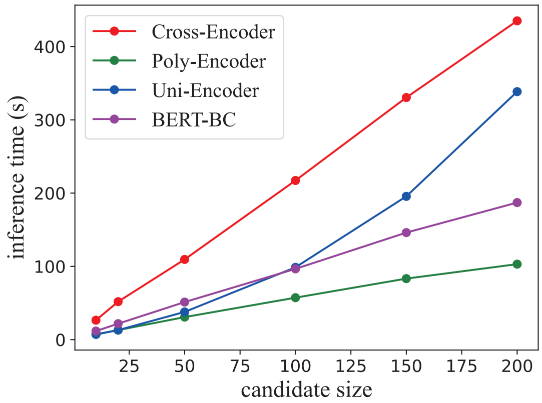


Figure 4: The inference time comparison.

5.4. Visualization of Alignment Matrix

In the above discussion, we assume that the alignment relationship between context and response can be learned through the contrastive learning mechanism.

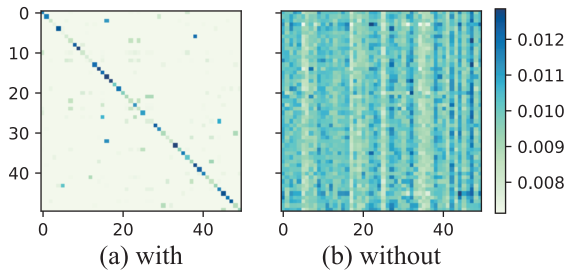


Figure 5: Visualization of the alignment matrix. (a) alignment matrix with contrastive learning, (b) alignment matrix without contrastive learning

As shown in Figure 5, we visualize the similarity matrix before and after using the contrastive learning mechanism. The horizontal and vertical axes of the graph represent context and response respectively, and darker colors represent higher alignment scores. It can be found that Figure 5 (a) shows a good alignment relationship between context and response, and Figure 5 (b) has almost no alignment between context and response, which demonstrates that semantically related relationship between context and response can be effectively learned through multiple contrastive learning. At the same time, Figure 5 (a) illustrates that not all context and response positive sample pairs can achieve good alignment. We conjecture that some samples cannot be discriminated only by simple semantic similarity and word overlap, which further illustrates the necessity of the Cross-Encoder reasoning model and the HNR strategy proposed in this paper.

6. Conclusion

In this paper, we propose a response selection model BERT-BC that combining Bi-Encoder and Cross-Encoder with three contrastive learning mechanisms and a hard-negative resampling strategy. The BERT-BC increases the Bi-encoder encoding ability of the model by multiple contrastive learning, which improves the recognition of semantically related samples. At the same time, the Cross-Encoder is able to focus on discriminating conversationally related samples through the hard negative resampling strategy. The experimental results demonstrate the superiority of our proposed BERT-BC model in the response selection task. In future work, we consider introducing common-sense knowledge in response selection.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62272092, No. 62172086) and the Fundamental Research Funds for the Central Universities of China (No. N2116008).

7. Bibliographical References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

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

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Rudolf Kadlec, Martin Schmid, and Jan Kleindienst. 2015. Improved deep learning baselines for ubuntu corpus dialogs. *arXiv preprint arXiv:1510.03753*.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*.

Jia Li, Chongyang Tao, Wei Wu, Yansong Feng, Dongyan Zhao, and Rui Yan. 2019. Sampling matters! an empirical study of negative sampling strategies for learning of matching models in retrieval-based dialogue systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1291–1296.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Xinze Li, Zhenghao Liu, Chenyan Xiong, Shi Yu, Yu Gu, Zhiyuan Liu, and Ge Yu. 2023. Structure-aware language model pretraining improves dense retrieval on structured data. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11560–11574.

Tzu-Hsiang Lin, Ta-Chung Chi, and Anna Rumshisky. 2022. On task-adaptive pretraining for dialogue response selection. *arXiv preprint arXiv:2210.04073*.

Zibo Lin, Deng Cai, Yan Wang, Xiaojiang Liu, Haitao Zheng, and Shuming Shi. 2020. The world is not binary: Learning to rank with grayscale data for dialogue response selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9220–9229.

Zhenghao Liu, Sen Mei, Chenyan Xiong, Xiaohua Li, Shi Yu, Zhiyuan Liu, Yu Gu, and Ge Yu. 2023. Text matching improves sequential recommendation by reducing popularity biases. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1534–1544.

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

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text matching as image recognition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016*, pages 2793–2799.

Lahari Poddar, Peiyao Wang, and Julia Reinspach. 2022. DialAug: Mixing up dialogue contexts in contrastive learning for robust conversational modeling. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 441–450.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992.

Chiyu Song, Hongliang He, Haofei Yu, Pengfei Fang, Leyang Cui, and Zhenzhong Lan. 2023. Uni-encoder: A fast and accurate response selection paradigm for generation-based dialogue systems. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6231–6244.

Yixuan Su, Deng Cai, Qingyu Zhou, Zibo Lin, Simon Baker, Yunbo Cao, Shuming Shi, Nigel Collier, and Yan Wang. 2021. Dialogue response selection with hierarchical curriculum learning. In *Proceedings of Annual Meeting of the Association for Computational Linguistics(Volume 1: Long Papers)*, pages 1740–1751.

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.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310.

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*, volume 30.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *Natural Language Processing and Chinese Computing: NLPCC 2020*, pages 91–103. Springer.

Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and Heuiseok Lim. 2020. An effective domain adaptive post-training method for BERT in response selection. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2020, pages 1585–1589.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14041–14049.

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

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14158–14166.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. 2017. Incorporating loose-structured knowledge into conversation modeling via recall-gate lstm. In *2017 international joint conference on neural networks (IJCNN)*, pages 3506–3513. IEEE.

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.

Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, and Nan Duan. 2022. Multi-view document representation learning for open-domain dense retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5990–6000.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. *arXiv preprint arXiv:1806.09102*.

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*, pages 372–381.

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.