DialogueCSE: Dialogue-based Contrastive Learning of Sentence Embeddings

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, Luo Si
DAMO Academy, Alibaba Group
{liuche.lc,wr224079,shaohan.ljh,jian.sun,f.huang,luo.si}@alibaba-inc.com

Abstract

Learning sentence embeddings from dialogues has drawn increasing attention due to its low annotation cost and high domain adaptability. Conventional approaches employ the siamese-network for this task, which obtains the sentence embeddings through modeling the context-response semantic relevance by applying a feed-forward network on top of the sentence encoders. However, as the semantic textual similarity is commonly measured through the element-wise distance metrics (e.g. cosine and L2 distance), such architecture yields a large gap between training and evaluating. In this paper, we propose DialogueCSE, a dialogue-based contrastive learning approach to tackle this issue. DialogueCSE first introduces a novel matching-guided embedding (MGE) mechanism, which generates a context-aware embedding for each candidate response embedding (i.e. the context-free embedding) according to the guidance of the multi-turn context-response matching matrices. Then it pairs each context-aware embedding with its corresponding context-free embedding and finally minimizes the contrastive loss across all pairs. We evaluate our model on three multi-turn dialogue datasets: the Microsoft Dialogue Corpus, the Jing Dong Dialogue Corpus, and the E-commerce Dialogue Corpus. Evaluation results show that our approach significantly outperforms the baselines across all three datasets in terms of MAP and Spearman's correlation measures, demonstrating its effectiveness. Further quantitative experiments show that our approach achieves better performance when leveraging more dialogue context and remains robust when less training data is provided.

1 Introduction

Sentence embeddings are used with success for a variety of NLP applications (Cer et al., 2018) and many prior methods have been proposed with different learning schemes. Kiros et al. (2015); Logeswaran and Lee (2018); Hill et al. (2016) train sentence encoders in a self-supervised manner with web pages and books. Conneau et al. (2017); Cer et al. (2018); Reimers and Gurevych (2019) propose to learn sentence embeddings on the supervised datasets such as SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). Although the supervised-learning approaches achieve better performance, they suffer from high cost of annotation in building the training dataset, which makes them hard to adapt to other domains or languages.

Recently, learning sentence embeddings from dialogues has begun to attract increasing attention. Dialogues provide strong semantic relationships among conversational utterances and are usually easy to collect in large amounts. Such advantages make the dialogue-based self-supervised learning methods promising to achieve competitive or even superior performance against the supervised-learning methods, especially under the low-resource conditions.

While promising, the issue of how to effectively exploit the dialogues for this task has not been sufficiently explored. Yang et al. (2018) propose to train an input-response prediction model on Reddit dataset (Al-Rfou et al., 2016). Since they build their architecture based on the single-turn dialogue, the multi-turn dialogue history is not fully exploited. Henderson et al. (2020) demonstrate that introducing the multi-turn dialogue context can improve the sentence embedding performance. However, they concatenate the multi-turn dialogue context into a long token sequence, failing to model inter-sentence semantic relationships among the utterances. Recently, more advanced methods such as (Reimers and Gurevych, 2019) achieve better performance by employing BERT (Devlin et al., 2019) as the sentence encoder. These works have in common that they employ a feed-forward network with a non-linear activation on top of the sentence en-

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2396–2406
November 7–11, 2021.

coders to model the context-response semantic relevance, thereby learning the sentence embeddings. However, such architecture presents two limitations: (1) It yields a large gap between training and evaluating, since the semantic textual similarity is commonly measured by the element-wise distance metrics such as cosine and L2 distance. (2) Concatenating all the utterances in the dialogue context inevitably introduces the noise as well as the redundant information, resulting in a poor result.

In this paper, we propose DialogueCSE, a dialogue-based contrastive learning approach to tackle these issues. We hold that the semantic matching relationships between the context and the response can be implicitly modeled through contrastive learning, thus making it possible to eliminate the gap between training and evaluating. To this end, we introduce a novel matching-guided embedding (MGE) mechanism. Specifically, MGE first pairs each utterance in the context with the response and performs a token-level dot-product operation across all the utterance-response pairs to obtain the multi-turn matching matrices. Then the multi-turn matching matrices are used as guidance to generate a context-aware embedding for the response embedding (i.e. the context-free embedding). Finally, the context-aware embedding and the context-free embedding are paired as a training sample, whose label is determined by whether the context and the response are originally from the same dialogue. Our motivation is that once the context semantically matches the response, it has the ability to distill the context-aware information from the context-free embedding, which is exactly the learning objective of the sentence encoder that aims to produce context-aware sentence embeddings.

We train our model on three multi-turn dialogue datasets: the Microsoft Dialogue Corpus (MDC) (Li et al., 2018), the Jing Dong Dialogue Corpus (JDDC) (Chen et al., 2020), and the E-commerce Dialogue Corpus (ECD) (Zhang et al., 2018). To evaluate our model, we introduce two types of tasks: the semantic retrieval (SR) task and the dialogue-based semantic textual similarity (D-STS) task. Here we do not adopt the standard semantic textual similarity (STS) task (Cer et al., 2017) for two reasons: (1) As revealed in (Zhang et al., 2020), the sentence embedding performance varies greatly as the domain of the training data changes. As a dialogue dataset is always about several certain domains, evaluating on the STS benchmark may mislead the evaluation of the model. (2) The dialogue-based sentence embeddings focus on context-aware rather than context-free semantic meanings, which may not be suitable to be evaluated through the context-free benchmarks. Since previous dialogue-based works have not set up a uniform benchmark, we construct two evaluation datasets for each dialogue corpus. A total of 18,964 retrieval samples and 4,000 sentence pairs are annotated by seven native speakers through the crowd-sourcing platform[1]. The evaluation results indicate that DialogueCSE significantly outperforms the baselines on the three datasets in terms of both MAP and Spearman's correlation metrics, demonstrating its effectiveness. Further quantitative experiments show that DialogueCSE achieves better performance when leveraging more dialogue context and remains robust when less training data is provided. To sum up, our contributions are threefold:

- We propose DialogueCSE, a dialogue-based contrastive learning approach with MGE mechanism for learning sentence embeddings from dialogues. As far as we know, this is the first attempt to apply contrastive learning in this area.
- We construct the dialogue-based sentence embedding evaluation benchmarks for three dialogue corpus. All of the datasets will be released to facilitate the follow-up researches.
- Extensive experiments show that DialogueCSE significantly outperforms the baselines, establishing the state-of-the-art results.

2 Related Work

2.1 Self-supervised Learning Approaches

Early works on sentence embeddings mainly focus on the self-supervised learning approaches. Kiros et al. (2015) train a seq2seq network by decoding the token-level sequences of the context in the corpus. Hill et al. (2016) propose to predict the neighboring sentences as bag-of-words instead of step-by-step decoding. Logeswaran and Lee (2018) perform sentence-level modeling by retrieving the ground-truth sentence from candidates under the given context, achieving consistently better performance compared to the previous token-level modeling approaches. The datasets used in these works

[1]All the datasets will be publicly available at https://github.com/wangruicn/DialogueCSE

are typically built upon the corpus of web pages and books (Zhu et al., 2015). As the semantic connections are relatively weak in these corpora, the model performances in these works are inherently limited and hard to achieve further improvement.

Recently, the pre-trained language models such as BERT (Devlin et al., 2019) and GPT (Radford et al.) yield strong performances across many downstream tasks (Wang et al., 2018). However, BERT's embeddings show poor performance without fine-tuning and many efforts have been devoted to alleviating this issue. Zhang et al. (2020) propose a self-supervised learning approach that derives meaningful BERT sentence embeddings by maximizing the mutual information between the global sentence embedding and all its local context embeddings. Li et al. (2020) argue that BERT induces a non-smooth anisotropic semantic space. They propose to use a flow-based generative module to transform BERT's embeddings into isotropic semantic space. Similar to this work, Su et al. (2021) replace the flow-based generative module with a simple but efficient linear mapping layer, achieving competitive results with reported experiments in BERT-flow.

Lately, the contrastive self-supervised learning approaches have shown their effectiveness and merit in this area. Wu et al. (2020); Giorgi et al. (2020); Meng et al. (2021) incorporate the data augmentation methods including the word-level deletion, reordering, substitution, and the sentence-level corruption into the pre-training of deep Transformer models to improve the sentence representation ability, achieving significantly better performance than BERT especially on the sentence-level tasks (Wang et al., 2018; Cer et al., 2017; Conneau and Kiela, 2018). Gao et al. (2021) apply a twice independent dropout to obtain two same-source embeddings from a single sentence as input. Through optimizing their cosine distance, SimCSE achieves remarkable gains over the previous baselines. Yan et al. (2021) empirically study more data augmentation strategies in learning sentence embeddings, and it also achieves remarkable performance as SimCSE. In this work, we propose the MGE mechanism to generate a context-aware embedding for each candidate response based on its context-free embedding. Different from previous methods built upon the data augmentation strategies, MGE leverages the context to accomplish this goal without any text corruption.

For dialogue, Yang et al. (2018) train a siamese transformer network with single-turn input-response pairs extracted from Reddit. Such architecture is further extended in (Reimers and Gurevych, 2019) by replacing the transformer encoder with BERT. Henderson et al. (2020) propose to leverage the dialogue context to improve the sentence embedding performance. They concatenate the multi-turn dialogue context into a long word sequence and adopt a similar architecture as (Yang et al., 2018) to model the context-response matching relationships. Our work is closely related to their works. We propose a novel dialogue-based contrastive learning approach, which directly models the context-response matching relationships without an intermediate MLP. We also consider the interactions between each utterance in the dialogue context and the response instead of simply treating the dialogue context as a long sequence.

2.2 Supervised Learning Approaches

The supervised learning approaches mainly focus on training classification models with the SNLI and the MNLI datasets (Bowman et al., 2015; Williams et al., 2018). Conneau et al. (2017) demonstrate the superior performance of the supervised learning model on both the STS-benchmark (Cer et al., 2017) and the SICK-R tasks (Marelli et al., 2014). Based on this observation, Cer et al. (2018) further extend the supervised learning to the multi-task learning by introducing the QA prediction task, the Skip-Thought-like task (Henderson et al., 2017; Kiros et al., 2015), and the NLI classification task, achieving significant improvement over InferSent. Reimers and Gurevych (2019) employ BERT as sentence encoders in the siamese-network and fine-tune them with the SNLI and the MNLI datasets, achieving the new state-of-the-art performance.

3 Problem Formulation

Suppose that we have a dialogue dataset $\mathcal{D} = \{S_i\}_{i=1}^{K}$, where $S_i = \{u_1, \cdots, u_{k-1}, r, u_{k+1}, \cdots, u_t\}$ is the i-th dialogue session in $\mathcal{D}$ with t turn utterances. r is the response and $C_i = \{u_1, \cdots u_{k-1}, u_{k+1}, \cdots, u_t\}$ is the bi-directional context around r. We omit the subscript i in the following paragraph and use S, C instead of S_i, C_i for brevity.

To generate the contrastive training pairs, we introduce two embedding matrices for r, named context-free embedding matrix and context-aware

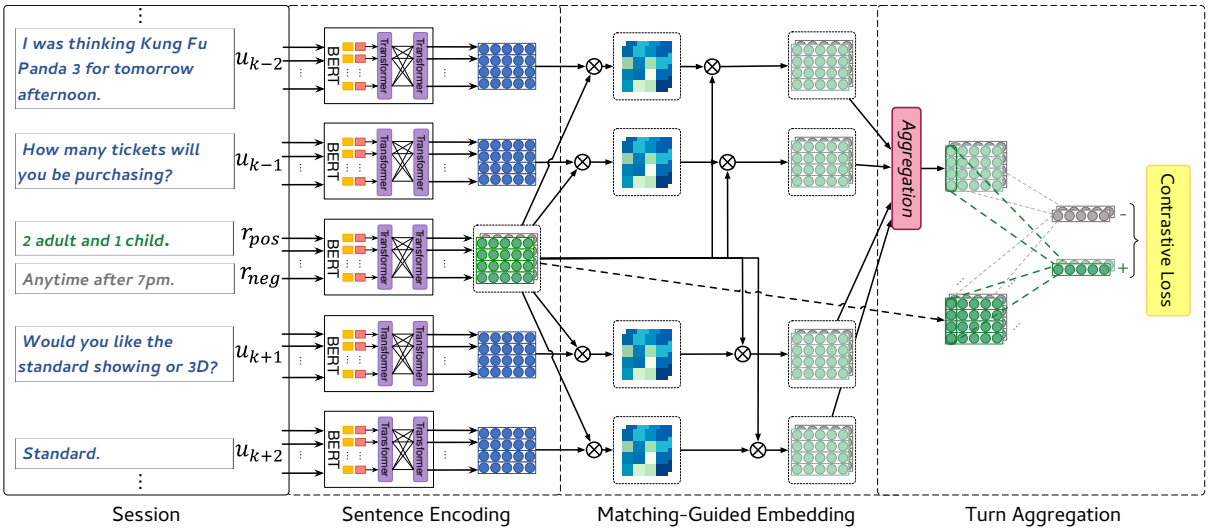

Figure 1: Model architecture. (1) We use BERT to encode the multi-turn dialogue context and the responses, all of the BERT encoders share the same parameters. (2) The matching-guided embedding (MGE) mechanism performs the token-level matching between each utterance and a response, generates multiple refined embeddings across turns. (3) All refined embedding matrices are aggregated to form a context-aware embedding matrix, which is further pooled along the sequence dimension.

embedding matrix. Specifically, we first encode r as an embedding matrix $\bar{\mathbf{R}}$. Since $\bar{\mathbf{R}}$ is encoded independently of the dialogue context, it is treated as the context-free embedding matrix. Then we generate a corresponding embedding matrix $\tilde{\mathbf{R}}$ based on $\bar{\mathbf{R}}$ according to the guidance of C. $\tilde{\mathbf{R}}$ is treated as the context-aware embedding matrix. As C and r are derived from the same dialogue, $(\bar{\mathbf{R}}, \tilde{\mathbf{R}})$ naturally forms a positive training pair. To construct a negative training pair, we first sample an utterance r' from a dialogue randomly selected from $\mathcal{D}$. r' is encoded as the context-free embedding matrix $\bar{\mathbf{R}}'$ based on which a context-aware embedding matrix $\tilde{\mathbf{R}}'$ is generated through the completely identical process. $(\bar{\mathbf{R}}', \tilde{\mathbf{R}}')$ is treated as a negative training pair. For each response r, we generate a positive training pair (since there is only one ground-truth response for each context) and multiple negative training pairs. All the training pairs are then passed through the contrastive learning module.

It is worth to mention that there is no difference between sampling the response or the context as they are symmetrical in constructing the negative training pairs. But we prefer the former as it is more straightforward and in accordance with the previous retrieval-based works for dialogues. With all the training samples at hand, our goal is to minimize their contrastive loss, thus fine-tuning BERT as a context-aware sentence encoder.

4 Our Approach

Figure 1 shows the model architecture. Our model is divided into three stages: sentence encoding, matching-guided embedding, and turn aggregation. We describe each part as below.

4.1 Sentence Encoding

We adopt BERT (Devlin et al., 2019) as the sentence encoder. Let u represent a certain utterance in C. u and r are first encoded as two sequences of output embeddings, which is formulated as:

$$\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n\} = \mathbf{BERT}(u), \quad (1)$$

$$\{\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_n\} = \mathbf{BERT}(r), \quad (2)$$

where $\mathbf{u}_i$, $\mathbf{r}_j$ represent the i-th and the j-th output embedding derived from u and r respectively. n is the maximum sequence length of both input sentences. $\forall i, j \in 1, 2, \cdots, n$, the shapes of $\mathbf{u}_i$ and $\mathbf{r}_j$ are $1 \times d$, where d is the dimension of BERT's outputs. We stack $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n\}$ and $\{\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_n\}$ to obtain the context-free embedding matrices $\bar{\mathbf{U}}$ and $\bar{\mathbf{R}}$, whose shapes are both $n \times d$.

4.2 Matching-Guided Embedding

The matching-guided embedding mechanism performs a token-level matching operation on $\bar{\mathbf{U}}$ and $\bar{\mathbf{R}}$ to form a matching matrix $\mathbf{M}$, which is formulated

as:

$$\mathbf{M} = \frac{\bar{\mathbf{U}}\left(\bar{\mathbf{R}}\right)^{\mathrm{T}}}{\sqrt{d}}, \tag{3}$$

Then it generates a refined embedding matrix $\hat{\mathbf{R}}$ based on the context-free embedding matrix $\bar{\mathbf{R}}$, which is given by:

$$\hat{\mathbf{R}} = \mathbf{M}\bar{\mathbf{R}} \tag{4}$$

$\hat{\mathbf{R}}$ is a new representation of r from the perspective of the utterance u. Note that as u is only a single turn utterance in C, we generate $t-1$ refined embedding matrices for r in total.

4.3 Turn Aggregation

After obtaining all of the refined embedding matrices across turns, we consider two strategies to fuse them to obtain the final context-aware embedding matrix $\tilde{\mathbf{R}}$. The first strategy adopts a weighted sum operation based on the attention mechanism, formulated by:

$$\tilde{\mathbf{R}} = \sum_i \alpha_i \hat{\mathbf{R}}_i, \tag{5}$$

where $i \in \{1, \cdots, k-1, k+1, \cdots, t\}$ and $\hat{\mathbf{R}}_i$ is the refined embedding matrix corresponding to the i-th turn utterance in the context. The attention weight α_i is decided by:

$$\alpha_i = \frac{\exp(\text{FFN}(\hat{\mathbf{R}}_i))}{\sum_j \exp(\text{FFN}(\hat{\mathbf{R}}_j))}, \tag{6}$$

where FFN is a two-layer feed-forward network with ReLU (Nair and Hinton, 2010) activation function. We denote this strategy as I_1. The second strategy I_2 directly sums up all the refined embeddings across turns, which is defined as:

$$\tilde{\mathbf{R}} = \frac{1}{t-1}\sum_i \hat{\mathbf{R}}_i, \tag{7}$$

For the negative sample r', we apply the same procedure to generate the context-free embedding matrix $\bar{\mathbf{R}}'$ and the context-aware embedding $\tilde{\mathbf{R}}'$. Each context-aware embedding matrix is then paired with its corresponding context-free embedding matrix to form a training pair.

As mentioned in the introduction, MGE holds several advantages in modeling the context-response semantic relationships. Firstly, the token-level matching operation acts as a guide to distill the context-aware information from the context-free embedding matrix. Meanwhile, it provides rich semantic matching information to assist the generation of the context-aware embedding matrix. Secondly, MGE is lightweight and computationally efficient, which makes the model easier to train than the siamese-network-based models. Finally and most importantly, the context-aware embedding $\tilde{\mathbf{R}}$ shares the same semantic space with $\bar{\mathbf{R}}$, which enables us to directly measure their cosine similarity. This is the key to successfully model the semantic matching relationships between the context and the response through contrastive learning.

4.4 Learning Objective

We adopt the NT-Xent loss proposed in (Oord et al., 2018) to train our model. The loss $\mathcal{L}$ is formulated as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\text{sim}(\bar{\mathbf{R}}_i,\tilde{\mathbf{R}}_i)/\tau}}{\sum_{j=1}^{M} e^{\text{sim}(\bar{\mathbf{R}}_j,\tilde{\mathbf{R}}_j)/\tau}}, \tag{8}$$

where N is the number of all the positive training samples and M is the number of all the training pairs associated with each positive training sample r. τ is the temperature hyper-parameter. $\text{sim}(\cdot,\cdot)$ is the similarity function, defined as a token-level pooling operation followed by the cosine similarity.

Once the model is trained, we take the mean pooling of BERT's output embeddings as the sentence embedding.

5 Experiments

We conduct experiments on three multi-turn dialogue datasets: the Microsoft Dialogue Corpus (MDC) (Li et al., 2018), the Jing Dong Dialogue Corpus (JDDC) (Chen et al., 2020), and the E-commerce Dialogue Corpus (ECD) (Zhang et al., 2018). Each utterance in these three datasets is originally assigned with an intent label, which is further leveraged by us in the heuristic strategy to construct the evaluation datasets.

5.1 Experimental Setup

5.1.1 Training

Table 1 shows the statistics information of these three datasets. The Microsoft Dialogue Corpus is a task-oriented dialogue dataset. It consists of three domains, each with 11 identical intents. The Jing Dong Dialogue Corpus is a large-scale customer service dialogue dataset publicly available from

Dataset	MDC	JDDC	ECD
# Total dialogues	10,087	1,024,196	1,020,000
# Total turns	74,685	20,451,337	7,500,000
# Total words	190,952	150,716,172	49,000,000
# Total intents	11	289	207

Table 1: Statistics of the datasets.

JD[2]. Although the dataset collected from the real-world scenario is quite large, it contains much noise which brings great challenges for our model. The E-commerce Dialogue Corpus is a large-scale dialogue dataset collected from Taobao[3]. The released dataset takes the form of the response selection task. We recover it to the dialogue sessions by dropping the negative samples and splitting the context into multiple utterances. We pre-process these datasets by the following steps: (1) We combine the consecutive utterances of the same speaker. (2) We discard the dialogues with less than 4 turns in JDDC and ECD since such dialogues are usually incomplete in practice.

5.1.2 Evaluation

We introduce the semantic retrieval (SR) and the dialogue-based STS (D-STS) tasks to evaluate our model. For the SR task, we construct evaluation datasets by the following steps: (1) we sample a large number of sentences with the intent labels as candidates. (2) the candidates are annotated with binary labels indicating whether the given sentence and its intent label are consistent. The inconsistent instances are directly discarded from the candidates. (3) for each sentence, we retrieval 100 sentences through BM25 (Robertson and Zaragoza, 2009) from the candidates, and assign each candidate sentence a label by whether its intent is consistent with the target sentence. We limit the number of positive samples to a maximum of 30 and keep approximately 7k, 7k, and 4k samples for MDC, JDDC, and ECD respectively.

For the D-STS task, we sample the sentence pairs from the dialogues following the heuristic strategies proposed by (Cer et al., 2017) to ensure there are enough semantically similar samples. The heuristic strategies include unigram-based and w2v-based KNN retrieval methods and random sampling from the candidates with the same intent labels. The sentence pairs are further annotated through the crowd-sourcing platform, with five degrees ranging from 1 to 5 according to their semantic relevance. We use the median number of annotated results as the semantic relevance degrees, obtaining 1k, 2k, and 1k sentence pairs for MDC, JDDC, and ECD respectively.

All annotations are carried out by seven native speakers. For the SR task, we adopt the Mean average precision (MAP) and the Mean reciprocal rank (MRR) metrics. Following previous works, we adopt Spearman's correlation metric for the D-STS task to assess the quality of the dialogue-based sentence embeddings.

5.2 Baselines

We evaluate our model against the two groups of baselines: self-supervised learning methods and dialogue-based self-supervised learning methods. The former is not designed for dialogues while the latter is.

5.2.1 Self-supervised learning methods

In this line, we consider the BERT-based methods, which include BERT (Devlin et al., 2019), domain-adaptive BERT (Gururangan et al., 2020), BERT-flow (Li et al., 2020), and BERT-whitening (Su et al., 2021). "Domain-adaptive BERT" means that we run continue pre-training with the dialogue datasets. BERT-flow and BERT-whitening are two BERT-based variants that transform BERT's sentence embedding to the isotropic semantic space.

For BERT, we use the `[CLS]` token embedding (denoted as BERT-CLS) and the average of the sequence output embeddings (denoted as BERT-avg) as the sentence embedding, and the same is true for domain-adaptive BERT. It should be noted that in related sentence embedding researches, domain-adaptive BERT is rarely considered since the training datasets are relatively small. Fortunately, the large-scale dialogue datasets allow us to explore whether the domain-adaptive pre-training is helpful for our tasks. We also adopt the average of GloVe word embeddings (Pennington et al., 2014) (denoted as Avg. GloVe) as the sentence embedding to compare with our results.

5.2.2 Dialogue-based self-supervised learning methods

In this line, we mainly consider the siamese-networks commonly applied in dialogue-based researches. Considering none of the previous works (Yang et al., 2018; Henderson et al., 2020) employs the pre-trained language model as encoder, we re-

[2]https://www.jd.com
[3]https://www.taobao.com

Model	Microsoft Corpus			Jing Dong Corpus			E-commerce Corpus		
	Corr.	MAP	MRR	Corr.	MAP	MRR	Corr.	MAP	MRR
Self-supervised models									
Avg. GloVe embeddings	36.64	31.59	40.91	39.61	45.94	59.53	19.80	46.14	63.68
BERT-CLS	22.34	29.54	35.94	21.40	45.05	59.58	16.61	47.75	65.91
BERT-avg	40.95	32.10	43.01	50.89	49.08	64.54	43.68	51.77	70.79
BERT-flow	45.56	33.13	40.86	65.11	49.53	64.30	55.04	52.16	71.06
BERT-whitening	26.70	32.09	43.01	61.57	49.08	64.54	47.64	51.77	70.80
BERT(adapt)-CLS	27.35	31.30	39.83	26.49	48.51	65.70	33.91	51.75	74.68
BERT(adapt)-avg	42.81	32.53	43.49	72.60	53.03	66.99	74.26	59.32	76.89
BERT(adapt)-flow	50.17	34.32	41.62	73.32	53.42	67.00	74.31	59.77	76.48
BERT(adapt)-whitening	29.68	32.53	43.48	67.18	53.04	67.01	57.22	59.33	76.84
Dialogue-based self-supervised models									
SiameseBERT$_S$	77.95	76.26	84.92	75.70	61.92	74.44	74.83	65.84	79.88
SiameseBERT$_M$	76.70	73.81	85.09	76.85	62.45	74.64	75.45	66.24	80.58
DialogueCSE$_{I_1}$	80.13	87.26	85.89	80.60	66.54	74.79	81.79	68.70	79.89
DialogueCSE$_{I_2}$	**82.36**	**91.40**	**90.45**	**81.22**	**68.02**	**79.52**	**83.94**	**69.32**	**81.20**

Table 2: Evaluation results on the dialogue-based semantic textual similarity (D-STS) task and the semantic retrieval (SR) task. Corr. refers to Spearman's correlation metric for the D-STS task. MAP and MRR are metrics for the SR task. Reported numbers are in percentages.

implement two BERT-based siamese-network models according to their original approaches. The first baseline SiameseBERT$_s$ is a siamese-network which shares the architecture with (Yang et al., 2018; Reimers and Gurevych, 2019). It is equipped with a non-linear activation function in the matching layer to model the heterogeneous matching relationships between the context and the response[4]. The second baseline SiameseBERT$_m$ has the similar architecture as (Henderson et al., 2020). It flattens the multi-turn context and takes the token sequence as input. There is also an MLP layer on top of the sentence encoders.

5.3 Implementation Details

Our approach is implemented in Tensorflow (Abadi et al., 2016) with CUDA 10.0 support. For all datasets, we continue pre-training BERT for approximately 0.5 epochs to improve its domain adaption ability as well as keeping the general domain information as much as possible. During the continue pre-training stage, we use a masking probability of 0.15, a learning rate of 2e-5, a batch size of 50, and a maximum of 10 masked LM predictions per sequence. During the contrastive learning stage, we freeze the bottom 6 layers of BERT to prevent catastrophic forgetting which simultaneously enables the model to be trained with larger batch size. Such a setting achieves the best performance in our experiments. The batch size, the learning rate, and the number of context turns are set to 20, 5e-5, and 3 respectively. The maximum sequence length is set to 100, 50, 50 for JDDC, MDC, and ECD for both continue pre-training stage and contrastive learning stage. All models are trained on 4 Tesla V100 GPUs.

5.4 Evaluation Results

Table 2 shows the main experimental results on the three datasets. From the table, we can observe that our model achieves the best performance in terms of all metrics across the three datasets. Compared to the results of the siamese-networks, our model achieves at least 4.41 points (77.95 $\rightarrow$ 82.36), 4.37 points (76.85 $\rightarrow$ 81.22), and 8.49 points (75.45 $\rightarrow$ 83.94) in terms of Spearman's correlation on MDC, JDDC, and ECD respectively. It also improves the MAP metric by 14.84 points (76.26 $\rightarrow$ 91.40), 5.57 points (62.45 $\rightarrow$ 68.02), and 3.08 points (66.24 $\rightarrow$ 69.32) in terms of MAP metric on the three datasets. There are even larger improvements between DialogueCSE and the domain-adaptive baselines including BERT(adapt) and its variants. We attribute this improvement to two main reasons: First, by introducing contrastive learning, DialogueCSE eliminates the gap between training and evaluating, gaining significant improvements on both SR and D-STS tasks. Second, DialogueCSE models the

[4]We use "heterogeneous" to describe the matching relationships for context-response pairs since they have different semantic meanings. As a comparison, the NIL-like sentence pairs have the "homogeneous" matching relationships.

semantic relationships in each utterance-response pair, which distills the important information at turn-level from the multi-turn dialogue context and achieves better performance.

Moreover, by comparing the performances of DialogueCSE$_{I_1}$ and DialogueCSE$_{I_2}$, we find that the weighted sum aggregation strategy surprisingly brings a significant deterioration on all metrics. We consider that this is because the weighted sum operation breaks down the turn-level unbiased aggregation process. Since the attention mechanism tends to provide shortcuts for the model to achieve its learning objective, the long-tail utterances in the context may be partially ignored, thus leading to a decline in embedding performance. We hold that we can completely dismiss the weighted sum aggregation strategy in DialogueCSE since the token-level matching operation in MGE has implicitly served this role.

We also notice that BERT(adapt) achieves significantly better performance than the original BERT, especially on JDDC and ECD. It demonstrates the importance of continued pre-training with the in-domain training data. Without such procedure, the in-domain data can't be fully exploited, making it difficult for the model to achieve satisfactory performance. This also indicates that the MLM pre-training task is indeed a powerful task to learn effective sentence embeddings from texts, especially when the domain training data is sufficient.

5.5 Discussion

We conduct comparison and hyper-parameter experiments in the following section to study how our model performs with different numbers of turns, data scales, temperature hyper-parameter, and numbers of negative samples.

5.5.1 Comparison with Baseline

In this section, we choose SiameseBERT$_m$ as a comparison method. MAP and Spearman's correlation metrics are adopted in these experiments.

Impact of turn number. Figure 2 shows the performance of our model and the baseline under different numbers of turns on all datasets. From the results, we observe that our model is indeed benefited from the multi-turn dialogue context, and it exhibits consistently better performance than the baseline. The performance of our model increases as the turn number increases until it approximately arrives at 3. When the turn number goes bigger, the performance of both models begins to drop. We believe that in this case, adding more dialogue context will bring too much noise. Since MGE acts as a noise filter at both token and turn level, it makes the model more robust when using more context turns.

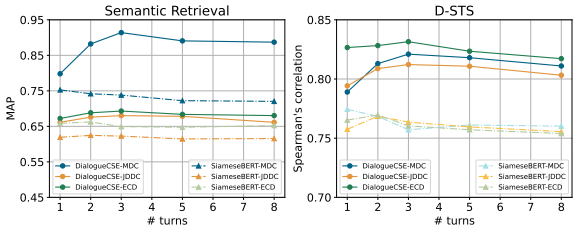

Figure 2: Impact of turn number.

Impact of data scale. We further explore whether our model is robust when fewer training samples are given. we select JDDC and ECD in this experiment since they are large-scale and topically diverse, which is suitable for simulating a few-shot learning scenario. Figure 3 shows the performances of our model and the baseline under different numbers of training dialogues.

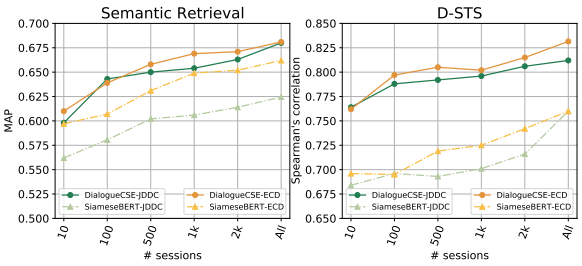

Figure 3: Impact of data scale.

As the figure reveals, the performance gaps between our model and the baseline are even larger when fewer training dialogue sessions are given. Particularly, when using only a few dialogues, our model can achieve even superior performance over the SiameseBERT trained on larger datasets, especially on the D-STS task. We think this is reasonable since the siamese-networks introduce a large amount parameters to model the semantic matching relationships, while our model accomplishes this goal without introducing any additional parameters.

5.5.2 Hyper-parameter Evaluations

We further conduct experiments on JDDC and EDC to study how our model is influenced by the temperature τ and the number of negative samples. The MDC dataset is excluded here since the semantics

Temperature		0.05	0.1	0.2	0.5
JDDC	Corr.	80.05	81.22	80.82	79.85
	MAP	67.19	68.02	67.55	68.63
ECD	Corr.	82.24	83.94	84.24	83.76
	MAP	67.98	69.32	69.63	69.11

Table 3: Impact of temperature.

# Negative samples		1	4	9	19
JDDC	Corr.	80.60	80.85	81.22	81.56
	MAP	67.48	67.69	68.02	68.63
ECD	Corr.	82.55	83.14	83.94	84.12
	MAP	68.56	68.87	69.32	69.56

Table 4: Impact of negative samples.

of its utterances are highly centralized around a few top intents.

Impact of temperature. Table 3 shows the experimental results with different τ values. We find that the Spearman's correlations increase monotonically as τ increases until 0.1 for JDDC and 0.2 for ECD, then they begin to drop. The MAP metrics also increase as τ increases until 0.1 for both datasets, but they remain stable as τ varies from 0.1 to 0.5. We consider this is due to the coarse-grained nature of the SR task. When τ approaches 0.1, our model can gradually distinguish among different fine-grained semantics, thus achieving better performance on both SR and D-STS tasks. As τ continues to increase, the model forces the sentence embeddings to be closer, resulting in a decrease in Spearman's correlation. However, as all positive samples in the candidates have identical labels, such degradation may not be fully reflected through the ranking metric (e.g. MAP) or even be covered as the number of retrieved positive samples changes.

Impact of negative samples. We vary the number of negative samples for each positive sample within $\{1, 4, 9, 19\}$. Table 4 shows the experimental results, from which we find that both metrics improve slightly when the number of negative samples increases. Considering the similar observation in (Gao et al., 2021; Yan et al., 2021), we conclude this phenomenon may be related to the discrete nature of language. Specifically, as the generation of the sentence embeddings in our approach is guided and constrained by the token-level interaction mechanism, our model is more robust than the other contrastive learning approaches and is even effective when only one negative sample is provided.

6 Conclusion

In this work, we propose DialogueCSE, a dialogue-based contrastive learning approach to learn sentence embeddings from dialogues. We also propose uniform evaluation benchmarks for evaluating the quality of the dialogue-based sentence embeddings. Evaluation results show that DialogueCSE achieves the best result over the baselines while adding no additional parameters. In the next step, we will study how to introduce more interaction information to learn the sentence embeddings and try to incorporate the contrast learning method into the pre-training stage.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283.

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *arXiv e-prints*, pages arXiv–1606.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. The jddc corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce

customer service. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 459–466.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

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

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

John M Giorgi, Osvald Nitski, Gary D Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. Convert: Efficient and accurate conversational representations from transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2161–2174.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems.*, page 3294—3302.

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.

Xiujun Li, Yu Wang, Siqi Sun, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv e-prints*, pages arXiv–1807.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223.

Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *arXiv preprint arXiv:2102.08473*.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 807–814.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Information Retrieval*, 3(4):333–389.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 164–174.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610.

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.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.