Small Changes Make Big Differences: Improving Multi-turn Response Selection in Dialogue Systems via Fine-Grained Contrastive Learning

Anonymous ACL submission

Abstract

Retrieve-based dialogue response selection aims to find a proper response from a candidate set given a multi-turn context. The sequence representations generated by pre-trained language models (PLMs) play key roles in the learning of matching degree between the dialogue contexts and the responses. However, we observe that different context-response pairs sharing the same context always have a greater similarity in the sequence representations calculated by PLMs, which makes it hard to distinguish positive responses from negative ones. Motivated by this, we propose a novel **F**ine-**G**rained **C**ontrastive (FGC) learning method for the response selection task based on PLMs. This FGC learning strategy helps PLMs to generate more distinguishable matching representations of each dialogue at fine grains, and further make better predictions on choosing positive responses. Empirical studies on two benchmark datasets demonstrate that the proposed FGC learning method can generally and significantly improve the model performance of existing PLM-based matching models.[1]

Dialogue Context Between Participants A and B
A: Hello. What kind of movies do you like?
B: Hi, I like action movies.
A: I really liked The Amazing Spider-Man 2 and Spectre they're different but both action filled.
Positive Response
B: I have not seen Spectre. I will have to look for it.
Negative Response
B: I love chocolate on my fried ice cream cake!

Table 1: A dialogue context with a positive and a negative response. The cosine similarity of matching representation between positive and negative context-response pairs is +0.870.

1 Introduction

Multi-turn response selection is the task of predicting the most appropriate response using a retrieval model by measuring the matching degree between a multi-turn dialogue context and a set of response candidates. Most recently, pre-trained language models (PLMs) have achieved substantial performance improvements in multi-turn response selection (Lu et al., 2020; Gu et al., 2020; Humeau et al., 2020). PLM-based models take the concatenation of dialogue context and response as the input, and utilize the matching representation (i.e., the embedding of [CLS] token on the top layer) to predict a score indicating the matching degree. By further post-training PLMs with in-domain data and auxiliary self-supervised tasks (Whang et al., 2020, 2021a; Xu et al., 2021), PLMs achieve state-of-the-art results on benchmarks.

Despite the success of PLM-based matching models and their various variants, the cross-entropy loss is widely used to tune the matching representations during the training of such models. Previous works (Elsayed et al., 2018; Soudry et al., 2018) demonstrate that the cross-entropy loss is inferior in establishing a large margin between positive and negative examples. We conduct prior analysis and find that the average BERT representation cosine similarity between all positive and negative examples is +0.074. A positive similarity indicates the matching representation learned with the cross entropy are not strong enough to distinctively separate the positive and negative examples. Contrastive learning (CL) (Wang and Isola, 2020) provides a way to address the problem described above. Employing the off-the-shelf constrative learning method (Chen et al., 2020; Fang and Xie, 2020) could reduce the similarity of average positive-negative representations to -0.033. However, the matching representation of two examples with the same context but different responses are still too similar, with an average similarity of +0.068. The matching representation is aggregated

[1] We will make the code public available later to facilitate reproducing the results.

from all embeddings of the dialogue context and response. Therefore, the shared context of the two examples above could cause the matching representations to have higher similarities. Table 1 shows an example, although there are great difference between positive and negative responses, the similarity of two context-response pairs is 0.870, which is a very positive number. This number only decreases to 0.246 with regular contrastive learning, which is still a positive number. This phenomenon makes the matching representations less distinguishable and makes it hard to separate positive pairs from negative ones.

To address the aforementioned issues, we propose a **F**ine-**G**rained **C**ontrastive learning (**FGC**) approach to fine-tune matching representations for the response selection task. FGC introduces contrastive learning on each example with the same context and different responses. In contrast to the off-the-shelf contrastive learning method, which takes every other context-response pair as negative examples, FGC takes context and response as separate parts and focuses on distinguishing between positive and negative examples with the same context. Each context-response pair is converted into an augmented pair during FGC learning via rule-based transformation on the response utterance. Each pair is asked to be close to its augmentation, while the one with a positive response should be far away from the one with a negative response. FGC works totally in a self-supervised way that no additional supervision is required besides the classification label used for response selection training.

We conduct experiments on two response selection benchmarks: the Ubuntu Dialogue Corpus (Lowe et al., 2015) and the Douban Corpus (Wu et al., 2017). Our empirical results demonstrate that FGC is able to consistently improve PLMs by up to 3.2% absolute improvement with an average of 1.7% absolute improvement in terms of $R_{10}@1$. Besides, We also compare our method with standard-contrastive-learning-enhanced PLMs, which demonstrates the effectiveness of our proposed fine-grained contrastive objective.

In summary, our contributions in the paper are three-fold:

- We propose FGC, a novel fine-grained contrastive learning method, which helps generate better representations of dialogues and improves the response selection task.
- FGC shows good generality of effectiveness with various pre-trained language models for enhancing performance.
- Experimental results on two benchmark datasets demonstrate that FGC can significantly improve the performance of various strong PLM-based matching models, including state-of-the-art ones.

2 Related Work

2.1 Multi-Turn Response Selection

Multi-turn response selection has been discussed for a long period. Modeling dialogues with RNN-based models has been tried first on this task (Lowe et al., 2015; Zhou et al., 2016). However, these methods ignore relationships among utterances by simply concatenating the utterances together or converting the whole context to a vector. To alleviate this, Wu et al. (2017) proposed the sequential matching network that each utterance in the context first interacts with the response candidate, and then the matching features are aggregated according to the sequential order of multi-turn context. With the rise of the self-attention mechanism, some studies (Zhou et al., 2018; Tao et al., 2019a) explored how to enhance representations with it. Besides, Yuan et al. (2019) proposed a multi-hop selector to select the relevant utterances in the dialogue context for response matching.

Most recently, pre-trained language models (e.g., BERT (Devlin et al., 2019)) have shown an impressive performance in the response selection. The post-training method, which helps transfer the representations of BERT from the general domain to the dialogue domain, was proposed by Whang et al. (2020) and obtained state-of-the-art results. Subsequent researches (Gu et al., 2020; Lu et al., 2020) focused on incorporating speaker information into BERT and showed its effectiveness in multi-turn response selection. Further, self-supervised learning has been introduced into this task. Whang et al. (2021a) and Xu et al. (2021) indicated that incorporating well-designed self-supervised tasks according to the characteristics of the dialogue data into BERT fine-tuning can help with the multi-turn response selection. Han et al. (2021) proposed a fine-grained post-training method for enhancing the pre-trained language model, while the post-training process is computationally expensive than fine-tuning a classification model. Su et al. (2020) proposed a hierarchical curriculum learning framework for improving response selection with PLMs.

2.2 Contrastive Learning for NLP

There have been several investigations for contrastive learning for neural models, primarily in the computer vision domain. Oord et al. (2018) proposed a framework for contrastive learning to learn visual representations based on contrastive predictive coding, which predicts the features in latent space by using powerful autoregressive models. Khosla et al. (2020) investigated supervised contrastive learning, allowing to leverage label information effectively. Following this trend, some researchers verified the effectiveness of constructive learning in specific NLP tasks. For eaxmple, Fang and Xie (2020) proposed pre-training language representation models with a contrastive self-supervised learning objective at the sentence level, outperforming previous methods on a subset of GLUE tasks. Gunel et al. (2021) combined the cross-entropy with a supervised contrastive learning objective, showing improvements over fine-tuning RoBERTa-Large on multiple datasets of the GLUE benchmark. Our work differs from previous works in that we do not directly make contrast on one example with all the other examples, as the granularity of negative samples constructed using this approach is too coarse to provide sufficient discrimination with the positive ones.

3 Background

3.1 Task Formalization

The response selection task is to select the best candidate to respond a given multi-turn dialogue context from a pool of candidate responses. Suppose that we have a dataset $D = \{c_i, r_i, y_i\}_{i=1}^N$, where $c_i = \{u_i^1, \cdots, u_i^{n_i}\}$ is a multi-turn dialogue context with n_i turns, r_i denotes a candidate response, and $y_i \in \{0, 1\}$ denotes a label with $y_i = 1$ indicating r_i a proper response for c_i and otherwise $y_i = 0$. Our goal is to estimate a matching model $y = f(\cdot, \cdot)$ from D. For any given context-response pair (c, r), $f(c, r)$ returns a score that reflects the matching degree between c and r.

3.2 Pre-trained Language Model for Response Selection

As a trend in these years, pre-trained language models, e.g., BERT(Devlin et al., 2019), have been widely studied and adapted into numerous NLP tasks, showing several state-of-the-art results. Dialogue response selection is one of them.

Applying a pre-trained language model into response selection usually involves two steps. The first step is to make domain-adaptive post-training, which continues to train a standard pre-trained language model with a domain-specific corpus. This step helps to transfer the original pre-trained language model into the target domain.

The second step is to fine-tune the post-trained model with the response selection task. Given a context $c = \{u_1, \cdots, u_m\}$ where u_t is the t-th turn of the dialog context, and a response r, the model is asked to predict a score $\hat{y}$ to represent the matching degree between c and r. To achieve this, a special token [EOT] is added at the end of each turn to distinguish them in the context c. Utterances from both the context c and response r are concatenated with separator [EOT] and [SEP] between them. Taking x as input, BERT returns a sequence of vectors with the same length as x. The output of the first place $\mathbf{s}_{\texttt{[CLS]}}$ is an aggregated representation vector that holds the information of interaction between context c and response r. A relevance score $\hat{y}$ is computed based on $\mathbf{s}_{\texttt{[CLS]}}$ and optimized through a binary classification loss.

$$\begin{aligned} \hat{y} &= \sigma(\mathbf{W}_{sel}\mathbf{s}_{\texttt{[CLS]}} + b) \\ \mathcal{L}_{\texttt{sel}} &= -\left(y \log \hat{y} + (1 - y) \log(1 - \hat{y})\right), \end{aligned} \tag{1}$$

where $\mathbf{W}_{sel}$ and b are parameters and $\hat{y}$ denotes for the ground truth binary label.

4 Methodology

4.1 Overview

In this paper, we propose the **F**ine-**G**rained **C**ontrastive Learning method (**FGC**) for learning PLMs-based matching models. It consists of two complementary contrastive objectives: (1) an instance-view contrastive objective (IVC); and (2) a category-view contrastive objective (CVC). Figure 1 demonstrates the joint effects of the two contrastive objectives on the space of matching representations. The IVC objective pushes away examples with the same context and different responses, making the model easier to distinguish between positive and negative responses. However, only pushing the examples with the same context away increases the risk of instances with different contexts getting closer in the representation space. As a remedy, the CVC objective further pulls all context-response pairs into two distinguishable clusters in the matching space according to whether the pair is positive or not. These two objectives are introduced in 4.3 and 4.4 respectively. For simplicity, we take BERT as an example in the following sections in the following sections.

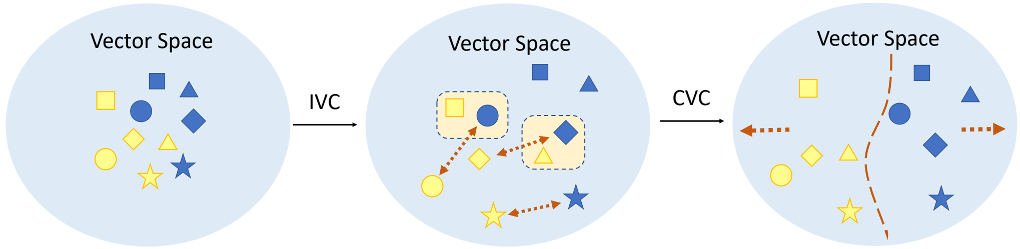

Figure 1: FGC contains two objectives. IVC pushes away examples with the same context but different responses (icons in the same shape), while examples that belong to different categories may still be similar (as shown in orange boxes). CVC further solves this problem by pulling all examples into two distinguishable clusters.

4.2 Dialogue Data Augmentation

Data augmentation takes an important role in contrastive learning (Zoph et al., 2020; Ho and Vasconcelos, 2020). Similar to standard contrastive learning (e.g., CERT), the first step of FGC is to create augmentations for every context-response pair. Given a context-response pair, we take an augmentation method on the response to generate an augmented response. The context and augmented response pair form the augmentation of the original context-response pair. In order to fine-grained control the difference between a pair and its corresponding augmentation and easily perform augmentation on various languages, a fully unsupervised rule-based utterance augmentation method is adopted for utterance augmentation. Inspired by (Wei and Zou, 2019), we adopt three types of augmentation operations:

- **Random deletion**: Each token in the utterance is randomly and independently deleted with a probability p_{del}.
- **Random swaping**: Each token in the utterance is randomly swapped with another token in the utterance with a probability p_{swap}.
- **Synonym replacing**: Randomly replace a non-stop-word token to one of its synonyms with a probability p_{syn}.

Given a response utterance r and an augmentation strength $p \in [0, 1]$, we randomly pick out one of these three augmentation methods and then apply the augmentation on the utterance with the probability being p. After augmentation, the response r is converted into another augmented response $\bar{r}$. The augmentation strength p is a hyper-parameter that controls how much difference is there between r and $\bar{r}$. A smaller p brings less variety to r, which makes it easier to cluster r and its augmentation $\bar{r}$, as well as distancing positive and negative response pairs, while a larger p may potentially introduce too much noise into model training, which harms the representation learning and further influence the response selection task.

4.3 Instance-View Contrastive Objective

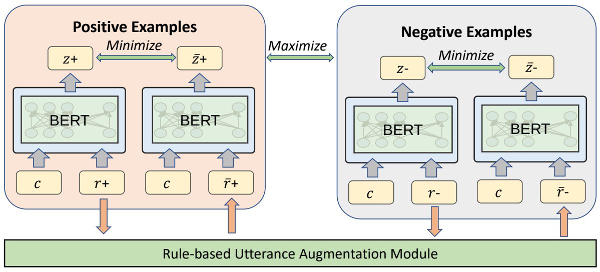

Figure 2: An overview of IVC. The input is a dialogue context $\mathbf{c}$ and a pair of positive and negative responses $(\mathbf{r}+, \mathbf{r}-)$. Both responses are augmented to form a new pair $(\bar{\mathbf{r}}+, \bar{\mathbf{r}}-)$. BERT takes four examples as input and outputs a projection vector $\mathbf{z}$ for each of them. IVC aims to maximize the dissimilarity of $\mathbf{z}$ between positive examples and negative examples, as well as maintains high cohesion within positive and negative cases.

The instance-view contrastive (IVC) objective aims at introducing more discrepancy between a pair of examples with the same context and positive/negative responses. Feeding a context-response pair into BERT, BERT helps to make internal interactions by attention mechanism and generate latent vectors representing the pair. The output vector of the `[CLS]` position $\mathbf{s}_{\texttt{[CLS]}}$ stands for a aggregated sequence representation of both context and response. We also take this vector as the matching representation used for contrastive learning. Moreover, we apply another projection layer to convert $\mathbf{s}_{\texttt{[CLS]}}$ into a smaller vector $\mathbf{z}$. This projection is made through an MLP with one hidden layer. Through this projection, each coherent pair with positive responses (c_i, r_i+) is transformed into a projection vector $\mathbf{z}_i+$, and each incoherent pair (c_i, r_i-) is transformed into $\mathbf{z}_i-$. The augmentations of the positive and negative pairs are also converted into two vectors, i.e., $\bar{\mathbf{z}}_i+$ and $\bar{\mathbf{z}}_i-$. Here $+$ and $-$ indicates the item belongs to the positive class or the negative class, and the bar indicates this item comes from an augmented example.

As illustrated by Ethayarajh (2019) and Li et al. (2020), the embedding vectors of different utterances are distributed in a narrow cone of the vector space, showing less distinguishability. This phenomenon is even worse when two utterances are semantically similar, e.g., two examples sharing the same context. Thus, we leverage the IVC objective on these projection vectors $\mathbf{z}$ to distinguish between positive and negative responses given the same context. IVC objective regards the projection vector $\mathbf{z}$ as a representation of response r given context c. This loss is applied on the projection vector $\mathbf{z}$, which helps to maximize the similarity between a response with its augmentation given the same context, as well as minimize the similarity between each positive response and negative response pair. The maximum and minimum are achieved as a set of pair-wise comparisons, i.e.,

$$\begin{aligned}\forall i \quad & \mathbf{sim}(\mathbf{z}_i+, \bar{\mathbf{z}}_\mathbf{i}+) > \\ & \quad \mathbf{sim}(\mathbf{z}_i+, \mathbf{z}_i-), \mathbf{sim}(\mathbf{z}_i+, \bar{\mathbf{z}}_i-) \\ & \quad \mathbf{sim}(\bar{\mathbf{z}}_i+, \mathbf{z}_i-), \mathbf{sim}(\bar{\mathbf{z}}_i+, \bar{\mathbf{z}}_i-) \\ \forall i \quad & \mathbf{sim}(\mathbf{z}_i-, \bar{\mathbf{z}}-_i) > \\ & \quad \mathbf{sim}(\mathbf{z}_i-, \mathbf{z}_i+), \mathbf{sim}(\mathbf{z}_i-, \bar{\mathbf{z}}_i+) \\ & \quad \mathbf{sim}(\bar{\mathbf{z}}_i-, \mathbf{z}_i+), \mathbf{sim}(\bar{\mathbf{z}}_i-, \bar{\mathbf{z}}_i+).\end{aligned} \tag{2}$$

Here we use the NT-Xent Loss (Chen et al., 2020) to model the similarities of projection vectors. By writing this pair-wise comparison into a loss function, the IVC loss is formulated as

$$\begin{aligned} l(\mathbf{z}, \bar{\mathbf{z}}) &= -\log \frac{\exp(\mathbf{sim}(\mathbf{z}, \bar{\mathbf{z}})/\tau)}{\sum\limits_{\mathbf{z}_k \neq \mathbf{z}} \exp(\mathbf{sim}(\mathbf{z}, \mathbf{z}_k)/\tau)} \\ \mathcal{L}_{ivc} &= \sum_{i=1}^{N} \left(l\left(\mathbf{z}_i+, \bar{\mathbf{z}}_i+\right) + l\left(\mathbf{z}_i-, \bar{\mathbf{z}}_i-\right) \right), \end{aligned} \tag{3}$$

where $\tau > 0$ is a scalar temperature parameter that controls the separation of positive and negative classes; $\mathbf{z}_k$ ranges from $\{\mathbf{z}+, \bar{\mathbf{z}}+, \mathbf{z}-, \bar{\mathbf{z}}-\}$; and N is the total number of examples.

Notice that the IVC objective aims to separate the representation of positive and negative responses given the same context, so that we do not take all other in-batch examples as negative examples in the same way as in standard contrastive learning.

4.4 Category-View Contrastive Objective

The IVC objective ensures a high difference between examples with the same context, while it cannot guarantee that the learned representations are suitable for classification. The representations of a positive example may be close to the representation of another negative example with a different context, as is shown in Figure 1. Thus, we introduce another category-view contrastive (CVC) objective into model training. The category-view contrastive objective aims at bunching examples that belong to the same category into a cluster and separate these two clusters.

There are two categories for the response selection task, i.e., the positive category that indicates the response is a proper response for the given context, and the negative category in vise versa. The CVC objective is applied between examples from the two classes. It captures the similarity of projection vectors of the same class and contrasts them with projection vectors from the other class, i.e.,

$$\begin{aligned} &\forall i, j, k, l \;\; \mathbf{sim}(\mathbf{z}_i+, \mathbf{z}_j+) > \mathbf{sim}(\mathbf{z}_k+, \mathbf{z}_l-) \\ &\forall i, j, k, l \;\; \mathbf{sim}(\mathbf{z}_i-, \mathbf{z}_j-) > \mathbf{sim}(\mathbf{z}_k+, \mathbf{z}_l-). \end{aligned} \tag{4}$$

This category-view contrastive loss works with a batch of representation vectors of size $2N$, where the number of both positive examples and negative examples is N. Denote $\{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_{2N-1}, \mathbf{z}_{2N}\}$ to be all representation vectors in a batch, where $\{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N\}$ are representation vectors for positive examples and their augmentations, and $\{\mathbf{z}_{N+1}, \mathbf{z}_{N+2}, \cdots, \mathbf{z}_{2N}\}$ are representation vectors for negative examples and their representations. The CVC objective works as an additional restriction to punish the high similarity between positive-negative pairs and low similarity within all positive and negative examples. The following formulas give this loss:

$$\begin{aligned} l(\mathbf{z}_i, \mathbf{z}_j) &= \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j/\tau)}{\sum_{i \neq r} \exp(\mathbf{z}_i \cdot \mathbf{z}_r/\tau)} \\ \mathcal{L}_{cvc} &= -\frac{1}{N-1} \sum_{i=1}^{2N} \sum_{i \neq j} \mathbf{1}_{\bar{y}_i = \bar{y}_j} l(\mathbf{z}_i, \mathbf{z}_j) \end{aligned} \tag{5}$$

.

Finally, the BERT model is fine-tuned with the standard response selection loss $\mathcal{L}_{\texttt{sel}}$ and both IVC and CVC loss. A weighted summation is computed as

$$\mathcal{L} = \mathcal{L}_{\texttt{sel}} + \lambda(\mathcal{L}_{ivc} + \mathcal{L}_{cvc}), \tag{6}$$

where λ is a hyper-parameter that controls the balance between response selection loss and contrastive loss. The model is optimized by minimizing the overall loss value.

5 Experiments

5.1 Dataset

- **Ubuntu Dialogue Corpus V1**
  The Ubuntu Dialogue Corpus V1 (Lowe et al., 2015) is a domain-specific multi-turn conversation dataset. Conversations in this dataset are dumped from the multi-party chat room whose topic is the Ubuntu operating system.
- **Douban Corpus**
  The Douban Corpus(Wu et al., 2017) is a Chinese dataset collected from an online social network website named Douban. Douban Corpus is an open-domain conversation corpus, whose topic is much wider than that of Ubuntu Corpus.

The statistics of these two datasets are shown in Table 2. These two datasets vary greatly in both language and topic. Following previous works, we take $R_{10}@k$s as evaluation metrics, which measures the probability of having the positive response in the top k ranked responses. We take $k = \{1, 2, 5\}$ for model evaluation.

Dataset	Ubuntu			Douban		
	Train	Val	Test	Train	Val	Test
# dialogues	1M	500K	500K	1M	50K	6670
#pos:#neg	1:1	1:9	1:9	1:1	1:1	1.2:8.8
# avg turns	10.13	10.11	10.11	6.69	6.75	6.45

Table 2: Statistics of two datasets.

5.2 Baseline Methods

We introduce FGC into several open-sourced PLM-based models, including BERT and ELECTRA. We also test the effectiveness of FGC on variants of BERT model, including BERT-small (H=4, L=4, H=512), BERT with domain-adaptive post training named BERT-DPT (Whang et al., 2020), and BERT with self-supervised tasks named BERT-UMS (Whang et al., 2021b). Several non-PLM-based models are also compared.[2]

5.3 Implementation Details

All models are implemented based on Pytorch and Huggingface's implementation. Each PLM model is trained for 5 epochs with a learning rate beginning from 3e-6 to 0 with a linear learning rate decay. Our model is trained on 8 Nvidia Tesla A100 GPUs with 40GB memory.

[2]A pre-trained Chinese BERT-Small is not available, thus we do not conduct experiments on it.

5.4 Experimental Results

The comparison between PLMs and FGC-enhanced PLMs is shown in Table 3. All PLM-based methods outperform non-PLM-based methods. By adding our proposed FGC into PLM-based models, the performance of all models is significantly improved. The maximum improvement of a standard-sized BERT for the two datasets are 1.9% and 3.2% respectively in terms of $R_{10}@1$. The average performance improvement also achieves 1.1% and 2.2%. Besides, our proposed method can also enhance the current state-of-the-art method BERT-UMS by 1.1% and 0.8% on two datasets in terms of $R_{10}@1$. In addition to a standard-sized BERT model, we also find an absolute gain of 0.9% by adding FGC on the BERT-Small model, which is about 10$\times$ smaller than a standard one. The success of these two datasets demonstrates the effectiveness of our proposed FGC across different models, languages, and dialogue topics on multi-turn response selection.

FGC separates representation vectors of examples into different latent spaces according to their type of relevance between contexts and responses. On the one hand, IVC helps distinguish between positive and negative responses given the same context. On the other hand, CVC separates representations of examples from two categories so that these representations can have better distinguishability. As a result, the matching representation of context-response pairs for positive and negative responses are forced to stay away from each other. These better representations ensures higher accuracy in selecting the positive response given a set of candidate responses.

6 Closer Analysis

We conduct closer analysis with BERT-DPT since combining post-training and fine-tuning is the most popular manner of applying BERT for downstreaming tasks. The Ubuntu Corpus is used in the following analysis.

6.1 Ablation Studies

As we add two contrastive learning objectives into training for response selection, we test the gain of each objective. The results are shown in Table 4. It can be observed from the table that both IVC and

Models	Ubuntu			Douban					
	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	P@1	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
non-PLM-based methods									
Multi-View (Zhou et al., 2016)	0.662	0.801	0.951	0.505	0.543	0.342	0.292	0.350	0.729
SMN (Wu et al., 2017)	0.726	0.847	0.961	0.529	0.569	0.397	0.233	0.396	0.724
DUA (Zhang et al., 2018)	0.752	0.868	0.961	0.551	0.599	0.421	0.243	0.421	0.780
DAM (Zhou et al., 2018)	0.767	0.874	0.961	0.550	0.601	0.427	0.254	0.410	0.757
MRFN (Tao et al., 2019a)	0.786	0.886	0.976	0.571	0.617	0.448	0.276	0.435	0.783
IoI (Tao et al., 2019b)	0.796	0.894	0.974	0.573	0.621	0.444	0.269	0.451	0.786
IMN (Gu et al., 2019)	0.794	0.889	0.974	0.576	0.618	0.441	0.268	0.458	0.796
MSN (Yuan et al., 2019)	0.800	0.899	0.978	0.587	0.632	0.470	0.295	0.452	0.788
PLM-based Methods									
BERT	0.820	0.906	0.978	0.597	0.634	0.448	0.279	0.489	0.823
BERT+FGC	**0.829**	0.910	0.980	**0.614**	**0.653**	**0.495**	**0.312**	0.495	**0.850**
BERT-DPT (Whang et al., 2020)	0.862	0.935	0.987	0.609	0.645	0.463	0.290	0.505	0.838
BERT-DPT+FGC	**0.881**	**0.945**	**0.990**	0.620	**0.660**	**0.495**	**0.322**	0.495	**0.850**
BERT-UMS (Whang et al., 2021b)	0.875	0.942	0.988	0.625	0.664	0.499	0.318	0.482	0.858
BERT-UMS+FGC	**0.886**	**0.948**	0.990	0.627	0.670	0.500	**0.326**	**0.512**	**0.869**
ELECTRA	0.826	0.908	0.978	0.602	0.642	0.465	0.287	0.483	0.839
ELECTRA+FGC	**0.832**	0.912	0.980	**0.625**	**0.668**	**0.499**	**0.313**	**0.502**	**0.850**
BERT-Small	0.792	0.888	0.972	N/A	N/A	N/A	N/A	N/A	N/A
BERT-Small+FGC	**0.800**	0.890	0.974	N/A	N/A	N/A	N/A	N/A	N/A

Table 3: Evaluation results on the two data sets. Numbers in bold indicate that the PLM-based models using FGC outperforms the original models with a significance level p-value < 0.05.

Strategy	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
BERT-DPT + FGC	**0.881**	**0.944**	**0.990**
- IVC	0.866	0.935	0.986
- CVC	0.877	0.941	0.988
BERT-DPT	0.862	0.935	0.987

Table 4: Ablation Analysis on the Ubuntu corpus.

CVC can enhance the performance on response selection, with an absolute improvement of 1.4% and 0.4% respectively in terms of $R_{10}@1$. By applying these two contrastive objectives together, we obtain an absolute improvement of 1.9% based on the post-trained BERT model. Both of the two contrastive objectives share the same purpose of separating the representation of examples with positive and negative responses, and thus there is a performance overlap by adding these two objectives.

6.2 Sensitive Analysis

Temperature Temperature τ works as a hyperparameter that controls the punishment on the degree of separation of positive and negative classes. A smaller τ gives more power to pull away examples from different classes. We test how this hyperparameter can influence the response selection performance. We test τ in the range of $\{0.1, 0.5, 1\}$ on FGC and the results are shown in Table 5. FGC achieves the best performance when τ is set to be 0.5, while the performance drops given a smaller or a bigger τ. A suitable τ can provide a proper differentiation that is neither too strong nor too weak, keeping a balance between contrastive and response selection objectives.

Temperature τ	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
BERT-DPT	0.862	0.935	0.987
+ FGC (τ=0.1)	0.872	0.939	**0.990**
+ FGC (τ=0.5)	**0.881**	**0.944**	**0.990**
+ FGC (τ=1.0)	0.876	0.938	**0.990**

Table 5: Influence of temperature τ in FGC.

Utterance Augmentation Strength Utterance augmentation plays an important role in contrastive learning. A dialogue with a context and a positive response is drawn closer to its augmentation while pushed far away from the dialogue with the same context but a negative response. The strength of utterance augmentation decides the boundary of each cluster. We conduct experiments to test how augmentation strength can influence response selection accuracy. We range the augmentation strength p from $\{0.1, 0.2, 0.5\}$, and the testing results are shown in Table 6. It achieves the best performance when $p = 0.2$. Augmentation strength being either too large or too small may harm the clustering. On the one hand, a too-large p brings too much noise into the clustering process, which blurs the boundary between positive and negative examples. On the other hand, a too-small p cannot provide enough variation to the utterance, which harms the generalization of identifying positive responses.

Augment Strength p	R_{10}@1	R_{10}@2	R_{10}@5
BERT-DPT	0.862	0.935	0.987
+ FGC (p=0.1)	0.874	0.938	0.989
+ FGC (p=0.2)	**0.881**	**0.944**	**0.990**
+ FGC (p=0.5)	0.872	0.935	**0.990**

Table 6: Influence of augmentation strength p in FGC.

6.3 Discussion

Effect of Data Augmentation Alone Data augmentation, working as a kind of data noise, shows a positive effect on training models with robustness in natural language processing. One may concern that can data augmentation alone help with the response selection task. We conducted experiments with data augmentation alone, i.e., no contrastive learning strategy is included. The results are shown in Table 7. It can be observed from the table that data augmentation alone cannot enhance the model but even harm the accuracy significantly. Data augmentation methods should work with fine-grained contrastive learning to make positive effects for the multi-turn response selection task.

	Ubuntu	Douban
BERT-DPT	0.862	0.290
+Aug	0.837 (-2.5%)	0.278 (-1.2%)
BERT-UMS	0.875	0.318
+Aug	0.851 (-2.4%)	0.292 (-2.6%)

Table 7: Performance with data augmentation alone.

Compare with Standard Contrastive Learning The main difference between our proposed FGC and standard contrastive learning (e.g., CERT (Fang and Xie, 2020) and SimCSE (Gao et al., 2021)) is that we only take examples with the same context but different responses as negative examples, instead of using in-batch examples as negative ones. We compare FGC with those methods, whose results are shown in Table 8. Standard contrastive learning can bring less gain (or even harm) on the response selection task, while contrastive learning with fine-grained negative examples leads to a significant gain on this task.

Similarity between Examples The goal of FGC is to enlarge distances between dialogue examples with the same context and different responses. To estimate how effective this target is achieved, we compute two average cosine similarities: (1) instance-level similarity, which is the average similarity between dialogue pairs with the same context but different responses; and (2) category-level similarity, which is the average similarity between all positive examples and negative examples. As can be seen from Table 9, both similarities are lowered from a positive value indicating positive correlation into a negative value indicating negative correlation by adding FGC. By introducing better distinguishability into dialogue representations, our proposed FGC helps to make better response predictions effectively. Though these two similarities can also be lowered by adding IVC alone, the category similarity is not small enough to separate the two categories well. This shortcoming is compensated by further applying CVC as an additional training objective. Besides, CVC alone can neither provide a sufficiently low level of instance-level similarity that separates examples with the same context.

Contrastive Method	R_{10}@1	R_{10}@2	R_{10}@5
BERT-DPT	0.862	0.935	0.987
BERT-DPT + CERT	0.855	0.931	0.985
BERT-DPT + SimCSE	0.864	0.936	0.987
BERT-DPT + FGC	**0.881**	**0.944**	**0.990**

Table 8: Result on comparing with regular contrastive learning methods.

Strategy	Ins Sim	Cat Sim
BERT-DPT	+0.074	+0.064
+ IVC	-0.178	-0.015
+ CVC	+0.052	-0.109
BERT-DPT + FGC	-0.111	-0.131

Table 9: Similarity Analysis on the Ubuntu corpus.

7 Conclusion

In this paper, we propose FGC, a fine-grained contrastive learning method, which helps to improve the multi-turn response selection task with PLM-based models. FGC consists of an instance-view contrastive (IVC) objective that helps to differentiate positive response and negative response with the same context, and a category-view contrastive (CVC) objective that separate positive examples and negative examples into two distinguishable clusters. Experiments and analysis on two benchmark datasets and five PLM-based models demonstrates the effectiveness of FGC to significantly improve the performance of multi-turn dialogue response selection.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Gamaleldin F Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. 2018. Large margin deep networks for classification. *arXiv preprint arXiv:1803.05598*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Hongchao Fang and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *CoRR*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *CIKM*, pages 2041–2044.

Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2019. Interactive matching network for multi-turn response selection in retrieval-based chatbots. In *CIKM*, pages 2321–2324.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. *ICLR*.

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.

Chih-Hui Ho and Nuno Vasconcelos. 2020. Contrastive learning with adversarial examples. *NeurIPS*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *ICLR*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *NeurIPS*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *ICLR*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems.

Junyu Lu, Xiancong Ren, Yazhou Ren, Ao Liu, and Zenglin Xu. 2020. Improving contextual language models for response retrieval in multi-turn conversation. In *SigIR*, pages 1805–1808.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. 2018. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878.

Yixuan Su, Deng Cai, Qingyu Zhou, Zibo Lin, Simon Baker, Yunbo Cao, Shuming Shi, Nigel Collier, and Yan Wang. 2020. Dialogue response selection with hierarchical curriculum learning. *arXiv preprint arXiv:2012.14756*.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019a. Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. In *WSDM*, pages 267–275.

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

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

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 *Proc. Interspeech 2020*.

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021a. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection.

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021b. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. In *AAAI*.

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, Vancouver, Canada. Association for Computational Linguistics.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2021. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues.

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

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

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

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. 2020. Rethinking pre-training and self-training. *NeurIPS*.

A More Implementation Details

Domain-adaptive Post-training For domain-adaptive post-training, we take the same hyper-parameter settings as BERT-DPT (Whang et al., 2020). Concretely, the maximum length of input dialogue is set to be 512. A full dialogue is randomly cut into a shorted token sequence with a probability of 10%. A masked language model loss and a next sentence prediction loss is optimized jointly during post-training. For the masked language model training, we masked each token with a probability of 15%. The post-training process traverses all the dialogues for 10 iterations, and the words that are masked during each iteration are independently sampled.

Fine-tuning for Response Selection The model is fine-tuned with the response selection task. The projection layer for transforming [CLS] vectors into projection vectors z is an MLP with one hidden layer with hidden size being 256. For dialogues longer than 512 (i.e. the maximum length supported by BERT), we discard the beginning of its context while keeps a complete response, as the latter part of the dialogue context may have stronger relevance with the response. We take an AdamW optimizer(Loshchilov and Hutter, 2019) with linear learning rate decay for fine-tuning. The initial learning rate is $3 * 10^{-5}$, and gradually decreases to 0 within 5 epochs. The λ for controlling the balance between response selection loss and contrastive loss is set to be 1.

All pre-trained language model checkpoints are downloaded from huggingface[3], with their names as the keys except for BERT-Small. For the BERT-Small model, the pre-trained model checkpoint is downloaded with model name “prajjwal1/bert-small”. Each model is trained by 3 times, and the best results among them are reported.

[3]https://huggingface.co/