Learning Dialogue Representations from Consecutive Utterances

Zhihan Zhou * **Dejiao Zhang**† **Wei Xiao**† **Nicholas Dingwall**†
Xiaofei Ma† **Andrew O. Arnold**† **Bing Xiang**†
*Northwestern University †AWS AI Labs

Abstract

Learning high-quality dialogue representations is essential for solving a variety of dialogue-oriented tasks, especially considering that dialogue systems often suffer from data scarcity. In this paper, we introduce Dialogue Sentence Embedding (DSE), a self-supervised contrastive learning method that learns effective dialogue representations suitable for a wide range of dialogue tasks. DSE learns from dialogues by taking consecutive utterances[1] of the same dialogue as positive pairs for contrastive learning. Despite its simplicity, DSE achieves significantly better representation capability than other dialogue representation and universal sentence representation models. We evaluate DSE on five downstream dialogue tasks that examine dialogue representation at different semantic granularities. Experiments in few-shot and zero-shot settings show that DSE outperforms baselines by a large margin. For example, it achieves 13% average performance improvement over the strongest unsupervised baseline in 1-shot intent classification on 6 datasets. [2] We also provide analyses on the benefits and limitations of our model.

1 Introduction

Due to the variety of domains and the high cost of data annotation, labeled data for task-oriented dialogue systems is often scarce or even unavailable. Therefore, learning universal dialogue representations that effectively capture dialogue semantics at different granularities (Hou et al., 2020; Krone et al., 2020; Yu et al., 2021) provides a good foundation for solving various downstream tasks (Snell et al., 2017; Vinyals et al., 2016).

Contrastive learning (Chen et al., 2020; He et al., 2020) has achieved widespread success in representations learning in both the image domain (Hjelm et al., 2018; Lee et al., 2020; Bachman et al., 2019) and the text domain (Gao et al., 2021; Zhang et al., 2021a,b; Wu et al., 2020a). Contrastive learning aims to reduce the distance between semantically similar (positive) pairs and increase the distance between semantically dissimilar (negative) pairs. These positive pairs can be either human-annotated or obtained through various data augmentations, while negative pairs are often collected through negative sampling in the mini-batch.

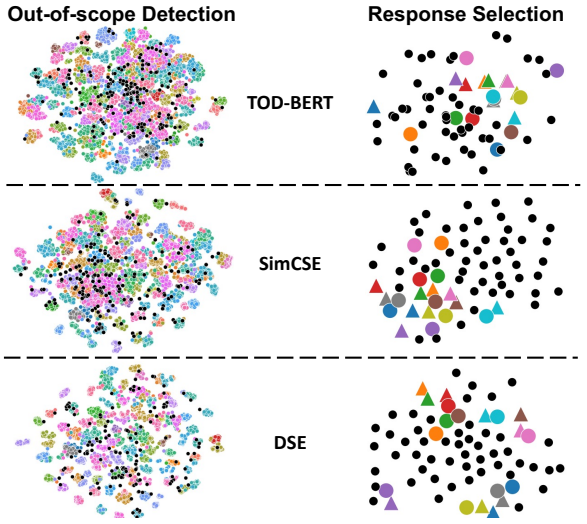

Figure 1: TSNE visualization of the dialogue representations provides by `TOD-BERT`, `SimCSE`, and `DSE`. **Left**: each color indicates one intent category, while the black circles represents out-of-scope samples. **Right**: items with the same color stands for query-response pairs, where triangles represent queries. The black circles represents randomly sampled responses.

In the supervised learning regime, Gao et al. (2021); Zhang et al. (2021a) demonstrate the effectiveness of leveraging the Natural Language Inference (NLI) datasets (Bowman et al., 2015; Williams et al., 2018) to support contrastive learning. Inspired by their success, a natural choice of dialogue representation learning is utilizing the

*Work done during an internship at AWS AI Labs.

[1]Throughout this paper, we use *utterance* to refer to all the sentences that belong to the same dialogue turn.

[2]The code and pre-trained models are publicly available at https://github.com/amazon-research/dse.

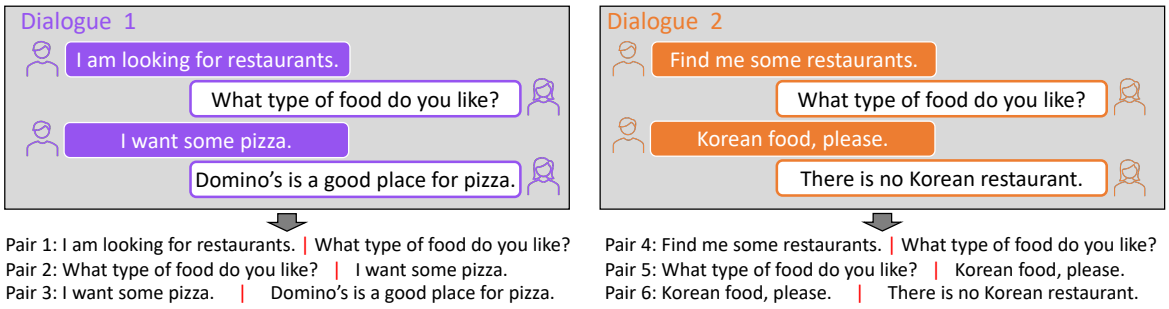

Figure 2: Illustration of the positive pair construction from dialogues.

Dialogue-NLI dataset (Welleck et al., 2018) that consists of both semantically entailed and contradicted pairs. However, due to its relatively limited scale and diversity, we found learning from this dataset leads to less satisfying performance, while the high cost of collecting additional human annotations precludes its scalability. On the other extreme, unsupervised representation learning has achieved encouraging results recently, among which SimCSE (Gao et al., 2021) and TOD-BERT (Wu et al., 2020a) set new state-of-the-art results on general texts and dialogues, respectively.

SimCSE uses Dropout (Srivastava et al., 2014) to construct positive pairs from any text by passing a sentence through the encoder twice to generate two different embeddings. Although SimCSE outperforms common data augmentations that directly operate on discrete text, we find it performs poorly in the dialogue domain (see Sec. 4.3). This motivates us to seek better positive pair constructions by leveraging the intrinsic properties of dialogue data. On the other hand, TOD-BERT takes an utterance and the concatenation of all the previous utterances in the dialogue as a positive pair. Despite promising performance on same tasks, we found TOD-BERT struggles on many other dialogue tasks where the semantic granularities or data statistics are different from those evaluated in their paper.

In this paper, inspired by the fact that dialogues consist of consecutive utterances that are often semantically related, we use consecutive utterances within the same dialogue as positive pairs for contrastive learning (See Figure 2). This simple strategy works surprisingly well. We evaluate DSE on a wide range of task-oriented dialogue applications, including intent classification, out-of-scope detection, response selection, and dialogue action prediction. We demonstrate that DSE substantially outperforms TOD-BERT, SimCSE, and some other sentence representation learning models in most scenarios. We assess the effectiveness of our approach by comparing DSE against its variants trained on other types of positive pairs (e.g., Dropout and Dialogue-NLI). We also discuss the trade-off in learning dialogue representation for tasks focusing on different semantic granularities and provide insights on the benefits and limitations of the proposed method. Additionally, we empirically demonstrate that using consecutive utterances as positive pairs can effectively improve the training stability (Appendix A.3).

2 Why Contrastive Learning on Consecutive Utterances?

When performing contrastive learning on consecutive utterances, we encourage the model to treat an utterance as similar to its adjacent utterances and dissimilar to utterances that are not consecutive to it or that belong to other dialogues.

On the one hand, this training process directly increases an utterance's similarity with its *true* response and decreases its similarities with other randomly sampled utterances. The ability to identify the appropriate response from many similar utterances is beneficial for dialogue ranking tasks (e.g., response selection). On the other hand, consecutive utterances also contain implicit categorical information, which benefits dialogue classification tasks (e.g., intent classification and out-of-scope detection). Consider pairs 1 and 4 in Figure 2: we implicitly learn similar representations of *I am looking for restaurants* and *Find me some restaurants*, since they are both consecutive with *What type of food do you like?*.

In contrast, SimCSE does not enjoy these benefits by simply using Dropout as data augmentation. Although TOD-BERT also leverages the intrinsic dialogue semantics by combining an utterance with its dialogue context as positive pair, the context is often the concatenation of 5 to 15 utterances.

Due to the large discrepancy in both semantics and data statistics between each utterance and its context, simply optimizing the similarity between them leads to less satisfying representations on many dialogue tasks. As shown in Section 4, TOD-BERT can even lead to degenerated representations on some downstream tasks when compared to the original BERT model.

3 Model

3.1 Notation

Let $\{(x_i, x_{i^+})\}_{i=1}^{M}$ be a batch of positive pairs, where M is the batch size. In our setting, each (x_i, x_{i^+}) denotes a pair of consecutive utterances sampled from a dialogue. Let e_i denote the representation of the text instance x_i that is obtained through an encoder. In this paper, we use mean pooling to obtain representations.

3.2 Training Target

Contrastive learning aims to maximize the similarity between positive samples and minimize the similarity between negative samples. For a contrastive anchor x_i, the contrastive loss aims to increase its similarity with its positive sample x_{i^+} and decrease its similarity with the other $2M-2$ negative samples within the same batch.

We adopt the *Hard-Negative* sampling strategy proposed by Zhang et al. (2021a), which puts higher weights on the samples that are close to the anchor in the representation space. The underlying hypothesis is that hard negatives are more likely to occur among those that are located close to the anchor in the representation space. Specifically, the Hard-Negative sampling based contrastive loss regarding anchor x_i is defined as follows:

$$\ell^{i,i^+} = -\log \frac{\exp(\text{sim}(e_i, e_{i^+})/\tau)}{\sum_{j \neq i} \exp(\alpha_{ij} \cdot \text{sim}(e_i, e_j)/\tau)} . \quad (1)$$

As mentioned above, here i and i^+ represent the indices of the anchor and its positive sample. We use τ to denote the temperature hyperparameter and $\text{sim}(e_i, e_j)$ represent the cosine similarity of e_i and e_j. In the above loss, α_{ij} is defined as follows,

$$\alpha_{ij} = \frac{\exp(\text{sim}(e_i, e_j)/\tau)}{\frac{1}{2M-2}\sum_{k \neq i^+} \exp(\text{sim}(e_i, e_k)/\tau)} . \quad (2)$$

Noted here, the denominator is averaged over all the other $2M$-2 negatives of x_i. Intuitively, samples that are close to the anchor in the representation space are assigned with higher weights. In other words, α_{ij} denotes the relative importance of instance x_j for optimizing the contrastive loss of the anchor x_i among all $2M$-2 negatives. For every positive pair (x_i, x_{i^+}), we respectively take x_i and x_{i^+} as the contrastive anchor to calculate the contrastive loss. Thereby, the contrastive loss over the batch is calculated as:

$$\mathcal{L} = \frac{1}{2M} \sum_{i=1}^{M} (\ell^{i,i^+} + \ell^{i^+,i}) \quad (3)$$

Here $\ell^{i^+,i}$ is defined by exchanging the roles of instances i and i^+ in Equation (1), respectively.

4 Experiments

We run experiments with five different backbones: $\text{BERT}_{\texttt{base}}$, $\text{BERT}_{\texttt{large}}$ (Devlin et al., 2018), $\text{RoBERTa}_{\texttt{base}}$, $\text{RoBERTa}_{\texttt{large}}$ (Liu et al., 2019b), $\text{DistilBERT}_{\texttt{base}}$ (Sanh et al., 2019). Due to the space limit, we only present the results on $\text{BERT}_{\texttt{base}}$ in the main text. The results of other models are summarized in Appendix D. We use the same training data as TOD-BERT for a fair comparison. We summarize the implementation details and data statistics of both pre-training and evaluation in Appendices A and B, respectively.

4.1 Baselines

We compare DSE against several representation learning models that attain state-of-the-art results on both general text and dialogue languages. We categorize them into the following two categories.

Supervised Learning `SimCSE-sup` (Gao et al., 2021) is the supervised version of SimCSE, which uses entailment and contradiction pairs in the NLI datasets (Bowman et al., 2015; Williams et al., 2018) to construct positive pair and hard negative pair accordingly. In a similar vein, `PairSupCon` (Zhang et al., 2021a) leverages the entailment pairs as positive pairs only while proposing an unsupervised hard negative sampling strategy that we summarized in Section 3.2. Following this line, we also evaluate DSE against its variant `DSE-dianli` trained on the Dialogue Natural Language Inference dataset (Welleck et al., 2018) by taking all the *entail* pairs as positive pairs.

Task	Dataset	Evaluation Setting	Num.
Intent Classification	Clinc150, Snips, Hwu64, Bank77, Appen-A, Appen-H	1-shot & 5-shot fine-tune 1-shot & 5-shot similarity	10 10
Out-of-scope Detection	Clinc150	1-shot & 5-shot similarity	10
Utterance-level Response Selection	AmazonQA	0-shot similarity 500-shot & 1000-shot fine-tune	1 5
Dialogue-level Response Selection	DSTC7-Ubuntu	0-shot similarity	1
Dialogue Action Prediction	DSTC2, GSIM	10-shot & 20-shot fine-tune	5

Table 1: Summarization of all the experimental settings. Please see Appendix B.2 for details of each dataset. The last column (**Num**) indicates the number of independent experiments with different random seeds (we report the averaged results). Since there is no randomness in 0-shot evaluations, we only run them once.

Unsupervised Learning TOD-BERT (Wu et al., 2020a) optimizes a contrastive response selection objective by treating an utterance and its dialogue context as positive pair. DialoGPT (Zhang et al., 2019) is a dialogue generation model that learns from consecutive utterance by optimizing a language modeling target.[3] SimCSE-unsup (Gao et al., 2021) uses Dropout (Srivastava et al., 2014) to construct positive pairs. In the general text domain, SimCSE-unsup has attained impressive performance over several explicit data augmentation strategies that directly operate on the discrete texts. To test its effectiveness in the dialogue domain, we compare DSE against its variant DSE-dropout where augmentations of every single utterance are obtained through Dropout.

The evaluations on DSE-dropout and DSE-dianli allow us to fairly compare our approach against the state-of-the-art approaches in both the supervised learning and the unsupervised learning regimes.

4.2 Evaluation Setting

To accommodate the fact that obtaining a large number of annotations is often time-consuming and expensive for solving the task-oriented dialogue applications, especially considering the variety of domains and certain privacy concerns, we mainly focus on few-shot or zero-shot based evaluations.

4.2.1 Evaluation Methods

Considering that only a few annotations are available in our setting, we mainly focus on the **similarity-based** evaluations, where predictions are made based on different similarity metrics applied in the embedding space without requiring updating the model.

We use different random seeds to independently construct multiple (See Table 1) few-shot train and validation sets from the original training data and use the original test data for performance evaluation. To examine whether the performance gap reported in the similarity-based evaluations is consistent with the associated fine-tuning approaches, we also report the **fine-tuning** results. We perform early stopping according to the validation set and report the testing performance averaged over different data splits.

4.2.2 Tasks and Metrics

We evaluate all models considered in this paper on two types of tasks: **utterance-level** and **dialogue-level**. The utterance-level tasks take a single dialogue utterance as input, while the dialogue-level tasks take the dialogue history as input. These two types of tasks assess representation quality on dialogue understanding at different semantic granularities, which are shared across a variety of downstream tasks.

Intent Classification is an utterance-level task that aims to classify user utterances into one of the pre-defined intent categories. We use Prototypical Networks (Snell et al., 2017) to perform the similarity-based evaluation. Specifically, we calculate a prototype embedding for each category by averaging the embedding of all the training samples that belong to this category. A sample is classified into the category whose prototype embedding is the most similar to its own. We report the classification accuracy for this task.

Out-of-scope Detection advances intent classification by detecting whether the sample is out-of-

[3] We use mean pooling of its hidden states as sentence representation, which leads to better performance than using only the last token. We use its *Medium* version that has twice as many parameters as $\textbf{BERT}_{\texttt{base}}$, since we found its *Small* version performs dramatically worse under our settings.

	$\text{BERT}_{\texttt{base}}$	Clinc150	Bank77	Snips	Hwu64	Appen-A	Appen-H	Ave.
1-shot	**SimCSE-sup**$^{\clubsuit}$	52.30	38.05	65.98	40.79	35.35	44.81	46.21
	PairSupCon$^{\clubsuit}$	55.34	41.30	65.20	41.43	37.55	47.55	48.06
	DSE-dianli$^{\clubsuit}$ **(ours)**	45.91	38.33	58.23	34.95	33.87	42.26	42.26
	BERT$^{\diamondsuit}$	36.98	22.05	62.51	27.74	13.19	18.74	30.20
	SimCSE-unsup$^{\diamondsuit}$	46.44	37.51	59.58	34.34	27.10	36.00	40.16
	DialoGPT$^{\diamondsuit}$	42.23	28.08	63.10	30.45	18.90	24.48	34.54
	TOD-BERT$^{\diamondsuit}$	36.67	27.11	62.52	29.52	20.61	26.68	33.85
	DSE-dropout$^{\diamondsuit}$ **(ours)**	46.48	30.02	65.03	33.25	16.94	21.77	35.58
	DSE$^{\diamondsuit}$ **(ours)**	**62.53**	**43.12**	**79.57**	**44.31**	**37.97**	**48.71**	**52.70**
5-shot	**SimCSE-sup**$^{\clubsuit}$	71.11	56.38	79.98	56.52	49.71	59.42	62.18
	PairSupCon$^{\clubsuit}$	73.88	60.07	76.14	55.75	**52.71**	62.23	63.46
	DSE-dianli$^{\clubsuit}$ **(ours)**	60.65	49.78	73.80	46.65	46.52	54.39	55.30
	BERT$^{\diamondsuit}$	59.48	38.73	78.65	43.15	21.39	27.61	44.83
	SimCSE-unsup$^{\diamondsuit}$	65.37	55.03	77.01	48.79	43.35	51.55	56.85
	DialoGPT$^{\diamondsuit}$	64.53	46.56	82.15	45.67	33.67	39.61	52.03
	TOD-BERT$^{\diamondsuit}$	57.74	42.98	79.68	42.32	33.58	42.52	49.80
	DSE-dropout$^{\diamondsuit}$ **(ours)**	70.46	49.95	80.10	52.16	30.00	37.48	53.36
	DSE$^{\diamondsuit}$ **(ours)**	**78.73**	**61.65**	**88.62**	**60.87**	52.32	**62.68**	**67.48**

Table 2: Results on similarity-based 1-shot and 5-shot Intent Classification. Predictions are made purely based on the embeddings provided by each model without any parameter tuning. All the models use $\text{BERT}_{\texttt{base}}$ as the backbone model. ♣: Supervised models. ♢: Unsupervised models.

scope, *i.e.,* does not belong to any pre-defined categories. We adapt the aforementioned Prototypical Networks to solve it. For a test sample, if its similarity with its most similar category is lower than a threshold, we classify it as out-of-scope. Otherwise, we assign it to its most similar category. For each model, we calculate the `mean` and `std` (standard deviation) of the similarity scores between every sample and its most similar prototype embedding, and take `mean−std` and `mean` as the threshold, respectively. The evaluation set contains both in-scope and out-of-scope examples. We evaluate this task with four metrics: 1) Accuracy: accuracy of both in-scope and out-of-scope detection. 2) In-Accuracy: accuracy reported on 150 in-scope intents. 3) OOS-Accuracy: out-of-scope detection accuracy. 4) OOS-Recall: recall of OOS detection.

Utterance-level Response Selection is an utterance-level task that aims to find the most appropriate response from a pool of candidates for the input user query, where both the query and response are single dialogue utterances. We formulate it as a ranking problem and evaluate it with Top-k-100 accuracy (a.k.a., k-to-100 accuracy), a standard metric for this ranking problem (Wu et al., 2020a). For every query, we combine its ground-truth response with 99 randomly sampled responses and rank these 100 responses based on their similarities with the query in the embedding space. The Top-k-100 accuracy represents the ratio of the ground-truth response being ranked at top-k, where k is an integer between 1 and 100. We report the Top-1, Top-3, and Top-10 accuracy of the models.

Dialogue-Level Response Selection is a dialogue-level task. The only difference with the *Utterance-level Response Selection* is that query in this task is dialogue history (e.g., concatenation of multiple dialogue utterances from different speakers). We also report the Top-1, Top-3, and Top-10 accuracy for this task.

Dialogue Action Prediction is a dialogue-level task that aims to predict the appropriate system action given the most recent dialogue history. We formulate it as a multi-label text classification problem and evaluate it with model fine-tuning. We report the Macro and Micro F1 scores for this task.

4.3 Main Results

Intent Classification & Out-of-scope Detection Tables 2 and 3 show the results of similarity-based intent classification and out-of-scope detection. The fine-tuning based results are presented in Appendix C. As we can see, DSE substantially outperforms all the baselines. In intent classification, it attains 13% average accuracy improvement over the strongest unsupervised baseline. More impor-

	BERT$_{\texttt{base}}$	Accuracy	In-Accuracy	OOS-Accuracy	OOS-Recall	Ave.
m-d	**SimCSE-sup**$^{\clubsuit}$	44.63	51.50	78.63	13.70	47.12
	PairSupCon$^{\clubsuit}$	51.87	54.34	82.33	40.75	57.32
	DSE-dianli$^{\clubsuit}$ **(ours)**	44.73	44.88	80.83	44.07	53.63
	BERT$^{\diamondsuit}$	33.96	36.01	80.58	24.77	43.83
	SimCSE-unsup$^{\diamondsuit}$	40.45	45.50	77.83	17.73	45.38
	DialoGPT$^{\diamondsuit}$	36.98	40.70	80.73	20.21	44.66
	TOD-BERT$^{\diamondsuit}$	34.77	36.28	79.74	27.98	44.69
	DSE-dropout$^{\diamondsuit}$	42.41	45.19	81.26	29.92	49.70
	DSE$^{\diamondsuit}$ **(ours)**	**58.74**	**60.52**	**84.07**	**50.72**	**63.51**
mean	**SimCSE-sup**$^{\clubsuit}$	36.90	29.07	47.04	72.12	46.28
	PairSupCon$^{\clubsuit}$	47.29	37.44	58.72	91.63	58.77
	DSE-dianli$^{\clubsuit}$ **(ours)**	40.70	30.78	57.67	85.30	53.61
	BERT$^{\diamondsuit}$	35.64	24.78	53.09	84.47	49.50
	SimCSE-unsup$^{\diamondsuit}$	37.65	28.99	49.38	76.62	48.16
	DialoGPT$^{\diamondsuit}$	38.04	27.00	52.87	87.75	51.42
	TOD-BERT$^{\diamondsuit}$	36.31	25.76	53.40	83.75	49.81
	DSE-dropout$^{\diamondsuit}$ **(ours)**	41.19	30.93	54.76	87.39	53.57
	DSE$^{\diamondsuit}$ **(ours)**	**50.88**	**41.72**	**60.64**	**92.11**	**61.34**

Table 3: Results on similarity-based 1-shot out-of-scope detection on `Clinc150` dataset. The out-of-scope threshold is respectively set as *mean* (`m`) and *mean-std* (`m-d`) of each sample's similarity with its closest category. See Sec. 4.2.2 for details. ♣: Supervised models. ♢: Unsupervised models.

BERT$_{\texttt{base}}$	AmazonQA			DSTC7-Ubuntu		
	Top-1 Acc.	Top-3 Acc.	Top-10 Acc.	Top-1 Acc.	Top-3 Acc.	Top-10 Acc.
SimCSE-sup$^{\clubsuit}$	47.03	62.40	76.80	11.37	19.40	33.53
PairSupCon$^{\clubsuit}$	52.22	65.09	76.85	**15.00**	23.02	**35.73**
DSE-dianli$^{\clubsuit}$ **(ours)**	49.16	63.36	76.66	14.92	22.73	34.72
BERT$^{\diamondsuit}$	29.70	43.86	60.36	6.75	12.97	24.20
SimCSE-unsup$^{\diamondsuit}$	48.02	62.45	76.00	10.03	17.13	29.37
DialoGPT$^{\diamondsuit}$	35.96	49.52	64.44	10.20	17.60	29.82
TOD-BERT$^{\diamondsuit}$	27.25	40.26	56.63	5.52	10.55	22.30
DSE-dropout$^{\diamondsuit}$ **(ours)**	37.80	51.64	66.58	9.55	16.97	28.80
DSE$^{\diamondsuit}$ **(ours)**	**56.62**	**70.54**	**81.90**	14.78	**23.10**	**35.73**

Table 4: Results on 0-shot response selection on `AmazonQA` (utterance-level) and `DSTC7-Ubuntu` (dialogue-level).

tantly, DSE achieves a 5%–10% average accuracy improvement over the supervised baselines that are trained on a large amount of expensively annotated data. The same trend was observed in out-of-scope detection, where DSE achieves 13%-20% average performance improvement over the strongest unsupervised baseline. The comparison between DSE, DSE-dropout, and DSE-dianli further demonstrates the effectiveness of using consecutive utterances as positive pairs in learning dialogue embeddings.

The left panel of Figure 1 visualizes the embeddings on the `Clinc150` dataset given by TOD-BERT, SimCSE, and DSE, which provides more intuitive insights into the performance gap. As shown in the figure, with the DSE embeddings, in-scope samples belonging to the same category are closely clustered together. Clusters of different categories are clearly separated with a large margin, while the out-of-scope samples are far away from those in-scope clusters.

Response Selection Table 4 shows the the results of similarity-based 0-shot response selection on utterance-level (AmazonQA) and dialogue-level (DSTC7-Ubuntu). Results of finetune-based evaluation on AmazonQA show similar trend and we summarize in Table 9 in Appendix. In Table 4, the large improvement attained by DSE over the baselines indicate our model's capability in dialogue response selection, in presence of both single-utterance query or using long dialogue history as query. The right panel of Figure 1 further illustrates this. It visualizes the embeddings of questions and

BERT$_{\text{base}}$	DSTC2		GSIM		
	10-shot	20-shot	10-shot	20-shot	Ave.
SimCSE-sup$^{\clubsuit}$	84.12 ‖ 36.62	86.15 ‖ 36.99	77.22 ‖ 35.03	84.75 ‖ 38.67	59.94
PairSupCon$^{\clubsuit}$	84.42 ‖ 36.52	86.22 ‖ 36.87	74.35 ‖ 33.44	82.26 ‖ 37.62	58.96
DSE-dianli$^{\clubsuit}$ **(ours)**	83.99 ‖ 36.20	86.74 ‖ 37.02	69.52 ‖ 31.36	79.97 ‖ 36.62	57.68
BERT$^{\diamondsuit}$	81.74 ‖ 34.78	86.98 ‖ 37.28	70.67 ‖ 31.24	77.60 ‖ 35.74	57.00
SimCSE-unsup$^{\diamondsuit}$	84.41 ‖ 36.62	87.84 ‖ 37.98	75.78 ‖ 34.47	81.73 ‖ 37.64	59.56
TOD-BERT$^{\diamondsuit}$	87.12 ‖ 36.83	88.59 ‖ 37.90	85.63 ‖ 38.53	92.15 ‖ 42.04	63.60
DSE-dropout$^{\diamondsuit}$ **(ours)**	83.23 ‖ 36.18	86.65 ‖ 36.95	72.25 ‖ 32.70	81.91 ‖ 37.33	58.62
DSE$^{\diamondsuit}$ **(ours)**	84.58 ‖ 36.02	88.01 ‖ 38.01	79.26 ‖ 35.89	86.73 ‖ 39.51	61.03
DSE$_{2\text{-}1}$ (ours)	84.47 ‖ 36.09	88.86 ‖ 38.41	83.81 ‖ 37.78	88.03 ‖ 40.29	62.22
DSE$_{3\text{-}1}$ (ours)	88.78 ‖ 38.52	89.59 ‖ 38.58	85.27 ‖ 39.10	88.65 ‖ 40.87	63.67
DSE$_{123\text{-}1}$ (ours)	**89.48 ‖ 38.60**	**90.97 ‖ 39.79**	**87.90 ‖ 40.05**	**92.48 ‖ 42.22**	**65.19**

Table 5: Results on 10-shot and 20-shot dialogue action prediction fine-tuning on DSTC2 amd GSIM. We use "‖" to separate the Micro F1 score and Macro F1 score. ♣: Supervised models. ♢: Unsupervised models.

answers in the AmazonQA dataset calculated by DSE, SimCSE, and TOD-BERT. With the DSE embedding, each question is placed close to its real answer while far away from other candidates.

Dialogue Action Prediction Table 5 shows that DSE outperforms all baselines except TOD-BERT, which indicates its capability in capturing dialogue-level semantics. To better understand TOD-BERT's superiority over DSE on this task, we further investigate this task and find its data format is special. Concretely, here each input consists of multiple utterances explicitly concatenated by using two special tokens [SYS] and [USR] to indicate the system and user inputs, respectively. For example, ([SYS] hi [USR] how are you? [SYS] I'm good). It follows the same format as the queries[4] used for training TOD-BERT, while DSE uses a single utterance as the query.

BERT$_{\text{base}}$	IC	OOS	u-RS	d-RS	DA
TOD-BERT	41.83	47.25	41.38	12.79	63.60
DSE	**60.09**	**62.43**	**69.69**	**24.54**	61.03
DSE$_{2\text{-}1}$	56.56	61.55	59.88	19.36	62.22
DSE$_{3\text{-}1}$	57.26	61.19	61.94	22.04	63.67
DSE$_{123\text{-}1}$	59.60	61.59	63.67	22.63	**65.19**

Table 6: Performance of TOD-BERT, DSE, and its variants on intent classification (IC), out-of-scope detection(OOS), response selection on utterance-level (u-RS) and dialogue-level (d-RS), dialogue action prediction (DA).

4.4 Trade-off in Query Construction

To understand the impact of using multiple utterances as queries, we train three new variants of DSE. Specifically, we construct positive pairs as: $(u_1$ [SEP] $u_2, u_3)$, $(u_2$ [SEP] $u_3, u_4)$, where u_i represents the i-th utterance in a dialogue. We use the [SEP] token to concatenate two consecutive utterances as query. We refer DSE trained with this data as DSE$_{2-1}$ since it uses 2 utterances as the query and 1 utterance as the response. Similarly, we train another variant DSE$_{3-1}$. Lastly, we also combine the positive pairs constructed for training DSE, DSE$_{2\text{-}1}$, and DSE$_{3\text{-}1}$ together to train another variant named DSE$_{123-1}$.

As shown in Table 5, by simply increasing the number of utterances within each query to three, DSE again outperforms TOD-BERT, and the improvement further expands when trained with the combined set, *i.e.,* DSE$_{123-1}$. Our results demonstrate that using long queries that consist of 5 to 15 utterances as what TOD-BERT does is not necessary even for dialogue action prediction. We further demonstrate this by evaluating DSE and its variants on all the other four tasks in Table 6, where our model outperforms TOD-BERT by a large margin. As it indicates, by using a single utterance as a query, DSE achieves a good balance among different dialogue tasks. In cases where dialogue action prediction is of great importance, augmenting the original training set of DSE with positive pairs constructed by using query consisting of 2 to 3 utterances is good enough to attain better performance while only incurring a slight performance drop on other tasks.

4.5 Potential Limitation

Considering the effectiveness of using consecutive utterances as positive pairs, a natural yet important question is: what are the potential limitations of

[4]We use *query* to refer the first utterance in a positive pair and use *response* to refer the other one

our proposed approach? When using consecutive utterances as positive pairs for contrastive learning, an assumption is that responses to the same query are semantically similar. Vice versa, queries that prompt the same answer are similar. This assumption holds in many scenarios, yet it fails sometimes.

It may fail when answers have different semantic meanings. Take the pairs 2 and 5 in Figure 2 as an example. Through our data construction, we implicitly consider *I want some pizza* and *Korean food, please* to be semantically similar since they are both positively paired with *What type of food do you like*. Although this may be correct in some coarse-grained classification tasks since these two sentences generally represent the same intent (e.g., order food), using them as positive pairs can introduce some noise when considering more fine-grained semantics. This problem is further elaborated when answers are general and ubiquitous, *e.g., Thank you*. Since these utterances can be used to respond to countless types of dissimilar queries, *e.g., I have booked a ticket for you* v.s. *Happy birthday*, we may implicitly increase the similarities among highly dissimilar utterances when training on these samples, which is undesirable.

We verify this on the NLI datasets, where the the task is to identify whether one sentence semantically entails or contradicts the anchor sentence. For each anchor sentence, we calculate its cosine similarities with both the true *entailment*, *contradiction* sentences in the representation space. We classify the sentence with higher cosine similarity with the anchor as entailment and the other as the contradiction. Despite DSE achieves better classification accuracy (76.62) than BERT (69.40) and TOD-BERT (70.51), it underperforms SimCSE-unsup (80.31). Although using dropout to construct positive pairs is not as effective as ours in many dialogue scenarios, this method better avoids introducing fine-grained semantic noise.

Despite the limitations, using consecutive utterances as positive pairs still leads to better dialogue representation than the elaborately labeled NLI datasets, indicating the great value of the information contained in dialogue utterances.

5 Related Work

Positive Pair Construction Popular supervised sentence representation learning often takes advantage of the human-annotated natural language inference (NLI) datasets (Bowman et al., 2015; Williams et al., 2018) for contrastive learning (Gao et al., 2021; Zhang et al., 2021a; Reimers and Gurevych, 2019; Cer et al., 2018). These sentence pairs either entail or contradict each other, making them the great choice for constructing positive and negative training pairs. Unsupervised sentence representation learning often relies on variant data augmentation strategies. Logeswaran and Lee (2018) and Giorgi et al. (2020) propose using sentences and their surrounding context as positive pairs. Other works resort to popular NLP augmentation methods such as word permutation (Wu et al., 2020b) and back-translation (Fang et al., 2020). Recently, Gao et al. (2021) demonstrates the superiority of using Dropout over other data augmentations that directly operate on the discrete texts.

Contrastive Learning Methods Contrastive learning is key to recent advances in learning sentence embeddings. Many contrastive learning approaches utilize memory-based methods, which draw negative samples from a memory bank of embeddings (Hjelm et al., 2018; Bachman et al., 2019; He et al., 2020). On the other hand, Chen et al. (2020) introduces a memory-free contrastive framework, SimCLR, that takes advantage of negative sampling within large mini-batches. Promising results were also reported in the NLP domain. To name a few, Gao et al. (2021) leverages both within batch negatives and the 'contradiction' annotations in NLI; and Zhang et al. (2021a) propose an unsupervised hard-negative sampling strategy.

Dialogue Language Model Learning dialogue-specific language models has attracted a lot of attention. Along this line, Zhang et al. (2019) adapts the pre-trained GPT-2 model (Radford et al., 2019) on Reddit data to perform open-domain dialogue response generation. Bao et al. (2019) evaluates multiple dialogue generation tasks after training on Twitter and Reddit data (Wolf et al., 2019; Peng et al., 2020). For dialogue understanding, Henderson et al. (2019b) propose a response selection approach using a dual-encoder model. They pre-train the response selection model on Reddit and then fine-tune it for different response selection tasks. Following this, Henderson et al. (2019a) introduces a more efficient conversational model that is pre-trained with a response selection target on the Reddit corpus. However, they did not release code or pre-trained models for comparison. Wu

et al. (2020a) combines nine dialogue datasets to obtain a large and high-quality task-oriented dialogue corpus. They introduce the TOD-BERT model by further pre-training BERT on this corpus with both the masked language modeling loss and the contrastive response selection loss.

6 Conclusion

In this paper, we introduce a simple contrastive learning method DSE that learns dialogue representations by leveraging consecutive utterances in dialogues as positive pairs. We conduct extensive experiments on five dialogue tasks to show that the proposed method greatly outperforms other state-of-the-art dialogue representation models and universal sentence representation methods. We provide ablation study and analysis on our proposed data construction from different perspectives, investigate the trade-off between different data construction variants, and discuss the potential limitation to motivate further exploration in representation learning on unlabeled dialogues. We believe DSE can serve as a drop-in replacement of the dialogue representation model (e.g., the text encoder) for a wide range of dialogue systems.

References

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. *arXiv preprint arXiv:1704.00057*.

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2019. Plato: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931*.

Samuel R. 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, Lisbon, Portugal. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. *arXiv preprint arXiv:1909.05358*.

Inigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mihail Eric and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414*.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

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*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Matthew Henderson, Inigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2019a. Convert: Efficient and accurate conversational representations from transformers. *arXiv preprint arXiv:1911.03688*.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019b. Training neural response selection for task-oriented dialogue systems. *arXiv preprint arXiv:1906.01543*.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. *arXiv preprint arXiv:2006.05702*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jason Krone, Yi Zhang, and Mona Diab. 2020. Learning to classify intents and slot labels given a handful of examples. *arXiv preprint arXiv:2004.10793*.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*.

Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. 2020. I-mix: A domain-agnostic strategy for contrastive representation learning. *arXiv preprint arXiv:2010.08887*.

Sungjin Lee, Hannes Schulz, Adam Atkinson, Jianfeng Gao, Kaheer Suleman, Layla El Asri, Mahmoud Adada, Minlie Huang, Shikhar Sharma, Wendy Tay, and Xiujun Li. 2019. Multi-domain task-completion dialog challenge. In *Dialog System Technology Challenges 8*.

Xiujun Li, Sarah Panda, JJ (Jingjing) Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. In *SLT 2018*.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019a. Benchmarking natural language understanding services for building conversational agents.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65.

Nikola Mrkšić, Diarmuid O Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2016. Neural belief tracker: Data-driven dialogue state tracking. *arXiv preprint arXiv:1606.03777*.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. *arXiv preprint arXiv:2002.12328*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Pararth Shah, Dilek Hakkani-Tur, Bing Liu, and Gokhan Tur. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51.

Jake Snell, Kevin Swersky, and Richard S Zemel. 2017. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638.

Mengting Wan and Julian McAuley. 2016. Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 489–498. IEEE.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2018. Dialogue natural language inference. *arXiv preprint arXiv:1811.00671*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Chien-Sheng Wu, Steven Hoi, Richard Socher, and Caiming Xiong. 2020a. Tod-bert: pre-trained natural language understanding for task-oriented dialogue. *arXiv preprint arXiv:2004.06871*.

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

Dian Yu, Luheng He, Yuan Zhang, Xinya Du, Panupong Pasupat, and Qi Li. 2021. Few-shot intent classification and slot filling with retrieved examples. *arXiv preprint arXiv:2104.05763*.

Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O Arnold, and Bing Xiang. 2021a. Pairwise supervised contrastive learning of sentence representations. *arXiv preprint arXiv:2109.05424*.

Dejiao Zhang, Wei Xiao, Henghui Zhu, Xiaofei Ma, and Andrew O Arnold. 2021b. Virtual augmentation supported contrastive learning of sentence representations. *arXiv preprint arXiv:2110.08552*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

A Pre-train

In this section, we present the training data, implementation details, and training stability of model pre-training.

A.1 Data

We utilize the corpus collected by TOD-BERT (Wu et al., 2020a) to construct positive pairs. This dataset is the combination of 9 publicly available task-oriented datasets: MetaLWOZ (Lee et al., 2019), Schema (Rastogi et al., 2020), Taskmaster (Byrne et al., 2019), MWOZ (Budzianowski et al., 2018), MSR-E2E (Li et al., 2018), SMD (Eric and Manning, 2017), Frames (Asri et al., 2017), WOZ (Mrkšić et al., 2016), CamRest676 (Wen et al., 2017). The combined dataset contains 100707 dialogues with 1388152 utterances over 60 domains. We filter out sentences with less or equal to 3 words and end up with 892835 consecutive utterances (for DSE) and 879185 unique sentences (for DSE-dropout). Note that, the training data of SimCSE-unsup consists of 1 million sentences from Wikipedia. That says, on the one hand, we use the same dataset as TOD-BERT but with our proposed data construction. On the other hand, we use a similar number of training samples as SimCSE-unsup. We believe such data construction makes the comparisons fair enough.

A.2 Hyperparameters

We add a contrastive head after the Transformer model and use the outputs of the contrastive head to perform contrastive learning. We use a two-layer MLP with size $(d \times d, d \times 128)$ as the contrastive head. We use Adam (Kingma and Ba, 2014) with a batch size of 1024 and a constant learning rate as the optimizer. We set the learning rate for contrastive head as $3e-4$ and the learning rate for the Transformer model as $3e-6$. The temperature hyperparameter τ is set as 0.05. We train the model for 15 epochs (see Appendix A.3 for more details) and save the model at the end for evaluation. We use the same hyperparameters across all the experiments for $\text{BERT}_{\texttt{base}}$, $\text{RoBERTa}_{\texttt{base}}$, and $\text{DistilBERT}_{\texttt{base}}$ models. For $\text{BERT}_{\texttt{large}}$ and $\text{RoBERTa}_{\texttt{large}}$, we change the batch size to 512 to fit it into the GPUs. Pre-training of the $\text{DistilBERT}_{\texttt{base}}$, $\text{BERT}_{\texttt{base}}$, and $\text{BERT}_{\texttt{large}}$ model respectively takes 3, 4, and 13 hours on 8 NVIDIA® V100 GPUs[5].

[5]Our codes and model are under the Apache-2.0 License

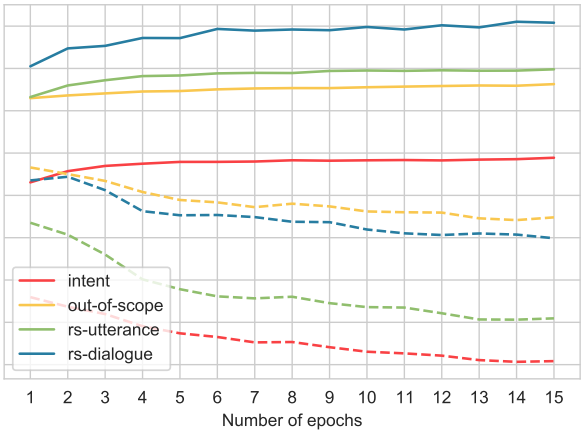

Figure 3: DSE and DSE-dropout's performance on each task at every epoch. The dashed lines represent the results of DSE-dropout.

A.3 Training Stability

In this section, we analyze the model's stability in terms of training steps when training with different type of positive pairs. We compare two data construction methods: consecutive utterances (DSE) and dropout (DSE-dropout). We first train each model for 15 epochs, save the checkpoint at the end of each epoch and evaluate each checkpoint with similarity-based methods. Figure 3 shows the two model's average performances on intent classification, out-of-scope detection, utterance-level response selection and dialogue-level response selection.

This result further illustrates the effectiveness of using consecutive utterances as positive pairs for learning dialogue representation. As shown in the figure, DSE's performance on all the tasks consistently improves during the training process, while DSE-dropout achieves the best performance at the first epoch and significantly loses performance afterwards. Besides, DSE's performance is less sensitive to the training steps. It achieves stable performance after about 5 epochs. In contrast, DSE-dropout's performance drops dramatically during the training process, yet it never surpasses DSE's performance. Therefore, we report DSE-dropout's performance at the first epoch in all the tables.

B Evaluation Setup

In this section, we present evaluation details and introduction to the evaluation dataset. Throughout this paper, we use *cosine similarity* as the similarity metric and *mean pooling of token embeddings* as the sentence representation. For baseline models, we report the better results of using its default setting (e.g., last hidden state of the [CLS] token as sentence embedding for SimCSE) and mean pooling.

B.1 Hyperparameters

We use the same hyperparameters for all the models. For similarity-based methods, the only hyperparameter is the max sequence length, we empirically choose a number that can fit at least 99% of the samples. We respectively set it as 64, 64, 128, and 128 for intent classification, out-of-scope detection, utterance-level response selection and dialogue-level response selection. Hyperparameters for fine-tune evaluations as listed as follows:

Intent Classification We fine-tune all the models for 50 epochs with a batch size of 16 and learning rate of $3e$-05. We evaluate the model on the few-shot validation set after every 10 steps. Early stopping is applied based on the model's validation results. The max sequence length is set as 64 and the dropout at the classification layer is set as 0.1.

Utterance-level Response Selection In this task, we set the max sequence length as 128 and batch size as 100. Other hyperparameters are same as those in Intent Classification. We use the original SimCLR loss (Chen et al., 2020) to optimize the model.

Dialogue Action Prediction In this task, we fine-tune all the models for 100 epochs with a batch size of 32 and learning rate of $5e$-05. We evaluate the model on the few-shot validation set after every 30 steps. Early stopping is also applied. The max sequence length is set as 32 since we find shorter inputs leads to much better performance for all the models. We truncate sentences from the head to keep the most recent dialogue utterances as model input. We set the dropout at the classification layer as 0.2.

B.2 Datasets

Intent Classification We use 4 popular publicly available datasets: Clinc150 (Larson et al., 2019) with 150 categories and 4500 test sample, Bank77 (Casanueva et al., 2020) with 77 categories and 3080 test sample, Snips (Coucke et al., 2018) with 7 categories and 1447 test sample, and Hwu64 (Liu et al., 2019a) with 64 categories and 3853 test sample. To apply the Clinc150 dataset

in intent classification, we remove all the out-of-scope samples. We also use an internal dataset named Appen, whose texts are transcribed from customer recording. This dataset contains 30 categories and 310 test samples. There are two versions of each sentence. One is transcribed by Automatic Speech Recognition (ASR), which includes some ASR noise (e.g., transcribe errors). The other is transcribed by human annotator. We refer them respectively as Appen-A and Appen-H.

Out-of-scope Detection We use the entire Clinc150 dataset, which contains 150 in-scope intents and one out-of-scope intent. There are 5500 test samples in total (4500 in-scope and 1000 out-of-scope).

Utterance-level Response Selection We use the AmazonQA dataset (Wan and McAuley, 2016), which contains 5334606 question-answer pairs about different products. Following Henderson et al. (2019a), we randomly select 300K pairs for model evaluation.

Dialogue-Level Response Selection We use the DSTC7-Ubuntu dataset (Lowe et al., 2017), which contains conversations about the Ubuntu system. Each query of this dataset comes together with one ground-truth response and 100 candidate responses. We combine the validation and test sets together for evaluation, which results in 6000 evaluation samples.

Dialogue Action Prediction We use the DSTC2 (Henderson et al., 2014) and GSIM (Shah et al., 2018) dataset processed by Wu et al. (2020a). These two datasets respectively contains 13/19 actions and 1117/1039 test samples. The average number of samples in 10-shot and 20-shot training is 79 and 149 for DSTC2; 60 and 120 for GSIM.

C Results of BERT$_{\texttt{base}}$

In this section, we present other evaluation results on the BERT$_{\texttt{base}}$ model, including 1-shot and 5-shot fine-tune on intent classification (Table 7), 5-shot similarity-based out-of-scope detection (Table 8), and 500-shot and 1000-shot fine-tune on AmazonQA response selection (Table 9).

D Results of Other Backbone Models

In this section, we present similarity-based evaluation results on other four backbone models: BERT$_{\texttt{large}}$, RoBERTa$_{\texttt{base}}$, RoBERTa$_{\texttt{large}}$, and DistilBERT$_{\texttt{base}}$. Table 10 shows the results of similarity-based intent classification and Table 11 shows the results of similarity based response selection on both utterance-level and dialogue-level. As shown in the tables, DSE leads to consistent and significant performance boost on all the backbone models.

	$\text{BERT}_{\texttt{base}}$	Clinc150	Bank77	Snips	Hwu64	Appen-A	Appen-H	Ave.
1-shot	**SimCSE-sup**♣	48.30	35.29	49.77	34.09	30.16	38.42	39.34
	PairSupCon♣	50.33	37.59	52.53	34.21	**33.58**	41.65	41.65
	DSE-dianli♣ **(ours)**	43.07	38.02	46.57	30.98	29.39	36.77	37.47
	BERT♢	37.01	24.28	52.05	26.36	17.87	19.71	29.55
	SimCSE-unsup♢	42.72	33.56	47.13	30.19	24.00	32.68	35.05
	TOD-BERT♢	39.48	26.12	46.13	26.81	13.45	23.26	29.21
	DSE-dropout♢ **(ours)**	41.89	30.10	44.46	28.06	18.68	20.48	30.61
	DSE♢ **(ours)**	**55.67**	**38.10**	**70.67**	**37.93**	32.68	**45.03**	**46.68**
5-shot	**SimCSE-sup**♣	85.49	70.16	88.90	68.86	61.71	73.29	74.74
	PairSupCon♣	85.15	71.00	86.01	68.12	**63.94**	73.68	74.65
	DSE-dianli♣ **(ours)**	81.87	69.33	82.49	64.40	58.90	68.52	70.92
	BERT♢	84.00	68.51	85.72	64.79	55.00	65.52	70.59
	SimCSE-unsup♢	83.35	70.08	86.82	65.61	59.68	70.03	72.60
	TOD-BERT♢	83.15	65.29	88.49	66.29	56.32	67.13	71.11
	DSE-dropout♢ **(ours)**	84.14	69.74	87.39	66.47	57.74	68.13	72.27
	DSE♢ **(ours)**	**86.67**	**71.52**	**92.56**	**70.71**	63.71	**75.10**	**76.71**

Table 7: Results of fine-tuning all the models for 1-shot and 5-shot Intent Classification for $\text{BERT}_{\texttt{base}}$ models. ♣: Supervised models. ♢: Unsupervised models

	$\text{BERT}_{\texttt{base}}$	Accuracy	In-Accuracy	OOS-Accuracy	OOS-Recall	Ave.
mean−std	**SimCSE-sup**♣	59.65	69.04	80.35	17.40	56.61
	PairSupCon♣	68.37	71.03	85.62	56.43	70.36
	DSE-dianli♣ **(ours)**	56.22	56.92	83.38	53.07	62.40
	BERT♢	53.92	57.25	81.94	38.95	58.01
	SimCSE-unsup♢	55.68	63.06	79.58	22.47	55.20
	DialoGPT♢	58.53	63.25	82.62	37.25	60.41
	TOD-BERT♢	53.49	56.12	81.75	41.64	58.25
	DSE-dropout♢ **(ours)**	64.21	67.39	83.59	49.92	66.28
	DSE♢ **(ours)**	**72.62**	**74.77**	**87.16**	**62.95**	**74.38**
mean	**SimCSE-sup**♣	42.50	34.95	48.26	76.50	50.55
	PairSupCon♣	56.19	47.53	62.79	**95.17**	65.42
	DSE-dianli♣ **(ours)**	46.87	37.75	59.96	87.92	58.13
	BERT♢	47.79	38.40	57.86	90.07	58.53
	SimCSE-unsup♢	45.13	36.82	52.01	82.51	54.12
	DialoGPT♢	49.64	40.26	57.49	91.81	59.80
	TOD-BERT♢	47.88	38.46	58.43	90.31	58.77
	DSE-dropout♢ **(ours)**	54.45	46.00	61.31	92.50	63.56
	DSE♢ **(ours)**	**59.20**	**51.35**	**64.79**	94.53	**67.47**

Table 8: Results on similarity-based 5-shot out-of-scope detection on Clinc150 dataset. The out-of-scope threshold is respectively set as *mean* (m) and *mean-std* (m−d) of each sample's similarity with its closest category. See Sec. 4.2.2 for details. All the models use $\text{BERT}_{\texttt{base}}$ as the backbone model. ♣: Supervised models. ♢: Unsupervised models.

$\text{BERT}_{\texttt{base}}$	AmazonQA 500-shot			AmazonQA 1000-shot		
	Top-1 Acc.	Top-3 Acc.	Top-10 Acc.	Top-1 Acc.	Top-3 Acc.	Top-10 Acc.
SimCSE-sup♣	59.02	72.90	84.40	60.24	73.91	85.17
PairSupCon♣	61.24	74.51	85.19	62.31	75.36	86.01
BERT♢	55.63	70.98	83.79	58.00	72.67	84.81
SimCSE-unsup♢	56.04	70.34	82.56	57.85	71.95	84.01
TOD-BERT♢	43.52	59.29	75.06	46.54	62.16	77.15
DSE-dropout♢ **(ours)**	57.66	72.02	83.72	58.66	72.86	84.67
DSE♢ **(ours)**	**61.71**	**75.66**	**86.49**	**63.02**	**76.47**	**87.55**

Table 9: Results on 500-shot and 1000-shot fine-tune evaluation on response selection on AmazonQA (utterance-level). All the models use $\text{BERT}_{\texttt{base}}$ as the backbone model. ♣: Supervised models. ♢: Unsupervised models.

		Clinc150	Bank77	Snips	Hwu64	Ave.
1-shot	**BERT$_{\texttt{large}}$**	31.71	20.47	54.31	25.24	32.93
	BERT$_{\texttt{large}}$-DSE	65.57	45.45	78.52	46.37	58.97
	RoBERTa$_{\texttt{base}}$	34.58	20.58	52.25	24.24	32.91
	RoBERTa$_{\texttt{base}}$-DSE	66.05	45.01	80.58	43.98	58.90
	RoBERTa$_{\texttt{large}}$	35.72	20.84	54.80	23.57	33.73
	RoBERTa$_{\texttt{large}}$-DSE	69.23	45.42	73.72	44.29	58.16
	DistilBERT$_{\texttt{base}}$	39.48	23.96	63.00	30.25	39.17
	DistilBERT$_{\texttt{base}}$-DSE	60.47	43.52	76.38	44.63	56.25
5-shot	**BERT$_{\texttt{large}}$**	46.78	33.53	70.89	37.06	47.06
	BERT$_{\texttt{large}}$-DSE	80.40	64.49	89.08	63.00	74.24
	RoBERTa$_{\texttt{base}}$	53.58	32.40	68.90	34.98	47.46
	RoBERTa$_{\texttt{base}}$-DSE	81.73	64.92	89.67	62.81	74.78
	RoBERTa$_{\texttt{large}}$	55.43	33.25	78.01	36.25	50.73
	RoBERTa$_{\texttt{large}}$-DSE	82.52	62.93	86.64	61.04	73.28
	DistilBERT$_{\texttt{base}}$	61.00	39.45	78.90	45.00	56.08
	DistilBERT$_{\texttt{base}}$-DSE	77.16	60.39	86.48	60.81	71.21

Table 10: Results on similarity-based 1-shot and 5-shot Intent Classification with different model as the backbone. DSE leads to significant and consistent performance boost for all the models.

BERT$_{\texttt{large}}$	**AmazonQA**			**DSTC7-Ubuntu**		
	Top-1 Acc.	**Top-3 Acc.**	**Top-10 Acc.**	**Top-1 Acc.**	**Top-3 Acc.**	**Top-10 Acc.**
BERT$_{\texttt{large}}$	27.97	41.30	57.04	6.10	11.08	22.31
BERT$_{\texttt{large}}$-DSE	59.63	73.46	84.12	16.40	24.56	36.51
RoBERTa$_{\texttt{base}}$	19.60	29.67	44.70	4.86	9.80	20.70
RoBERTa$_{\texttt{base}}$-DSE	55.69	70.01	81.68	15.86	24.25	37.38
RoBERTa$_{\texttt{large}}$	26.68	37.73	51.70	7.65	14.50	26.10
RoBERTa$_{\texttt{large}}$-DSE	58.13	71.65	82.20	18.66	27.70	40.93
DistilBERT$_{\texttt{base}}$	31.73	46.47	63.23	6.65	12.46	24.98
DistilBERT$_{\texttt{base}}$-DSE	56.36	70.11	81.51	14.56	22.78	35.63

Table 11: Results on 0-shot response selection on AmazonQA (utterance-level) and DSTC7-Ubuntu (dialogue-level). DSE leads to significant and consistent performance improvements on all the models.