Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Zhenfei Yang
Northeastern University
Beiming Yu
Northeastern University
Yuan Cui
Shenyang Polytechnic College
Daling Wang
Northeastern University
Ge Yu
Northeastern University
Shi Feng (fengshi@cse.neu.edu.cn)
Northeastern University

Research Article

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

Posted Date: July 20th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3156776/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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

Zhenfei Yang[1], Beiming Yu[1], Yuan Cui[2], Daling Wang[1], Ge Yu[1], Shi Feng[1*]

[1*]School of Computer Science and Engineering, Northeastern University, Shenyang China.
[2] Shenyang Polytechnic College, Shenyang China.

*Corresponding author(s). E-mail(s): fengshi@cse.neu.edu.cn;
Contributing authors: zhenfyang@163.com; beiming403@163.com;
cyuan401@163.com; wangdaling@cse.neu.edu.cn; yuge@cse.neu.edu.cn;

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 designed 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 minimal 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, Pretrained Model, Contrastive Learning, Bi-Encoder, Cross-Encoder

1 Introduction

Building a dialogue system that can interact naturally and continuously with humans is a popular research topic nowadays. As an important task in dialogue system, response selection as an important task in dialogue systems, it aims to find the best-matched response from a set of candidates given a dialogue context [1]. Due to the richness of the context-response corpus and the complexity of natural language generation, the retrieved responses are usually more fluent and informative. Therefore, response selection techniques have attracted widespread interest from industry and academia.

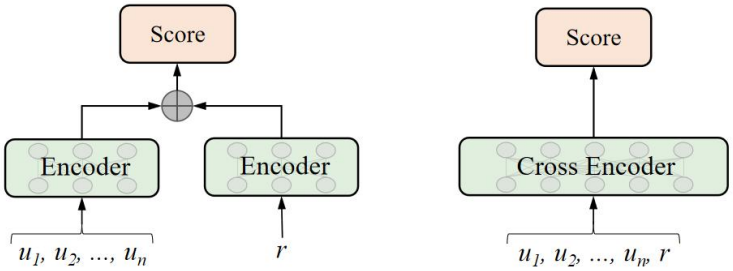
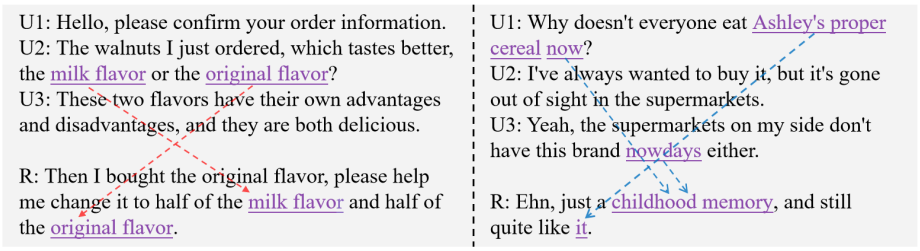


Fig. 1 Matching paradigms of Bi-Encoder and Cross-Encoder. $u_1, u_2, ..., u_n$ denotes the context and r denotes the response.

Pretrained response selection can be mainly divided into two approaches, namely representation-based Bi-Encoder model and interaction-based Cross-Encoder model [2]. 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 the context and the response, and then employs a similarity function to obtain the matching score. The Cross-Encoder model concatenates all utterances in a dialogue using special tokens, which are jointly encoded by a Cross-Encoder to capture the rich interactions between them.

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, Bi-Encoder is restricted by the single vector representation, so it has to face the upper bound of representation capacity [3]. Researchers have studied post-interaction methods to exploit the potential of Bi-Encoder. Khattab and Zaharia [4] proposes a ranking model ColBERT to model the relationship between context token and response token at a finer granularity. Poly-Encoder [5] encodes the context into multiple potential vectors and uses a simple attention mechanism to post-interactively match the context with candidate responses. Although the above approaches optimize the Bi-Encoder to some extent, they still lack comprehensive contextual understanding and logical reasoning ability, and perform poorly on **conversationally related** samples, such as Figure 2 (b). There is no explicit semantic similarity between the context and response

of conversationally related samples, but they have the coherent relationship between dialogue contextual utterances.



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

Fig. 2 An illustration of the semantically related sample and conversationally related sample. The examples are translated from Chinese. (a) In this sample, context and response have same topic, they are both talking about the flavor of walnuts and have the same keywords "plain flavor" and "milk flavor". (b) In this example, both U1 and U2 mention that they cannot find a certain flavor of snack nowadays, and the response is implicitly related to the nostalgia for the taste of childhood. In addition, "it" also refers to Ashley' proper cereal above.

In order to model conversational coherence relationship in multi-turn dialogue, recent research has focused on pre-trained Cross-Encoder models, by designing auxiliary self-supervised tasks related to dialogue features and post-training to learn knowledge from the dialogue domain [6, 7]. Han et al. [7] devises a fine-grained pretraining method BERT-FP to learn the dependencies and coherence between utterances in multi-turn dialogue. BERT-FP partitioned all utterances into multiple short context-response pairs and pretrained them within the domain using a sentence relationship classification task, which is the state-of-the-art method so far. Although the above approaches have achieved promising results, two shortcomings remain. Firstly, previous pretrained models for multi-turn response selection usually input all utterances in a conversation concatenated into the BERT model, leading to 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 conversational contexts. This negative sample selection strategy is straightforward, which makes it challenging to distinguish stronger distractors in realistic scenarios if training using such trivial negative samples [8, 9].

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. Specially, we split the pretrained BERT response selection model into two complementary modules, the lower-layer as Bi-Encoder and the higher-layer as Cross-Encoder.

In the proposed hierarchical BERT-like framework, (i) the low-layer Bi-Encoder performs representation learning on context and response, and optimizes the Bi-Encoder at a multi-grained level with three contrastive learning objectives. On the

one hand, by comparing context-response pairs in the same batch, the representation ability of the encoder to model context and response separately can be enhanced. On the other hand, by pulling closer the distance between context embedding and response embedding in the feature space, the encoder can extract critical features for distinguishing semantically related samples and reduce the learning difficulty of the upper-layer Cross-Encoder. (ii) The high-layer Cross-Encoder is mainly utilized to learn conversationally related samples that require conversation-level understanding and logical reasoning.

In order to improve the discriminative ability of the Cross-Encoder for conversationally related samples, we devise a minimal 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 [10], 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 [11], our model does not need training negative sample difficulty scoring function, which can significantly reduce the cost required for training. In addition, the proposed model can dynamically calculate difficulty scores for different samples based on the new knowledge gained in each iteration.

Our main contributions are summarized as follows:

(i) We propose a pretrained 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 by align context and response, 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(*i.e.*, Ubuntu, Douban, E-Commerce). [1].

2 Related Work

2.1 Multi-turn Response Selection

In terms of modeling approaches, retrieval-based models can be divided into representation-based and interaction-based methods, and in terms of encoder skeletons they can be further divided into traditional deep learning based encoding models and pretrained models.

2.1.1 Traditional Models

The traditional response selection model mainly encodes the context and response by encoders such as RNN [12], CNN [13], etc. Lowe et al. [14] proposes a representation based model to encode the context and response with RNN respectively and calculate

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

the similarity between them by a bilinear function. Zhou et al. [15] proposes a multi-view model which jointly models information from word sequence view and utterance sequence view by a deep neural network. However, the representation-based approach ignores the dependencies between response and multiple context utterances, including co-reference of dialogue elements and conversationally coherence between context. With the emergence of attention mechanisms [16–18], researchers have proposed interaction-based approaches. Zhou et al. [19] proposes a deep attention matching network that captures matching information between the context and the response from word-level to sentence-level through the stacked self-attention layer. Tao et al. [20] performes shallow to deep matching between the context and the response through multiple interaction modules. Tao et al. [21] studies multiple granularity representations (word, n-gram, utterance) for context-response matching and compared the performance of fusing to matching at early stage, intermediate stage or last stage. Yuan et al. [22] improves the performance of response selection by using a multi-hop selector to select dialogue context information.

2.1.2 Pretrained Models

With the successful application of pretrained models (BERT) in many downstream tasks, BERT-based retrieval model have become the mainstream of research. Reimers and Gurevych [23] employes BERT as a text encoder to encode the input as a single vector and explores the impact of different pooling methods on performance. Khattab and Zaharia [4] devises a fine-grained alignment method that balances both performance and query speed. Zhang et al. [24] improves the performance by representing documents as multiple latent semantic units corresponding to different latent queries for interaction. Humeau et al. [5] introduces an attention interaction layer after Bi-Encoder, which is the first effort to combine representation and interaction. However, these works essentially design a better similarity function for Bi-Encoder and cannot solve difficult samples that require conversational understanding and relational reasoning. Whang et al. [25] first applies BERT to interaction-based response selection and post-pretrain BERT on a dialogue corpus. Whang et al. [6] proposes three utterance manipulation strategy tasks to learn the sequential nature and dialogue coherence of multi-turn dialogue systems. Xu et al. [26] devises four auxiliary tasks to endow the pretrained models with coherence and consistency in dialogues, achieving more advanced results. Liu et al. [27] decouples the encoded dialogue representation to model the relative location information between speakers and the semantic information within the utterances through attention mask mechanism.

Previous work mainly focuses on how to improve the performance of Bi-encoder or Cross-Encoder. We argue that Bi-Encoder and Cross-Encoder play different roles in the retrieval process, which has been rarely explores in previous work. In this paper, we design a model that combines Bi-Encoder and Cross-Encoder. Different from previous work, we do not introduce additional parameters, simply by splitting the BERT model into two parts. 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. Contrastive learning learns a representation model by automatically constructing similar and dissimilar samples, allowing similar samples to be pushed closer in the projection space and dissimilar samples to be pushed farther away in the projection space. The concept of contrastive learning was first proposed in computer vision [28], with the successful application of contrastive learning in the computer vision, researchers also leverage the framework to the natural language processing tasks. SimCSE proposed by Gao et al. [29] applies contrastive learning to text representation by constructing positive samples by twice forward propagation with dropout. Yan et al. [30] increases data augmentation methods such as adversarial attack and token shuffle based on dropout. liang et al. [31] suggests a R-Drop method to obtain different feature representation of the same input by dropout twice for supervised learning, achieving a significant improvement in many downstream tasks. Li et al. [32] proposes a multimodal pretrained model to maximize the mutual information between text-image pairs by contrastive learning. Poddar et al. [33] devises a ConMix method to construct positive and negative samples by mixing up the context token within the same batch, which enhanced 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 previous work, which is also crucial for semantic alignment. In this paper, we introduce a multi-grained contrastive learning method to enhance the representation capability of Bi-Encoder by comparing the context-response pair in multiple views.

3 Preliminary

3.1 Problem Formulation

The dialogue usually consists of two parts, a context including multiple utterances and a response with one utterance. Response selection models typically use information retrieval and semantic matching techniques to compute a match between the dialogue context history and each response candidate from a large-scale corpus. The key issue of response selection is the deeply modeling of the input multiple conversational utterances and mining the semantic coherence relationships among them.

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 $g(c_i, r_i)$.

The traditional models such as RNN and CNN are limited by the encoder capability and lack of common sense knowledge for long document modeling, and thus cannot model utterances relationship and long distance dependency in multi-turn dialogues well. Recent models on response selection have mainly focused on Transformer-based

pretrained language models and so does ours. In this section, we provide a brief overview of the Transformer and Pretrained Language models.

3.2 Multi-Head Attention and Transformer

The attention mechanism in deep learning aids in making a neural network learn by focusing on the most important part of the current task goal among the large number of input information [34]. With the development and practicality of attention mechanisms, Vaswani et al. [18] proposes the multi-head self-attention mechanism. The self-attention mechanism has outstanding long-distance dependency feature capturing ability because any two words in a sentence can be related in one step during the computation.

The input of the self-attentive mechanism is the Query Q, Key K and Value V related to attention target. The attention score of Value is represented by the similarity between Query and Key, and then the attention score is used to weight the sum of Value to obtain the attention output of the task target.

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

where d denotes the dimension of Q and K.

In order to allow the model to learn relevant feature information from different representation subspaces, Vaswani et al. [18] extends the self-attention mechanism to a multi-headed attention mechanism.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_n)W^O \tag{2}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

Where W_i^Q, W_i^K and W_i^V are the parameter matrices.

Benefiting from the advantages of the multi-head attention mechanism, Vaswani et al. [18] further proposed a model Transformer that is entirely based on the multi-head attention mechanism. In the Transformer encoder, each layer consists of two sub-layers. The first layer is a Multi-Head Attention layer and the second layer is a Fully Connected Feedforward Neural Network FFN. Residual Connection is used between the two sub-layers and then layer normalization is performed.

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{4}$$

3.3 Pretrained Language Model

Since the generalization ability of small-scale datasets on deep neural network models is hardly guaranteed, the scale of datasets for natural language processing tasks limits the depth of neural network models to a large extent. However, the dataset scale of specific tasks in real applications is limited. To solve this problem, scholars have proposed pretrained language models [35–38], the idea of which is to first pretrain a task using a large-scale corpus and fine-tune on the specific task data.

BERT [39] is a bidirectional encoder representation model stacked by multiple Transformers, which is different from previous unidirectional language models. BERT uses Masked Language Model (MLM) and Next Sentence Prediction (NSP), two unsupervised tasks for pretraining and obtaining deep contextual representation. The input

of BERT is the concatenation of all sentence tokens and two special tokens. The two special tokens are the [CLS] token located at the beginning of the sequence to extract the global information of the input sequence, and the [SEP] token located at the end of each sentence to concatenate the two sentences. The input embedding of BERT consists of three components, including token embeddings, segment embeddings and position embeddings.

MLM is one of the pretraining tasks of BERT to obtain the deep bidirectional representation. MLM first randomly masks 15% of the input tokens with [MASK], then the masked tokens will be predicted by the output of BERT at the masked position. Another pretraining task of BERT is NSP to understand inter-sentence relations, the goal of the NSP task is to predict whether a sentence is the next sentence of a sentence. Based on the characteristics of multi-turn dialogue response selection [6, 7, 26], we choose BERT [39] to model multi-utterances relationship in dialogue and mitigate the impact caused by the small size of the training dataset.

4 Method

4.1 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.

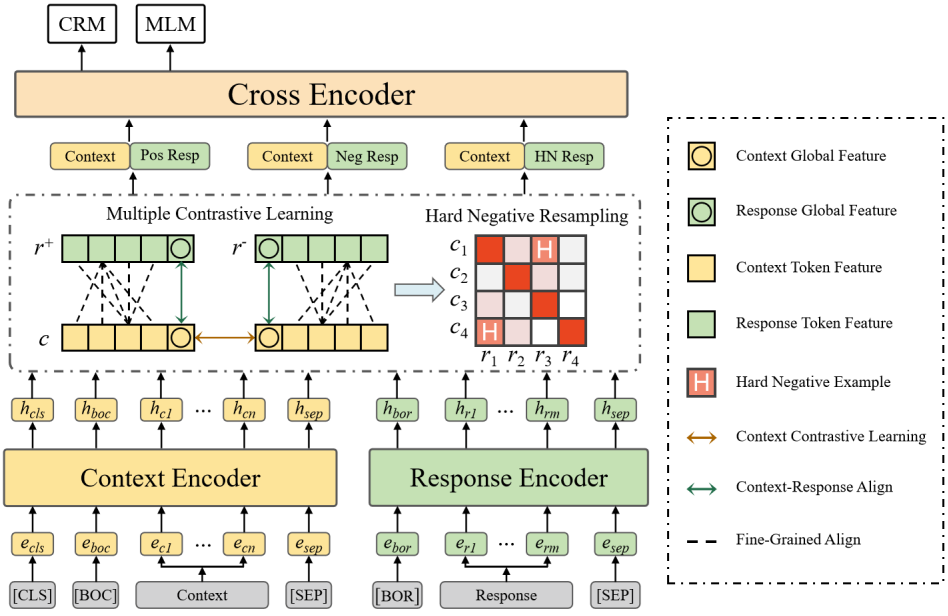


Fig. 3 The overall framework of our BERT-BC model.

The Bi-Encoder module consists of a context encoder and a response encoder for context and response feature extraction and enhances representation for response selection by multiple contrastive learning. 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. Following previous work [7], we leverage the BERT model to initialize the weights for the above network, and perform domain pretraining on the dialogue corpus.

4.2 Context-Response Encoder

BERT-BC employs the low-layer BERT as 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 r_i, it can be represented as $[BOR]r_i[SEP]$, and $[BOR]$ indicates 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} represents the global feature representation of context, and h_{bor} represents the global feature representation of 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.

4.3 Multiple Contrastive Learning

In order to effectively model context and response, and learn explicit semantic alignment information, we propose a multiple contrastive learning mechanism to optimize the context and response encoder. By pulling the distance between the context and response close in the feature space, features for discriminating semantically related samples are extracted. Thus, the subsequent Cross-Encoder could focus on reasoning about conversationally related samples that require deep semantic interactions. Specifically, our alignment method contains three contrastive learning objectives, i.e., CRA (Context Response Alignment), FGA (Fine-Grained Alignment), and CCL (Context Contrastive Learning).

4.3.1 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 [32]. We use cosine similarity to compute the global alignment score S_g between the context and response.

$$S_g = \frac{g_c(h_{boc})^T g_r\,(h_{bor})}{||g_c\,(h_{boc})\,|| \cdot ||g_r(h_{bor})||} \tag{5}$$

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 of context and response into a low-dimensional space (256-d) representation.

4.3.2 Fine-Grained Alignment

The above CRA aligns the context and response global representation, but the utterances in dialogue are usually relatively long, Both context and response are composed of many tokens. For a token in context, not all tokens in response have a match; some tokens are less matched, and vice versa, thus bringing in many noisy signals and unnecessary information [15, 22]. We propose a fine-grained alignment mechanism (FGA) for learning the similarity between context and response at the token level for context-response alignment.

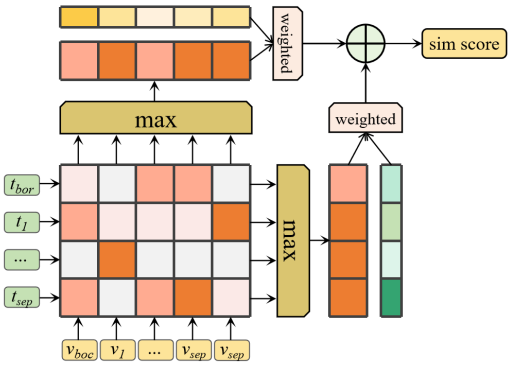


Fig. 4 Schematic overview of Fine-Grained Alignment.

First, we map the token-level features (H_c, H_r) to a lower dimensional space (256-d) using two linear transformation in order to obtain V_c and V_r.

$$V_c = f_c(H_c) \tag{6}$$

$$V_r = f_r(H_r) \tag{7}$$

where f_c, f_r is a linear layer.

After that, the similarity matrix M^w is calculated between the transformed context token feature V_c and response token feature V_r.

$$M^w = V_c^T V_r \tag{8}$$

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 weight size of each token during the matching process, we use a combination of linear layers and functions to measure the contribution of different features in similarity matching:

$$\alpha^c = softmax(l_c(V_c)) \tag{9}$$

$$\alpha^r = softmax(l_r(V_r)) \tag{10}$$

We use max pooling to select the highest matching response token to the current context token, and then obtain the final context-to-response fine-grained score sim_f^{c-r}

by averaging pooling. The same calculation method is employed for the response-to-context fine-grained alignment score sim_f^{r-c}:

$$sim_f^{c-r} = \sum_{i=1}^{n_c} \alpha_i^c max(M_{i,*}^w) \tag{11}$$

$$sim_f^{r-c} = \sum_{j=1}^{n_r} \alpha_j^r max(M_{*,j}^w) \tag{12}$$

$$sim_f = sim_f^{c-r} + sim_f^{r-c} \tag{13}$$

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 global alignment score (from CRA) and the fine-grained alignment score (from FGA) as the semantically related score S of context and response:

$$S = S_g + S_f \tag{14}$$

where semantically related score $S \in R^{B \times B}$, $S_{i,j}$ represents the semantical similarity between the i-th context and the j-th response within the same batch.

We use InfoNCE Loss to optimize the above two objectives [28]:

$$L_{cr} = -E_{p(C,R)} \left[log \frac{\exp(S_{i,+}/\tau)}{\sum_{j=1}^{B} \exp(S_{i,j}/\tau)} \right] \tag{15}$$

$$L_{rc} = -E_{p(C,R)} \left[log \frac{\exp(S_{+,j}/\tau)}{\sum_{i=1}^{B} \exp(S_{i,j}/\tau)} \right] \tag{16}$$

$$L_a = L_{cr} + L_{rc} \tag{17}$$

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

4.3.3 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. [29] 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, which mainly focus on the comparison with context and response, CCL mainly learns the differences between positive and negative samples among contexts to enhance the consistency of the same context across data-enhanced views, while allowing a more uniform distribution of the context's embedding space. The final calculated CCL loss is denoted as L_c.

4.4 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 consistency, this pair is easily misclassified as a positive sample by the Bi-Encoder.

However, such matching clues can be learned more easily by 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) with minimal computational cost.

We use the semantically related score S of context and response in Equation 14 to find difficult samples within the same batch, where the negative response with higher similarity to context has a higher chance of being sampled. 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.

$$p_{i,j} = \begin{cases} \frac{exp(S_{i,j})}{\sum_{j \in B} exp(S_{i,j})}, & j \neq i \\ 0, & j = i \end{cases} \tag{18}$$

For each training step, we construct a batch of difficult negative response samples which are fed into Cross-Encoder to obtain the hard negative resampling matching loss L_{hm}. Details of our learning framework are presented in Algorithm 1.

Algorithm 1: Hard Negative Response Resampling

Input: Dataset $D = (c_i, r_i, y_i)_{(i=1)}^{N}$, model trainer T, that takes batches of training data as input to optimize the Cross-Encoder model, Batch size is B

```
Initialize the parameter set and prepare the dataloader;
for train step t=1, ... do
    Encode context and response via Bi-Encoder;
    Calculate semantically alignment scores S by contrastive learning;
    Calculate the sampling probability p from semantically related score S;
    for i ← 1 to B do
        Sample negative responses r_i^-, from all responses in same batch, where
         r_i^- ≠ r_i^+;
        Concat c_i and r_i^- and feed into Cross-Encoder;
        Calculate the match probability of this context-response sample pair;
    end
    Calculate the matching loss L_hm for the hard negative sample in this batch;
    Invoke the trainer T, using matching loss to optimize the model;
end
```

Output: enhanced Cross-Encoder Model

4.5 Context Response Matching

In order to fuse context and response features for interactive reasoning about the conversational coherence of dialogues, we adopt the high-layer BERT as Cross-Encoder interaction model, which is widely used in response selection studies. Given a context-response pair, the model predicts whether they match (positive or negative), which can be considered as a binary classification problem. After the context and response representation, Cross-Encoder takes concatenation of $\{h_{cls}, h_{boc}, h_{c1} \dots h_{sep}\}$ and $\{h_{bor}, h_{r1}, \dots, h_{sep}\}$ 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 $y^{(C,R)}$. We assume the samples constructed by hard negative resampling are labeled as 0. The context-response matching loss L_{crm} is defined as:

$$L_{crm} = E_{p(C,R)} H\left(\phi(C, R),\ y^{(C,R)}\right) \tag{19}$$

where $H(\cdot, \cdot)$ is the cross-entropy loss function. The loss L_{hm} of hard negative samples is calculated using the same method.

The overall training objective of our model is:

$$L = L_a + L_c + L_{crm} + L_{hm} \tag{20}$$

4.6 Dialogue Domain Pretraining

Since most pretrained language models (PrLMs) are trained on general corpora (e.g., Wikipedia), direct fine-tuning of these models is usually less effective if domain transfer exists. Therefore, existing 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 corpus of dialogue domains [6, 7, 26]. In this paper, we adopt a fine-grained dialogue domain pretraining 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. In addition, a new training target utterance relevance classification is leveraged to distinguish whether the negative samples of responses are from the same conversational context or other conversational contexts, and to improve the model's understanding of semantic relevance and consistency among conversational utterances. We train the three contrastive learning and the hard negative resampling together with the dialogue domain pretraining method, and subsequently fine-tune them on the corresponding dataset. The framework for dialogue domain pretraining is shown in Figure 5.

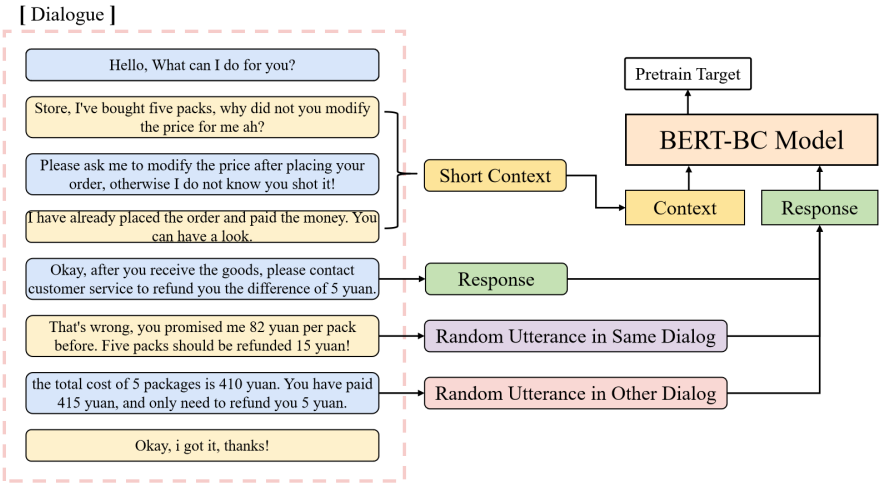


Fig. 5 The overall framework of dialogue domain pretraining

5 Experiment

5.1 Datasets

We test our model on three widely used benchmark datasets, including Ubuntu Corpus V1 [40], Douban Corpus [41], and the E-Commerce Corpus [42]. The statistics of these three datasets are shown in Table 1.

- **Ubuntu Corpus:** The Ubuntu Corpus V1 [40] is a large multi-turn conversation corpus constructed based on log records of chats in Ubuntu forums, a dataset that focuses on troubleshooting and technical support for the Ubuntu operating system. In our study, we use the pre-processed data from Xu et al. [26], where numbers, URLs, and paths are replaced with special placeholders.
- **Douban Corpus:** The Douban corpus [41] is an open domain dataset crawled from Douban, a popular social networking service in China. It consists of conversations between two people that are longer than two turns.
- **E-commerce corpus:** The E-commerce corpus [42] is a multi-turn conversation in Chinese collected from Taobao, the largest e-commerce platform in China. It contains real-world conversations between customers and customer service staff. It includes several kinds of conversations between customers and customer service staff (e.g., commodity consultations, recommendations and negotiations).

Table 1 Corpus statistics of response selection datasets

Dataset	Ubuntu			Douban			E-Commerce		
	Train	Val	Test	Train	Val	Test	Train	Val	Test
# pairs	1M	500K	500K	1M	50K	6670	1M	10K	10K
pos:neg	1:1	1:9	1:9	1:1	1:1	1.2:8.8	1:1	1:9	1:9
# avg turns	10.13	10.11	10.11	6.69	6.75	6.45	5.51	5.48	5.64

Inspired by the previous work [7], we reconstruct the pretraining 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. Based on creating multiple short context response pairs from a multi-turn dialogue, we finally build 12M, 9M and 6M sub-context response pairs for Ubuntu corpus, Douban corpus and E-commerce corpus respectively.

5.2 Evaluation Metrics

Following previous studies [22, 25], 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.

5.3 Experiment setup

In this paper, we use AdamW optimizer to optimize the BERT-BC model, and the train batch size is set to 96, and the test batch size is set to 100. The maximum number of context token is set to 190 and the maximum number of candidate response token is set to 70. The initial learning rate in the pretraining and fine-tuning stages is set to 1e-5, and gradually decays during the training process. The hidden layer dimension of BERT is set to 768, and the hidden dimension of global contrastive learning and fine-grained contrastive learning is set to 256. The number of hard negative samples in HNR is set to 2. To speed up the model convergence, DDP uses the checkpoint provided by Han et al. [7] as the initial checkpoint. The temperature coefficient of contrastive learning is set to 0.05. The BERT-BC model is trained on an A6000 graphics card for 50 epochs. The layer of contrastive learning added to the model is 9, 6 and 9 on Ubuntu, Douban and E-Commerce, respectively. Dialogue domain pretraining is performed on Ubuntu, Douban and E-Commerce datasets for 3, 5 and 10 epochs, respectively.

5.4 Baseline Methods

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

- **Single-turn matching models:** Lowe et al. [40], Kadlec et al. [43] proposed basic models based on RNN, CNN and LSTM.
- **Multi-turn matching models:** SMN [41] matches candidate responses and each utterance of the context interactively at multiple granularities (words and sentences). DAM [19] computes matching features between context and response by self-attention and cross-attention based on the Transformer model. DUA [42] represents the previous utterances into a fine-grained context representation by using a deep aggregation model. MSN [22] uses a multi-hop selector to filter out unnecessary information and matched the filtered contexts with candidate responses to obtain a match score.
- **Pretrained Models:** BERT concatenates contexts and candidate responses by special tokens, feeding them into the BERT model for fine-tuning. BERT-DPT [25] is a domain post training (DPT) approach using BERT pretraining task on a conversational corpus. SA-BERT [44] integrated speaker-aware embeddings into the BERT model; UMS_{BERT+} [6] devises three utterance manipulation strategies (i.e., utterance insertion, deletion, and search) to learn the temporal dependencies between utterances of a conversation. BERT-SL [26] implements four self-supervised auxiliary tasks based on the characteristics of dialogue to learn potential signals in dialogue. BERT-FGC [45] introduces a fine-grained contrast learning method to construct positive samples by data augmentation method to compare the same context, different response conversation pairs. BERT-TACM [46] improves the BERT model by masking some of the input sequences in the Transformer layer and using a better contextual modeling approach. BERT-FP [7] implements a post pretraining method including short context-response pair training and utterance relevance classification. BERT-TAP [47] explores the source of improvements from recent task-adaptive pre-training methods and highlights the criticality of the NSP task for dialogue response selection pretraining.

5.5 Experimental Results

Table 2 shows the performance of baselines and proposed BERT-BC model evaluated on the Ubuntu and E-Commerce datasets.

Models	Ubuntu			E-commerce		
	R_{10}@1	R_{10}@2	R_{10}@5	R_{10}@1	R_{10}@2	R_{10}@5
TF-IDF [40]	0.410	0.545	0.708	0.159	0.256	0.477
RNN [40]	0.403	0.547	0.819	0.325	0.463	0.775
CNN [43]	0.549	0.684	0.896	0.328	0.515	0.792
LSTM [43]	0.638	0.784	0.949	0.365	0.536	0.828
SMN [41]	0.726	0.847	0.961	0.453	0.654	0.886
DUA [42]	0.752	0.868	0.962	0.501	0.700	0.921
DAM [19]	0.767	0.874	0.969	0.526	0.727	0.933
MSN [22]	0.800	0.899	0.978	0.606	0.770	0.937
BERT [44]	0.808	0.897	0.975	0.610	0.814	0.973
BERT-DPT* [25]	0.851	0.924	0.984	-	-	-
SA-BERT* [44]	0.855	0.928	0.983	0.704	0.879	0.985
UMS$_{BERT+}$* [6]	0.875	0.942	0.988	0.762	0.905	0.986
BERT-SL* [26]	0.884	0.946	0.990	0.776	0.919	0.991
BERT-FGC* [45]	0.886	0.948	0.990	-	-	-
BERT-TACM* [46]	0.879	0.944	0.988	0.856	0.946	0.988
BERT-FP* [7]	0.911	0.962	0.994	0.870	0.956	0.993
BERT-TAP* [47]	0.912	0.966	0.994	0.926	0.980	0.998
BERT-BC* (ours)	**0.924**	**0.968**	**0.995**	**0.957**	**0.981**	**0.998**

Table 2 Evaluation results on Ubuntu and E-Commerce. * denotes pretrain on dialogue corpus.

The first group of results and the second group of results in Table 2 are reported for the single-turn matching model and the multi-turn matching model. It is evident that the multi-turn dialogue response selection models consistently outperform the single-turn response selection models. This indicates that incorporating multi-turn dialogue history information provides valuable benefits for the response selection task. In addition, the MSN model and DAM model based on Transformer encoder generally outperform the SMN model and DUA model based on traditional encoder, which suggests that the multi-head attention mechanism can alleviate the long-distance dependency problem in long text modeling. The BERT-BC model also adopts the multi-head attention based Transformer in modeling the context and response process. Comparing the Transformer-based DAM model and MSN model, DAM achieves more competitive performance, which can be attributed to that the context information cropping selection module helps the model to filter the useless information in selecting more appropriate responses. Our BERT-BC model also adopts the similar idea to align the context and response by a fine-grained approach, and uses feature weighting to select the context information.

The third group of Table 2 demonstrates the performance of the pretrained model on the response selection task. The vanilla BERT model also outperforms the traditional multi-turn response selection model across the board, indicating that the linguistic knowledge learned from large-scale text can better model the multi utterance relationship in multi-turn dialogue. Furthermore, the BERT-DPT model improves 43% over the vanilla BERT model on R_{10}@1, highlighting the significance of domain domain pretraining and fine-tuning on dialogue corpus in enhancing the performance of pretrained models on downstream tasks. SA-BERT introduces speaker-aware embedding on the basis of BERT-DPT to emphasize different dialogue roles. However, it does not achieve visible improvement compared with the ordinary BERT-DPT model, which suggests that the BERT modeling input can already adequately express the difference in dialogue speaker. The results demonstrate that BERT-BC outperforms all other baseline models. Compared to the vanilla BERT model, BERT-BC achieves absolute improvements of 10.3% and 34.7% in R_{10}@1 on the Ubuntu Corpus V1 and E-Commerce Corpus, respectively. Compared to the previous state-of-the-art model BERT-TAP, BERT-BC consistently achieves notable performance improvements across all metrics.

Models	Douban					
	MAP	MRR	P@1	R_{10}@1	R_{10}@2	R_{10}@5
TF-IDF [40]	0.331	0.359	0.180	0.096	0.172	0.405
RNN [40]	0.390	0.422	0.208	0.118	0.223	0.589
CNN [43]	0.417	0.440	0.226	0.121	0.252	0.647
LSTM [43]	0.485	0.537	0.320	0.187	0.343	0.720
SMN [41]	0.529	0.569	0.397	0.233	0.396	0.724
DUA [42]	0.551	0.599	0.421	0.243	0.421	0.780
DAM [19]	0.550	0.601	0.427	0.254	0.410	0.757
MSN [22]	0.587	0.632	0.470	0.295	0.452	0.788
BERT [44]	0.591	0.633	0.454	0.280	0.470	0.828
SA-BERT* [44]	0.619	0.659	0.496	0.313	0.481	0.847
UMS$_{BERT+}$* [6]	0.625	0.664	0.499	0.318	0.482	0.858
BERT-FGC* [45]	0.627	0.670	0.500	0.326	0.512	0.869
BERT-TACM* [46]	0.636	0.673	0.502	0.320	0.526	0.856
BERT-FP* [7]	0.644	0.680	0.512	0.324	0.542	0.870
BERT-TAP* [47]	0.644	0.684	0.511	0.323	0.548	0.853
BERT-BC* (ours)	**0.667**	**0.710**	**0.547**	**0.356**	**0.575**	**0.873**

Table 3 Evaluation results on Douban dataset. * denotes pretrain on dialogue corpus.

Table 3 presents the performance of the BERT-BC model evaluated on the Douban dataset. As anticipated, our proposed BERT-BC model outperforms the other baseline models obviously, achieving absolute improvements of 2.3%, 3.0%, and 3.2% in MAP, MRR, and R_{10}@1 compared to the previous best model BERT-FP. Furthermore, when comparing Table 2 and Table 3, we observe that the retrieval performance of the

Douban dataset is generally lower than that of Ubuntu and E-Commerce dataset. This discrepancy may be attributed to the construction methodology of the Douban dataset. Douban primarily consists of postings and replies from social networking sites, representing a more open domain, whereas Ubuntu and E-Commerce datasets originate from Ubuntu system forum logs and e-commerce customer service chats, respectively, focusing on a single domain.

These results demonstrate that the three contrastive learning mechanisms proposed in this paper can effectively enhance the representation ability of the encoder, and the semantic relevance and consistency of the dialogue can be learned more effectively through the method of first aligning and then interacting. 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 strategy. The performance of our BERT-BC achieves the current best performance (the state-of-art) on all three datasets.

5.6 Ablation Study

In this section, we conduct a series of ablation experiments on the E-Commerce dataset to examine the impact of the three contrast learning methods and hard negative resampling strategies proposed in this paper. 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. Subsequently, our method is gradually applied to the Base model. CRA, FGA, and CCL represent the three contrastive learning methods employed to enhance Bi-Encoder. HNR represents the hard negative sample resampling strategy. DDP refers to the dialogue domain post pretraining and fine-tuning.

Methods						Metric			
Base	CRA	FGA	CCL	HNR	DDP	R_{10}@1	R_{10}@2	R_{10}@5	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**

Table 4 Ablation study on E-commerce

As presented in Table 4, the addition of CRA or FGA separately yields a significant improvement in the metrics to the Base model. This illustrates that better representation can be learned through the alignment of context and response, improving the accuracy of the model in recognizing samples with high semantic relevance. When CRA and FGA are utilized in combination, they bring in better results than alone, demonstrating that the global and fine-grained alignment approach are complementary.

The inclusion of CCL also leads to an improvement in model performance, indicating that CCL enhances the model's ability to encode long text sequences within the context. Without training an additional sample difficulty evaluator, HNR effectively increases the recognition accuracy of the Cross-Encoder module for samples requiring conversational-level understanding and reasoning, achieving results that exceed the dialogue domain pretraining method, BERT-FP. Finally, by dialogue domain pretraining on a dialogue corpus, BERT-BC outperforms the current state-of-the-art model BERT-TAP on the E-Commerce dataset.

6 Further Analysis

6.1 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 (i.e., the percentage of representation and interaction in the model) on model performance. The Base model initializes the BERT-BC with BERT as the checkpoint and without applying multiple contrastive learning and HNR hard negative resampling strategy. The results of the experiments on the three dataset are shown in Table 5. In this table, "3-layer" represents the model adding multiple contrastive learning at layer 3, *i.e.* model uses a 3 layer Bi-Encoder and a 9 layer Cross-Encoder.

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

Table 5 Impact of contrastive learning at different layer

The results from the E-Commerce experiments in Table 5 demonstrate an increasing trend from 3 layers to 6 layers, and further to 9 layers. This suggests that better representation of context and response leads to improved recall performance. In contrast, the comparative results of 11-layer and 9-layer show a decrease, which suggests that too few context and response interactions can also adversely affect the performance of the model, and both representation and interaction play an important role

in the discrimination of response selection. Meanwhile, the result of 11-layer reaches 0.822, which illustrates that BERT-BC can reach a competitive performance on the E-commerce dataset using only few interactions.

It is easy to notice from observing the layer experiments in Table 5 on Douban dataset that the results for layers 6, 9 and 11 show a decreasing trend in $R10@1$ and MAP, which suggests that the explicit alignment signal between context and response is less on Douban dataset, and the model relies more on the interaction and reasoning with the text. Finally, the 6-layer MAP achieves the highest value, while E-Commerce performs best by applying contrastive learning at 9-layer. 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.

6.2 Effectiveness of Hard Negative Resample Strategy

Table 6 compares the effects of the HNR method 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 samples 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, and continuously increases the difficulty of negative samples 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. [11].

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

Table 6 Performance of different negative sampling strategies

As shown in Table 6, the results of Random improve 1.4% over $R_{10}@1$ of Base+MCL, which suggests that more negative samples based on the existing can enhance the ability of the model 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 difficulty evaluator in advance as HCL does, which has a minimal computation cost, and can dynamically select the appropriate negative samples according to the learning state of the model.

6.3 Performance Impact of Temperature Parameters

The temperature parameter is a important hyperparameter in contrastive learning, and the correct setting for temperature coefficient can better improve the efficiency of the model in learning negative samples. We explore the effect of different temperature coefficient on our proposed multiple contrastive learning method on the E-commerce dataset, and the results are shown in Table 7. The experimental results show that either too low or too high temperature coefficients lead to the degradation of the model performance. 0.05 is an optimal hyperparameter.

Temp	R_{10}@1	R_{10}@2	R_{10}@5	MAP
0.1	0.843	0.942	0.989	0.908
0.07	0.850	0.940	0.989	0.910
0.05	**0.855**	0.942	**0.993**	**0.914**
0.1	0.847	**0.948**	0.989	0.911

Table 7 Performance at different temperature

6.4 Visualization of the Alignment Matrix

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

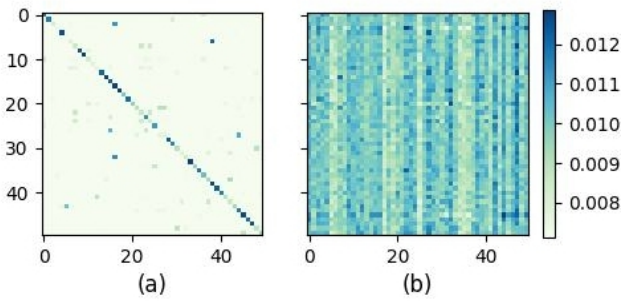


Fig. 6 Visualization of the alignment matrix. (a) alignment matrix with contrastive learning, (b) alignment matrix without contrastive learning

As shown in the Figure 6, we visualize the similarity matrix before and after using the comparative learning mechanism. It can be found that Figure 6 (a) shows a good alignment relationship between context and response, and Figure 6 (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 6 (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 hard negative resampling strategy and the Cross-Encoder reasoning model proposed in this paper.

6.5 Case Study

Table 8 shows examples of different models for multi-turn dialogue response selection. It can be found that the vanilla BERT is difficult to model the semantic alignment relationship between response and context in multi-turn dialogue, and retrieves an irrelevant response "Well, we will send you S.F. Express". With semantic alignment, the model is able to retrieve a response "Usually the goods will be delivered to Xinjiang within three days" that is more semantically related to the context of the conversation history, but lacks the coherence of the conversation. Finally, through the hard negative resampling strategy, the model proposed in this paper obtains a response that is semantically related to the dialogue context and has conversational coherence.

Context	**A:** Hello, I haven't received my deliver parcels, any news yet? **B:** What province are you from? **A:** I am in Xinjiang, It's been two days and I still haven't received the parcels. **B:** Today I'll help you to urge the courier. **A:** Let me know when the time comes, I'm waiting to use it! **B:** Well, I got it. **A:** Otherwise, I will deal with the case as not received the goods.
BERT	Well, we will send you S.F. Express.
MCL	Usually the goods will be delivered to Xinjiang within three days.
Ours	Sorry, I'll help you to urge the courier company.
Ground Truth	Sorry, I'll help you to urge the courier company.

Table 8 Case Study, A indicates customer and B indicates customer service staff.

7 Conclusion

In this paper, we propose a response selection model BERT-BC combining Bi-Encoder and Cross-Encoder with three contrastive learning mechanisms and a minimal-cost hard-negative resampling strategy. The BERT-BC increases the ability of the model to encode context and response through a contrastive learning mechanism, which improves the recognition of semantically related samples. At the same time, 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 response selection task. In the future work, we consider introducing common-sense knowledge graph in response selection, either by introducing knowledge-based pretrained models (ERNIE [48], K-BERT [49], etc.)

or by adding explicit knowledge embedding through graph neural networks, to enhance the model's understanding of open domain dialogue.

8 Declarations

8.1 Ethical Approval

This paper does not contain any research involving humans or animals, and ethical approvals do not apply to this paper.

8.2 Competing interests

We declare that the authors have no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

8.3 Authors' contributions

The contributions of each author in this work are as follows:

Zhenfei Yang (First Author):Conceptualization, Methodology, Experiment, Visualization, Writing;

Beiming Yu: Conceptualization, Methodology,Writing;

Cui Yuan: Investigation, Supervision;

Daling Wang: Resources, Supervision, Writing;

Ge Yu: Resources, Supervision;

Shi Feng (Corresponding Author): Conceptualization, Funding Acquisition, Resources, Supervision, Writing - Review & Editing.

8.4 Funding

This work was supported by the National Science Foundation of China (62106039, 62272092).

8.5 Availability of data and materials

The datasets used in this paper are all publicly available online(Ubuntu dataset[2], Douban dataset[3], E-Commerce dataset[4]), and the code used in this paper is open-sourced on github.

Acknowledgments. We would like to thank the reviewers for their constructive comments. This work was supported by the National Science Foundation of China (62106039, 62272092).

[2] https://www.dropbox.com/s/2fdn26rj6h9bpvl/ubuntu_data.zip
[3] https://www.dropbox.com/s/90t0qtji9ow20ca/DoubanConversaionCorpus.zip?dl=0
[4] https://drive.google.com/file/d/154J-neBo20ABtSmJDvm7DK0eTuieAuvw/view

References

[1] Huang, M., Zhu, X., Gao, J.: Challenges in building intelligent open-domain dialog systems. ACM Transactions on Information Systems (TOIS) **38**(3), 1–32 (2020)

[2] Thakur, N., Reimers, N., Daxenberger, J., Gurevych, I.: 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, pp. 296–310 (2021)

[3] Luan, Y., Eisenstein, J., Toutanova, K., Collins, M.: Sparse, dense, and attentional representations for text retrieval. Transactions of the Association for Computational Linguistics **9**, 329–345 (2021)

[4] Khattab, O., Zaharia, M.: 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, pp. 39–48 (2020)

[5] Humeau, S., Shuster, K., Lachaux, M.-A., Weston, J.: Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In: International Conference on Learning Representations (2020)

[6] Whang, T., Lee, D., Oh, D., Lee, C., Han, K., Lee, D.-h., Lee, S.: 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, vol. 35, pp. 14041–14049 (2021)

[7] Han, J., Hong, T., Kim, B., Ko, Y., Seo, J.: 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, pp. 1549–1558 (2021)

[8] Li, J., Tao, C., Wu, W., Feng, Y., Zhao, D., Yan, R.: 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, pp. 1291–1296 (2019)

[9] Lin, Z., Cai, D., Wang, Y., Liu, X., Zheng, H., Shi, S.: 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), pp. 9220–9229 (2020)

[10] Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009. ACM International

Conference Proceeding Series, vol. 382, pp. 41–48 (2009)

[11] Su, Y., Cai, D., Zhou, Q., Lin, Z., Baker, S., Cao, Y., Shi, S., Collier, N., Wang, Y.: Dialogue response selection with hierarchical curriculum learning. In: Proceedings of Annual Meeting of the Association for Computational Linguistics(Volume 1: Long Papers), pp. 1740–1751 (2021)

[12] Graves, A., Graves, A.: Long short-term memory. Supervised sequence labelling with recurrent neural networks, 37–45 (2012)

[13] Pang, L., Lan, Y., Guo, J., Xu, J., Wan, S., Cheng, X.: Text matching as image recognition. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA, pp. 2793–2799 (2016)

[14] Lowe, R., Pow, N., Serban, I.V., Pineau, J.: 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, pp. 285–294 (2015)

[15] Zhou, X., Dong, D., Wu, H., Zhao, S., Yu, D., Tian, H., Liu, X., Yan, R.: Multi-view response selection for human-computer conversation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 372–381 (2016)

[16] Bahdanau, D., Cho, K.H., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 (2015)

[17] Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 1412–1421 (2015)

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

[19] Zhou, X., Li, L., Dong, D., Liu, Y., Chen, Y., Zhao, W.X., Yu, D., Wu, H.: 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), pp. 1118–1127 (2018)

[20] Tao, C., Wu, W., Xu, C., Hu, W., Zhao, D., Yan, R.: 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, pp. 1–11 (2019)

[21] Tao, C., Wu, W., Xu, C., Hu, W., Zhao, D., Yan, R.: Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019, pp. 267–275 (2019)

[22] Yuan, C., Zhou, W., Li, M., Lv, S., Zhu, F., Han, J., Hu, S.: 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), pp. 111–120 (2019)

[23] Reimers, N., Gurevych, I.: 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), pp. 3982–3992 (2019)

[24] Zhang, S., Liang, Y., Gong, M., Jiang, D., Duan, N.: 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), pp. 5990–6000 (2022)

[25] Whang, T., Lee, D., Lee, C., Yang, K., Oh, D., Lim, H.: 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, vol. 2020, pp. 1585–1589 (2020)

[26] Xu, R., Tao, C., Jiang, D., Zhao, X., Zhao, D., Yan, R.: Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 14158–14166 (2021)

[27] Liu, L., Zhang, Z., Zhao, H., Zhou, X., Zhou, X.: Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 13406–13414 (2021)

[28] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)

[29] Gao, T., Yao, X., Chen, D.: SimCSE: Simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 6894–6910 (2021)

[30] Yan, Y., Li, R., Wang, S., Zhang, F., Wu, W., Xu, W.: ConSERT: A contrastive framework for self-supervised sentence representation transfer. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1:

Long Papers), pp. 5065–5075 (2021)

[31] liang, x., Wu, L., Li, J., Wang, Y., Meng, Q., Qin, T., Chen, W., Zhang, M., Liu, T.-Y.: R-drop: Regularized dropout for neural networks. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 10890–10905 (2021)

[32] Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)

[33] Poddar, L., Wang, P., Reinspach, J.: DialAug: Mixing up dialogue contexts in contrastive learning for robust conversational modeling. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 441–450 (2022)

[34] Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. Advances in neural information processing systems **27** (2014)

[35] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, Volume 1 (Long Papers), New Orleans, Louisiana, pp. 2227–2237 (2018)

[36] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)

[37] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems, vol. 32 (2019)

[38] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A lite BERT for self-supervised learning of language representations. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020 (2020)

[39] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: 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), pp. 4171–4186 (2019)

[40] Lowe, R., Pow, N., Serban, I.V., Pineau, J.: 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, pp. 285–294 (2015)

[41] Wu, Y., Wu, W., Xing, C., Zhou, M., Li, Z.: 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), pp. 496–505 (2017)

[42] Zhang, Z., Li, J., Zhu, P., Zhao, H., Liu, G.: Modeling multi-turn conversation with deep utterance aggregation. arXiv preprint arXiv:1806.09102 (2018)

[43] Kadlec, R., Schmid, M., Kleindienst, J.: Improved deep learning baselines for ubuntu corpus dialogs. arXiv preprint arXiv:1510.03753 (2015)

[44] Gu, J.-C., Li, T., Liu, Q., Ling, Z.-H., Su, Z., Wei, S., Zhu, X.: Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 2041–2044 (2020)

[45] Li, Y., Xu, C., Hu, H., Sha, L., Zhang, Y., Jiang, D.: Small changes make big differences: Improving multi-turn response selection in dialogue systems via fine-grained contrastive learning. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 2723–2727 (2022)

[46] Feng, J., Tao, C., Liu, C., Yan, R., Zhao, D.: How to represent context better? an empirical study on context modeling for multi-turn response selection. In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 7285–7298 (2022)

[47] Lin, T.-H., Chi, T.-C., Rumshisky, A.: On task-adaptive pretraining for dialogue response selection. arXiv preprint arXiv:2210.04073 (2022)

[48] Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: ERNIE: enhanced language representation with informative entities. In: Proceedings of the Conference of the Association for Computational Linguistics, Volume 1: Long Papers, pp. 1441–1451 (2019)

[49] Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., Wang, P.: K-BERT: Enabling language representation with knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 2901–2908 (2020)