Dialogue Representation Learning: A New Benchmark and Weighted Contrastive Learning Approach

Anonymous ACL submission

Abstract

High-quality pre-trained text representations are powerful tools for various downstream tasks. However, dialogue representation learning has received less attention compared to tasks such as sentence representation learning. This can be attributed to two main challenges: 1) the lack of standard evaluation benchmarks on dialogue representation learning, and 2) the complexity of incorporating dialogue corpus into existing representation learning paradigms. To overcome these challenges, we present the first comprehensive evaluation benchmark called **DiaEval** (**Dia**logue Representation **Eval**uation Benchmark), which covers 5 datasets across 3 tasks including action prediction, dialogue inference, and response retrieval. These datasets are meticulously selected to ensure their comprehensiveness and representativeness. Second, we propose a new dialogue embedding method called **WMDC** (**W**eighted **M**ulti-window-sized **D**ialogue **C**ontrastive learning). WMDC leverages multiple context windows and sample reweighting with contrastive learning to obtain universal dialogue embeddings. The use of multiple context windows allows flexible encoding with multiple granularity while the reweighting method addresses the anisotropy and lack of informativeness issues within the learned dialogue embedding space. Through extensive comparison with various competitive baselines, WMDC achieves state-of-the-art performance on all tasks demonstrating its effectiveness and scalability.

1 Introduction

Learning universal text representation (Pennington et al., 2014; Conneau et al., 2017; Cer et al., 2018; Reimers and Gurevych, 2019) is proven to be effective for various downstream tasks. State-of-the-art text representation techniques can attain competitive performance with significantly reduced training data, in contrast to less effective ones (Xiong et al., 2022; Sarkar et al., 2022). Prior research (Wu et al., 2020; Zhang et al., 2020) has highlighted the inherent disparities in linguistic patterns between conversations and plain texts, underscoring the potential shortcomings in dialogue representations. Consequently, addressing this shortfall entails a greater demand for annotated data to achieve comparable performance in natural language tasks involving conversational contexts. However, dialogue-related human annotations are generally much harder and more expensive to conduct due to their complex and nuanced form compared with plain texts or sentences. Therefore, learning proficient general dialogue representation (Liu et al., 2021a, 2022; Bai et al., 2022) becomes an important task, even though it has received relatively less attention and remains underexplored in comparison to the domain of sentence representation learning.

The availability of a standardized evaluation benchmark (Wang et al., 2018; Nie et al., 2020; Zhang et al., 2022a) is crucial to facilitate the development of its research field. In the realm of sentence representation learning, SentEval (Conneau and Kiela, 2018) plays a vital role by facilitating assessments of sentence embeddings' capacity to encapsulate diverse aspects of meaning related to textual similarity, inference, and so on. Nonetheless, there is currently no established benchmark for evaluating dialogue representations. Recent works (Xu and Zhao, 2021; Liu et al., 2021a; Xu and Zhao, 2021) in this area rely on diverse evaluation tasks and metrics to assess their efficacy. The diversity of evaluation methods leads to a lack of consistency in evaluation, underscoring the necessity for the establishment of a standardized benchmark.

Furthermore, prior dialogue representation learning techniques often lack generality. They suffer from performance degradation when confronted with out-of-domain data, primarily because they

are optimized for specific domains and datasets, thereby greatly limiting their applicability in real-world situations. Therefore, an imminent need arises for universal dialogue representation, which is capable of capturing generic information applicable across a wide spectrum of tasks and domains. TODBERT (Wu et al., 2020) and DSE (Zhou et al., 2022b) have demonstrated enhanced generalizability by leveraging a diverse set of nine different dialogue corpora spanning 60 distinct domains, totaling approximately 0.1 million utterances. They employ a contrastive learning mechanism to differentiate dialogue contexts between correct and incorrect responses. Nevertheless, both approaches fall short in fully harnessing the training dialogue datasets, failing to acquire a fine-grained multi-granularity representation, due to their approach of contrasting a single utterance with a fixed window size context.

Dialogue corpora frequently display imbalanced distributions. An examination of several widely used dialogue datasets reveals that the top 1% most frequent utterances represent a substantial portion, ranging from 15% to 30% of the entire corpus (refer to Figure 1). These frequently occurring utterances typically lack substantial information and are applicable across a wide array of dialogue contexts. Optimizing the standard contrastive learning object on these datasets can have detrimental effects, including: a) encouraging a significant portion of dialogue embeddings clustering around high-frequency utterances, resulting in an unsatisfying anisotropic embedding space; b) implicitly drawing distinct dialogue contexts closer together, leading to a less informative embedding space; c) introducing a high false negative rate within a single batch, impeding the optimization process.

To solve the problems mentioned above, in this paper, we first present a new **Dia**logue Representation **Eval**uation Benchmark called **DiaEval**. This benchmark consists of five datasets, covering three distinct tasks designed to assess the dialogue understanding capabilities of different representation learning methods. Furthermore, we propose **WMDC** – a new **W**eighted **M**ulti-window-sized **D**ialogue **C**ontrastive learning method. We better utilize existing dialogue corpora by contrasting unfixed window-size contexts that share the same response. We project diverse window-sized embeddings into different representation spaces to alleviate the inherent semantic gap. Additionally, we identify and tackle the issue of distribution imbalance in existing dialogue corpora by utilizing a reweighted contrastive learning object based on inverse response frequency. Through extensive experiments, our proposed method achieves SOTA performances on all tasks among a broad range of baselines and shows strong scalability.[1]

Our main contributions can be summarized as follows:

- We introduce a novel dialogue representation evaluation benchmark (DiaEval). This benchmark encompasses five datasets, covering three tasks commonly employed for assessing dialogue understanding.
- We propose WMDC, a sample-reweighted multi-window-sized dialogue contrastive learning method.WMDC not only optimizes corpus utilization but also effectively tackles the challenges presented by distribution imbalances in dialogue corpora, including anisotropic uninformative embedding spaces and optimization issues.
- We collect a dialogue corpus consisting of 17 million dialogues and 37 million utterances used for unsupervised dialogue representation learning. We demonstrate the efficacy and the scalability of our proposed method through comprehensive experiments.

2 Related Work

2.1 Contrastive Learning

Learned embeddings in PLMs can cluster at a narrow cone in the vector space rather than distribute uniformly, which can severely limit representation quality. This anisotropy problem is naturally connected to uniformity (Wang and Isola, 2020) and can be intuitively eased by optimizing the contrastive learning object. Contrastive learning aims to learn effective representation by pulling together semantically close neighbors (positive pair) while pushing away the unrelated ones (negative pair) (Hadsell et al., 2006). SimCSE (Gao et al., 2021) greatly advances state-of-art sentence embeddings by simply leveraging dropout as the minimal positive pair construction method and in-batch negatives. In the same vein, DiffCSE (Chuang et al., 2022) is an instance of equivariant contrastive learning (Dangovski et al., 2022). It learns sentence

[1]The code will be available here.

embeddings sensitive to specific types of augmentations. The edited sentence is obtained by firstly stochastically masking out the original one and then sampling from a masked language model. Liu et al. (2023); Seonwoo et al. (2023) extend contrastive learning with ranking information among sentences to learn more fine-grained semantics. Other recent works (Wang et al., 2022; Yan et al., 2021; Liu et al., 2021b; Carlsson et al., 2021) explore to contrast different views of the same sentence or document, by data augmentation or different copies of models. Contrastive learning can be applied to various topics beyond text embedding, like classification (Zhou et al., 2022a; Chen et al., 2022), named entity recognition (Ying et al., 2022; Huang et al., 2022), multi-modal alignment (Radford et al., 2021; Guzhov et al., 2022; Zhang et al., 2022b).

2.2 Dialogue Representation Learning

Pretrained with response selection task, TODBERT (Wu et al., 2020) and DSE (Zhou et al., 2022b) learn universal dialogue representation by employing contrastive objectives on massive dialogue corpora. They differ in positive pair construction methods. TODBERT uses an utterance and the concatenation of all its previous utterances in the same dialogue as positive pairs, while DSE only uses two consecutive utterances to enhance the embedding. Also, there are studies addressing the non-flat nature of dialogues (Bonial et al., 2020; Bai et al., 2021, 2022; Banarescu et al., 2013). They excel in abstracting core semantic knowledge and reducing data sparsity by leveraging AMR (Abstract Meaning Representation) or designing new parsing schemes. However, parsing unstructured dialogues into structured data is computationally expensive and prone to errors, sacrificing its scalability for dialogue pretraining.

3 DiaEval: Dialogue Representation Evaluation Benchmark

In this section, we introduce DiaEval, a benchmark designed for assessing the quality of universal dialogue representations. Drawing inspiration from SentEval, DiaEval comprises 5 datasets encompassing a variety of dialogue-level tasks, including action prediction, dialogue inference, and response selection. Furthermore, we deliberately select datasets with diverse topics and domains. The choice of these tasks is based on a consensus within the community regarding the most suitable evaluations for assessing universal dialogue understanding. Additionally, these three tasks are pivotal in the context of industrial applications of dialogue systems. The action prediction task informs the system about the aspect in which to respond, the response retrieval task seeks an appropriate answer, and the inference tasks assess the consistency of the answer with the dialogue history.

DiaEval offers a comprehensive evaluation framework for assessing the quality of dialogue representations and aims to facilitate the development of dialogue representation learning methods. The detailed statistics are shown in Table 1.

3.1 Evaluation Settings

Our benchmark focuses on dialogue-level tasks and comprises two types: classification and retrieval. For classification tasks, including action prediction and dialogue inference, we add one MLP layer on top of the fixed encoder and only tune this added layer. We use grid-search on the validation set to find the best hyperparameters to avoid the randomnesses (Conneau and Kiela, 2018). To evaluate the encoder's performance on a specific task, we calculate the average score across all datasets in that task, allowing for different hyperparameters for each dataset. For retrieval tasks, DSTC7-Ubuntu, we calculate the cosine similarity between two representations without the need for any additional parameters.

3.2 Action Prediction

Action prediction uses the most recent dialogue history as input to predict the action labels of the next utterance. This task is formulated as a multi-label classification problem because a dialogue system response can contain multiple actions, such as informing and inquiring simultaneously.

We concatenate the most recent dialogue history X, pass it through the encoder F, and classify the encoding using MLP before applying a Sigmoid layer. The output of this process is the action label A.

$$A = \text{Sigmoid}\left(MLP(F(X))\right). \quad (1)$$

The model predicts an action label if its probability exceeds 0.5.

DSTC2 (Henderson et al., 2014), introduced as the second dialogue state tracking challenge, is a human-machine interactive dataset labeled with 19

Task Category	Datasets	# Samples (train / val / test)	# Labels	Metrics
Action Prediction	GSIM	11,831 / 2,837 / 6,505	13	Macro F1
	DSTC2	20,130 / 6,856 / 17,546	19	Macro F1
Dialogue Inference	IC-TOD	2,553 / 319 / 318	2	Accuracy
	DECODE	31,011 / 1,650 / 1,650	3	
Response Retrieval	DSTC7-Ubuntu	- / - / 6,000	-	TOP@N

Table 1: Summarization of datasets and tasks included in DiaEval.

actions related to restaurant search. It consists of dialogues under a spoken dialogue system, where the utterances are automatically transcribed using ASR (Automatic Speech Recognition) techniques. This can result in transcription errors and make natural dialogue understanding challenging. We utilize the processed dataset by Wu et al. (2020).

GSIM (Shah et al., 2018a) is a human-rewrote machine-machine interactive dataset, with 3k dialogues in the restaurant and movie domains. It contains 13 different system dialogue acts and was collected using an M2M approach (Shah et al., 2018b), which combines self-play and crowdsourcing steps to obtain high-quality dialogues with considerable diversity, coverage, and correctness. Similar to the above dataset, we use the processed data provided by Wu et al. (2020).

3.3 Dialogue Inference

Dialogue inference involves detecting contradictions within a dialogue, such as inconsistencies in persona, logic, and knowledge. The goal is to help dialogue systems determine whether a response aligns with the dialogue history.

We approach this task as a multi-class classification problem. We use a softmax function applied to the output of a multi-layer perceptron (MLP) to obtain the prediction probabilities. Denoting the label set as L, the input text as X, and the dialogue encoder F, the formulation is the same as the action prediction task:

$$L = \text{Softmax}\left(MLP(F(X))\right). \quad (2)$$

The label with the highest probability is considered the predicted label.

IC-TOD (Qin et al., 2021) is proposed to detect various types of consistency issues in task-oriented dialog systems. It spans three distinct tasks in the in-car personal assistant space: calendar scheduling, weather information retrieval, and point-of-interest navigation. The type of inconsistency is annotated. These types of inconsistencies include dialogue history inconsistency, user query inconsistency, and knowledge base inconsistency. However, in our benchmark, we only focus on the first two types and disregard the knowledge base. A dialogue is considered inconsistent if it contradicts either the dialogue history or the user query.

DECODE (Nie et al., 2021) consists of dialogues labeled as either contradiction or non-contradiction. The dataset is collected from four pre-selected open-domain dialogue source corpora, encompassing both human-human and human-bot interactions. These dialogues cover a wide range of conversational topics and require logical and context-related reasoning beyond personal facts. To ensure computational efficiency, we only consider dialogues with less than 256 tokens for this benchmark.

3.4 Response Retrieval

Response retrieval assesses models' ability to select the most appropriate response from a candidate pool based on a given dialogue history. The model takes the concatenated dialogue history as input and aims to retrieve the next utterance that is contextually relevant and coherent with the previous dialogue turns.

We frame it as a ranking problem. We calculate similarity scores between the given history C and response R_i from the candidate pool as follows:

$$r_i = \text{Cosine}\left(F(C), F\left(R_i\right)\right). \quad (3)$$

DSTC7-Ubuntu (Lowe et al., 2017) is an updated version of Ubuntu Dialogue Corpus, which is a large dataset consisting of multi-turn dialogues between two participants. Training, valid, and test sets are separated based on time. This dataset has gained popularity due to its long context, large size, and approximate power law relationship between the number of dialogues and turns per dialogue. In our evaluation, We utilize both the validation and

test sets to assess the model's ability to retrieve suitable responses in a zero-shot retrieval scenario. This means that the model is evaluated without any prior training on the dataset, testing its capability to select appropriate responses solely based on the given dialogue context.

4 Methods

In this section, we delve into the datasets employed for pretraining and introduce our dialogue embedding method, WMDC. WMDC is composed of two pivotal elements: positive pair construction techniques and the contrastive loss function reweighted with inverse response frequency.

4.1 Unsupervised Training Corpus

To ensure a thorough and impartial comparison, we employ the identical training corpus as TOD-BERT, which contains approximately 0.1 million dialogues, and 1.4 million utterances, spanning across 60 domains.

Scalability becomes a crucial consideration in the progress of dialogue modeling. To assess our method's scalability, we've compiled a significantly larger dialogue corpus, consisting of 17 million dialogues and 37 million utterances. In comparison to the dataset mentioned earlier, this training corpus contains approximately 172 times more dialogues and 27 times more utterances.

Details of both corpora can be found in Appendix C.

4.2 Our Method: WMDC

In this sub-section, we present WMDC, a sample-reweighted multi-window-sized dialogue contrastive learning method. It contains two main parts: a new weighted contrastive loss, and a new positive pair construction method.

4.2.1 Weighted Contrastive Object

Imbalanced data distribution is a common occurrence in real-world scenarios. We analyze several most popular dialogue datasets, finding no exceptions. As shown in figure 1, the top 1% of most repetitive expressions account for a substantial proportion (15% to 30%) of the entire dataset.

NT-Xent loss (Chen et al., 2020) is a widely adopted loss function for contrastive learning. It optimizes the model by bringing together the semantically related pairs while pushing away the unrelated ones. However, it encounters challenges

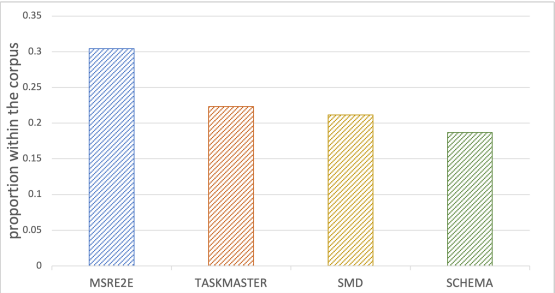

Figure 1: Data imbalance in the existing dialogue corpus. We arrange the utterances in descending order of their frequency. The top 1% of the most frequent utterances constitute a substantial portion of the entire corpus.

when dealing with imbalanced data. In the embedding space, the more frequent an utterance is, the more utterances will be pulled into its vicinity, resulting in the formation of numerous large clusters of vectors. This leads to an anisotropic embedding space where vectors are unevenly distributed in terms of directions, significantly limiting representation expressiveness, as noted in (Gao et al., 2021). Furthermore, these embedding clusters can contain distinct dialogue contexts, due to shared responses. Most common expressions like "yes" or "thank you" carry limited information, which further diminishes the informativeness of the embeddings as illustrated in Appendix 5. Finally, unsupervised contrastive learning always replies to big batch sizes to boost its performance. However, as we increase batch sizes, the potential for an elevated false negative rate arises due to the presence of a growing number of appropriate responses within a single batch. Optimization becomes challenging when the data is riddled with false labels.

To tackle this challenge, we enhance the NT-Xent loss by incorporating sample reweighting based on the inverse response frequency. The primary goal is to allocate less optimization emphasis to text pairs that involve frequently occurring utterances. This approach offers a dual advantage: it encourages the dispersion of clustering vectors, resulting in a more isotropic embedding space, while also increasing the informativeness of embeddings by increasing the separation between distinct dialogues. In unsupervised learning datasets, the presence of false negatives is inevitable. However, since we assign relatively low weights to false negatives, their impact on the optimization process is mitigated.

The formulation of the weighted contrastive

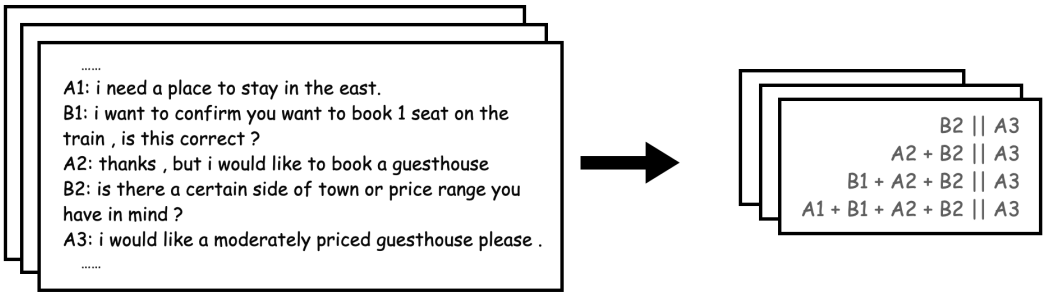

Figure 2: An illustration of MWPP. MWPP constructs multiple positive samples for the same response A3 by concatenating different window-sized consecutive turns adjacent to it. '+' means the concatenation of utterances.

learning object can be found in Appendix A.

4.2.2 Positive Pair Construction

We consider consecutive dialogue turns as context and define the positive pair as the context and its immediately following response. (See Figure 2)

In the initial step, we define a set of window sizes denoted as $W = \{1, 2, 3, \ldots\}$, and we select the window size iteratively. Following this, we pick one utterance from the dialogue as the response and establish a positive context by considering the last window-sized turns leading up to that response. All other combinations are categorized as negatives. This approach to generating contrastive pairs is named MWPP, an abbreviation for Multiple Window-sized Positive Pairs.

Directly contrasting encodings of the same response with those of different window-sized contexts can be problematic since these contexts differ in information richness. We must retain the inherent semantic distinctions among contexts from various turns while preserving their inferential similarity. Drawing inspiration from Wang and Li (2022), we incorporate linear layers to map different window-sized text pairs onto distinct embedding spaces. This enables flexible semantic matching during the training phase, enhancing our ability to capture semantics at various levels of granularity within the dialogue data. It is important to emphasize that projection layers are omitted during the evaluation phase.

5 Experiments

We initialize our training checkpoint with the pretrained BERT$_{base}$ and RoBERTa$_{base}$ models. To derive dialogue representations, we compute the average of the input token encodings from the final layer of the transformer encoder. During the training phase, we introduce multiple contrastive heads at the upper part of the encoder to enable contrastive learning at various levels of semantic granularity. Training details can be seen in D

5.1 Baselines

We compare WMDC against several text representation models, including BERT and RoBERTa, which serve as widely adopted baselines for language understanding tasks. These models are pretrained on extensive general text corpora. Additionally, we compare WMDC to SimCSE-unsup (Gao et al., 2021), which employs contrastive learning to acquire representations. It utilizes dropout as a minimal data augmentation strategy to construct positive pairs and in-batch negatives. Similarly, SimCSE-sup (Gao et al., 2021) leverages entailments as positives and contradictions as hard negatives. For dialogue understanding, TOD-BERT (Wu et al., 2020) employs contrastive learning by considering a random response and the concatenation of all its history as positive pairs. In contrast, DSE (Zhou et al., 2022b) defines positive pairs as two consecutive utterances within the same dialogue and treats all other pairs as negatives.

5.2 Results and Analysis

Action Prediction Table 2 shows the results for the action prediction task. Notably, WMDC consistently outperforms all baselines on both datasets by a substantial margin. WMDC surpasses the strongest baselines by 10.3 points on micro F1 and 6.3 points on macro F1, highlighting its superiority in capturing the overall meaning and anticipating future information within dialogue contexts. However, it's worth noting that despite these promising results, the overall performance of the action prediction task across all models remains relatively low, underscoring the task's difficulty and the potential for further enhancements.

Model	DSTC2		GSIM		AVG.	
	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
BERT	53.1	10.6	49.1	22.4	51.1	16.5
SimCSE-BERT$_{unsup}$	52.2	10.5	46.4	22.4	49.4	16.4
SimCSE-BERT$_{sup}$	52.9	10.9	47.2	21.7	50.1	19.0
TOD-JNT-BERT	52.9	12.7	55.1	30.0	54.0	21.3
DSE-BERT	52.0	12.6	50.4	25.7	51.2	19.1
WMDC-BERT	**56.9**	**16.8**	**66.1**	**38.3**	**61.5**	**27.6**
RoBERTa	48.7	6.5	37.0	10.5	42.9	8.5
SimCSE-RoBERTa$_{unsup}$	49.6	7.0	37.2	11.7	43.4	9.3
SimCSE-RoBERTa$_{sup}$	49.9	6.7	37.6	11.8	43.7	9.2
DSE-RoBERTa	48.0	6.8	39.6	12.3	43.8	9.6
WMDC-RoBERTa	**50.0**	**7.3**	**39.8**	**13.0**	**44.9**	**10.2**

Table 2: Results on action prediction task.

Model	IC-TOD	DECODE	AVG.
BERT	69.8	70.4	70.1
SimCSE-BERT$_{unsup}$	71.1	66.3	68.7
SimCSE-BERT$_{sup}$	71.4	69.1	70.3
TOD-JNT-BERT	70.8	69.7	70.3
DSE-BERT	70.8	69.6	70.2
WMDC-BERT	**73.3**	**70.7**	**72.0**
RoBERTa	70.1	76.0	73.1
SimCSE-RoBERTa$_{unsup}$	65.7	72.5	69.1
SimCSE-RoBERTa$_{sup}$	67.3	71.4	69.4
DSE-RoBERTa	73.3	72.5	72.9
WMDC-RoBERTa	**74.8**	**80.7**	**77.8**

Table 3: Results on dialogue inference task.

Model	TOP@1	TOP@3	TOP@5	TOP@10
BERT	8.2	14.8	18.8	26.7
SimCSE-BERT$_{unsup}$	15.3	23.9	28.8	36.6
SimCSE-BERT$_{sup}$	14.7	23.7	29.3	37.7
TOD-JNT-BERT	7.2	13.8	18.3	26.4
DSE-BERT	16.7	25.9	30.4	38.5
WMDC-BERT	**17.3**	**27.3**	**33.3**	**42.4**
RoBERTa	5.7	11.8	15.9	23.6
SimCSE-RoBERTa$_{unsup}$	16.7	26.6	32.5	41.9
SimCSE-RoBERTa$_{sup}$	14.5	24.0	29.4	38.7
DSE-RoBERTa	18.3	27.1	32.2	40.9
WMDC-RoBERTa	**19.6**	**29.7**	**36.0**	**46.2**

Table 4: Results on response selection task.

Dialogue Inference Table 3 presents the results for the dialogue inference task. Our model achieves state-of-the-art results on both datasets. Surprisingly, SimCSE-RoBERTa$_{sup}$, despite being trained on extensive NLI data, exhibits relatively lower performance in this task. This observation not only indicates a large disparity in patterns between plain texts and dialogues but also highlights WMDC's capability to comprehend the intrinsic conversational semantics and capture the nuances within dialogues.

Response Selection Table 4 showcases the results for response selection task. Our method outperforms all baselines across all evaluation metrics by a large margin. Notably, the performance gap widens as the value of N in TOP@N increases. We observe that even when contrasting plain texts, there is a significant enhancement in performance at the dialogue level, a contrast to the findings in the dialogue inference task.

5.3 Ablation Study

In this subsection, we analyze the influence of MWPP and sample reweighting on the performance of action prediction, dialogue inference, and response retrieval tasks. We evaluate the micro F1, accuracy, and TOP@3 performance for each task. However, we exclude the TOP@1 evaluation metric for the response retrieval task due to the presence of multiple false negatives in the candidate pool, which renders this metric unreliable.

Sample Reweighting. As illustrated in figure 3, it is evident that irrespective of the context window size, the exclusion of sample reweighting from the contrastive learning objective leads to a performance decline in all tasks. This observation underscores the importance of mitigating frequency imbalance issues and emphasizes the effectiveness of our sample reweighting strategy in improving the quality of dialogue representation.

MWPP. The arrow line in figure 3 reveals evident positive trends in performance among all tasks with increasing context window size. This phenomenon can be attributed to the fact that as the window size increases, more dialogue-level information is integrated. Notably, we observe that

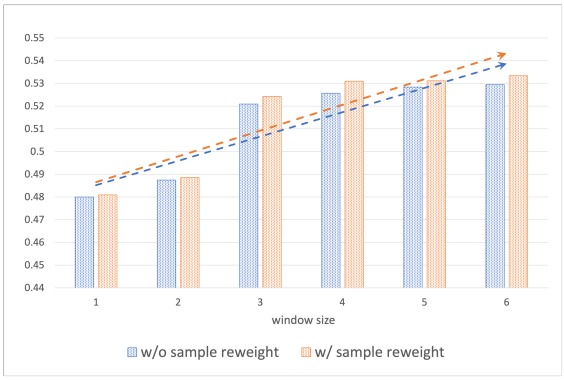

Figure 3: Ablation study on MWPP and sample reweighting. We present the averaged performance scores across all tasks, demonstrating that both methods consistently enhance performance. Notably, performance levels off as the window size increases.

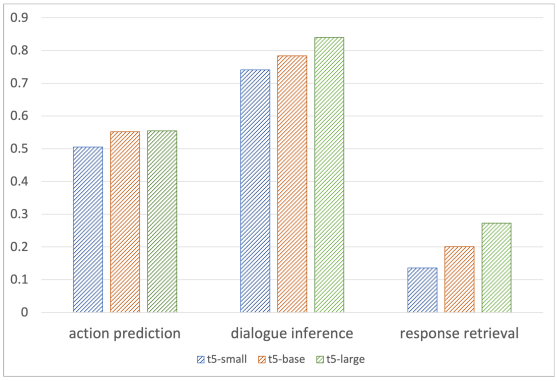

Figure 4: With the increase in model size, there is a consistent improvement in performance. This improvement shows no signs of saturation, suggesting the potential for further enhancements with even larger model sizes.

the performance experiences the most significant change when the window size transitions from 2 to 3, after which the rate of change becomes less pronounced. This suggests that if computational resources are constrained, a window size of 3 can be a viable and efficient choice.

5.4 Scalability

In this subsection, we study the scalability of our method regarding the size of the model and the training corpus.

Model Size. To scale up the model, we employ the pre-trained encoder of T5 series (Raffel et al., 2020), including three different sizes: T5-small, T5-base, and T5-large. We fine-tune these models using a comparatively smaller corpus obtained from TODBERT. From figure 4, the performance across all tasks consistently improves as models scale up. This improvement can be attributed to the increased capacity and generalization ability stemming from the greater number of parameters in the larger models.

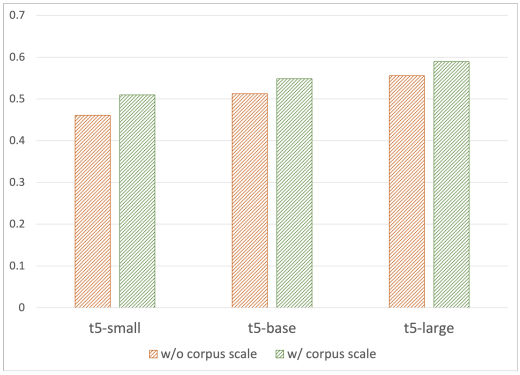

Figure 5: We present the average performance score across all tasks. As the corpus size expands, there is a continuous improvement in performance. This indicates that increasing the size of the corpus can result in enhanced performance across a range of tasks.

Corpus Size. We train the T5 series with a considerably larger corpus, as detailed in Appendix C. Figure 5 demonstrates that as the corpus size scales up, the performance on each task consistently improves, indicating that the model has not reached saturation yet. This enhancement can be attributed to the greater volume of dialogue information acquired during the training phase.

We believe that further gains can be achieved by scaling up both the model size and corpus size.

6 Conclusion

In this paper, we introduce DiaEval, a novel benchmark designed to assess dialogue representation models' ability to capture general dialogue semantic information. DiaEval consists of 5 datasets covering 3 distinct tasks, namely, action prediction, dialogue inference, and response selection. Furthermore, we have identified a frequency imbalance issue within existing dialogue corpora that can adversely affect the quality of dialogue representation. To address this issue and leverage the dialogue corpus more effectively, we propose WMDC, a Weighted Multi-window-sized Dialogue Contrastive learning method. It adjusts sample weights based on response frequency and contrasts the response with multiple adjacent window-sized contexts. Additionally, we conduct extensive experiments to demonstrate the effectiveness of our approach and to demonstrate its promising scalability on the extensive dialogue corpus we've collected.

Limitations

Further investigations are necessary to address the limitations associated with this approach. Firstly, WMDC, along with other universal dialogue representation methods, is data-thirsty. Besides a considerable carbon footprint, this poses a challenge in some languages where data may be scarce. Moreover, experiments in this paper solely employ encoder-only architecture. There is no warranty on its performance under other model architectures, such as the promising decoder-only GPT series.

As for future directions, we acknowledge that real-world conversations often involve multi-modal inputs, including audio and images, which are not currently included in our benchmark. Furthermore, although our method permits an unfixed number of turns as context, there remains a fixed hyperparameter for maximum input length. In actual conversations, however, dialogue can be much longer. Hence, further research is necessary to explore dialogue representation for extremely long, or even unlimited input lengths. Lastly, our proposed method is limited to English dialogue datasets. The effectiveness of our approach on dialogue datasets in other languages remains uncertain and warrants further investigation.

Ethics Statement

All the datasets utilized in this paper are sourced from publicly available repositories. However, it is important to acknowledge that inherent biases may still exist due to the nature of the data being collected from the Internet. It is crucial to emphasize that this paper does not involve any data collection or release, thereby eliminating any privacy concerns. Additionally, it is worth noting that our model has been trained on GPU, which may have environmental implications. Furthermore, this research does not involve any form of experimentation involving human subjects.

References

Xuefeng Bai, Yulong Chen, Linfeng Song, and Yue Zhang. 2021. Semantic representation for dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4430–4445, Online. Association for Computational Linguistics.

Xuefeng Bai, Linfeng Song, and Yue Zhang. 2022. Semantic-based pre-training for dialogue understanding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 592–607, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-AMR: Abstract Meaning Representation for dialogue. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 684–695, Marseille, France. European Language Resources Association.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. 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 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the*

2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3002–3017, Online. Association for Computational Linguistics.

Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. 2022. Contrastnet: A contrastive learning framework for few-shot text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10492–10500.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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, Copenhagen, Denmark. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljacic. 2022. Equivariant self-supervised learning: Encouraging equivariance in representations. In *International Conference on Learning Representations*.

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. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany. Association for Computational Linguistics.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Joachim Fainberg, Ben Krause, Mihai Dobre, Marco Damonte, Emmanuel Kahembwe, Daniel Duma, Bonnie Webber, and Federico Fancellu. 2018. Talking to myself: self-dialogues as data for conversational agents.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. 2019. Amazonqa: A review-based question answering task. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4996–5002. International Joint Conferences on Artificial Intelligence Organization.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

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, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Yucheng Huang, Kai He, Yige Wang, Xianli Zhang, Tieliang Gong, Rui Mao, and Chen Li. 2022. COPNER: Contrastive learning with prompt guiding for few-shot named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2515–2527, Gyeongju, Republic

of Korea. International Committee on Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

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.

Che Liu, Rui Wang, Junfeng Jiang, Yongbin Li, and Fei Huang. 2022. Dial2vec: Self-guided contrastive learning of unsupervised dialogue embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7272–7282, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021a. DialogueCSE: Dialogue-based contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2396–2406, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. 2023. Rankcse: Unsupervised sentence representations learning via learning to rank.

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 Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

William Myers, Tyler Etchart, and Nancy Fulda. 2020. Conversational scaffolding: An analogy-based approach to response prioritization in open-domain dialogs. In *ICAART (2)*, pages 69–78.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher 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, Doha, Qatar. Association for Computational Linguistics.

Libo Qin, Tianbao Xie, Shijue Huang, Qiguang Chen, Xiao Xu, and Wanxiang Che. 2021. Don't be contradicted with anything! CI-ToD: Towards benchmarking consistency for task-oriented dialogue system. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2367, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 353–360, Stockholm, Sweden. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Souvika Sarkar, Dongji Feng, and Shubhra Kanti Karmaker Santu. 2022. Exploring universal sentence

encoders for zero-shot text classification. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 135–147, Online only. Association for Computational Linguistics.

Hannes Schulz, Adam Atkinson, Mahmoud Adada, Kaheer Suleman, and Shikhar Sharma. 2019. Metalwoz: A dataset of multi-domain dialogues for the fast adaptation of conversation models. *Microsoft Research*.

Yeon Seonwoo, Guoyin Wang, Changmin Seo, Sajal Choudhary, Jiwei Li, Xiang Li, Puyang Xu, Sunghyun Park, and Alice Oh. 2023. Ranking-enhanced unsupervised sentence representation learning.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018a. 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, New Orleans - Louisiana. Association for Computational Linguistics.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018b. Building a conversational agent overnight with dialogue self-play.

Leslie N. Smith and Nicholay Topin. 2018. Super-convergence: Very fast training of neural networks using large learning rates.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Bin Wang and Haizhou Li. 2022. Relational sentence embedding for flexible semantic matching.

Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao. 2022. Sncse: Contrastive learning for unsupervised sentence embedding with soft negative samples.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

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.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Yuanhao Xiong, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, and Inderjit Dhillon. 2022. Extreme Zero-Shot learning for extreme text classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5455–5468, Seattle, United States. Association for Computational Linguistics.

Yi Xu and Hai Zhao. 2021. Dialogue-oriented pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2663–2673, Online. Association for Computational Linguistics.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Huaiyuan Ying, Shengxuan Luo, Tiantian Dang, and Sheng Yu. 2022. Label refinement via contrastive learning for distantly-supervised named entity recognition. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2656–2666, Seattle, United States. Association for Computational Linguistics.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022a. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. 2022b. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Yunhua Zhou, Peiju Liu, and Xipeng Qiu. 2022a. KNN-contrastive learning for out-of-domain intent classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5129–5141, Dublin, Ireland. Association for Computational Linguistics.

Zhihan Zhou, Dejiao Zhang, Wei Xiao, Nicholas Dingwall, Xiaofei Ma, Andrew Arnold, and Bing Xiang. 2022b. Learning dialogue representations from consecutive utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 754–768, Seattle, United States. Association for Computational Linguistics.

A Weighted Contrastive object

Adapting from NT-Xent loss, let $\mathbf{c}$ and $\mathbf{r}$ denote the embedding of context and response, respectively. The training objective for a single text pair i in a mini-batch of N pairs is given by:

$$\ell_i = -\frac{1}{2}\log\left(\frac{e^{\text{sim}(\mathbf{c}_i,\mathbf{r}_i)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{c}_i,\mathbf{r}_j)/\tau}}\right)$$

$$-\frac{1}{2}\log\left(\frac{e^{\text{sim}(\mathbf{r}_i,\mathbf{c}_i)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{r}_i,\mathbf{c}_j)/\tau}}\right)$$

where τ is the temperature hyper-parameter and sim is cosine similarity.

To address the issue of frequency imbalance, we extend this objective function with sample reweighting. The weight assigned to each pair is determined by an inverse function of its response frequency. By assigning a lower weight, utterances positive to the frequent ones will receive less optimization strength, allowing them to remain relatively distant.

Let $freq_{\mathbf{r}}$ denote the frequency of response $\mathbf{r}$ in the training data. The inverse response frequency weight IRF for text pair i can be calculated as:

$$IRF_i = \frac{1}{\log freq_{\mathbf{r}_i} + 1}, \tag{4}$$

The weighted contrastive loss function can then be defined as:

$$IRF_i * \ell_i \tag{5}$$

B Most common utterances

We present the top ten most common utterances in Table 5. A majority of these utterances offer limited information and are versatile in various dialogue contexts.

Utterances	Frequecy
yes.	5371
yes	4543
thank you.	3680
thanks	2971
okay.	2629
thank you	2347
thanks.	1910
thanks!	1678
anything else?	1594
what date and time would you like to go?	1481

Table 5: Most common utterances in TODBERT training corpus. All utterances are lowercase.

C Training Corpus

We utilize the same corpus as TODBERT to ensure a valid comparison. This dataset is a composition of nine sub-datasets, including 1) MetaLWOZ (Schulz et al., 2019), 2) Schema (Rastogi et al., 2020), 3) Taskmaster (Byrne et al., 2019), 4) MWOZ (Budzianowski et al., 2018), 5) MSR-E2E (Li et al., 2018), 6) SMD (Eric et al., 2017), 7) Frames (El Asri et al., 2017), 8) WOZ (Mrkšić et al., 2017) and 9) CamRest676 (Wen et al., 2017). See Table 6 for detailed information.

To assess the scalability of our proposed method, we have assembled a larger corpus. This dataset encompasses approximately 20 sub-datasets, including 1) Reddit (Zhang et al., 2020), 2) AmazonQA (Gupta et al., 2019), 3) Movie-Dialogs (Danescu-Niculescu-Mizil and Lee, 2011), 4) MetaLWOZ (Schulz et al., 2019), 5) Self-Dialog (Fainberg et al.,

Datasets	# Dialogue	# Utterance	Avg. Turn	Domain
MetaLWOZ (Schulz et al., 2019)	37,884	432,036	11.4	47
Schema (Rastogi et al., 2020)	22,825	463,284	20.3	17
Taskmaster (Byrne et al., 2019)	13,215	303,066	22.9	6
MWOZ (Budzianowski et al., 2018)	10,420	71,410	6.9	7
MSR-E2E (Li et al., 2018)	10,087	74,686	7.4	3
SMD (Eric et al., 2017)	3,031	15,928	5.3	3
Frames (El Asri et al., 2017)	1,369	19,986	14.6	3
WOZ (Mrkšić et al., 2017)	1,200	5,012	4.2	1
CamRest676 (Wen et al., 2017)	676	2,744	4.1	1
TOTAL	100,707	1,388,152	13.8	60

Table 6: Data statistics of the training corpus. We keep the original table from (Wu et al., 2020) and only add the last line.

2018), 6) TaskMaster1 (Byrne et al., 2019), 7) TaskMaster2 (Byrne et al., 2019), 8) TaskMaster3 (Byrne et al., 2019), 9) Schema (Rastogi et al., 2020), 10) PersonaChat (Zhang et al., 2018), 11) MWOZ (Budzianowski et al., 2018), 12) MSR-E2E (Li et al., 2018), 13) TopicChat (Gopalakrishnan et al., 2019), 14) ABCD (Chen et al., 2021), 15) ChitChat (Myers et al., 2020), 16) SMD (Eric et al., 2017), 17) Frames (El Asri et al., 2017), 18) WOZ (Mrkšić et al., 2017), 19) CCPE-M (Radlinski et al., 2019), and 20) CamRest676 (Wen et al., 2017). See Table 7 for data statistic information.

D Hyper-parameters

Each head is a linear layer with a size of (d * d), where d is the hidden size of the model. We set the batch size to 256, and use the AdamW optimizer (Kingma and Ba, 2017) along with the OneCycleLR learning rate scheduler (Smith and Topin, 2018). The learning rate for the encoder is set to 3e-5, while the learning rate for the contrastive heads is amplified by a factor of 40. We set the contrastive temperature τ to 0.05.

Datasets	# Dialogue	# Utterance	Avg. Turn
Reddit (Zhang et al., 2020)	15,914,021	31,908,317	2.0
AmazonQA (Gupta et al., 2019)	962,260	1,924,520	2.0
Movie-Dialogs (Danescu-Niculescu-Mizil and Lee, 2011)	220,579	441,158	2.0
MetaLWOZ (Schulz et al., 2019)	37,884	356,268	9.4
Self-Dialog (Fainberg et al., 2018)	24,165	348,554	14.4
TaskMaster1 (Byrne et al., 2019)	13,215	135,176	10
TaskMaster2 (Byrne et al., 2019)	17,289	137,064	7.9
TaskMaster3 (Byrne et al., 2019)	23,789	237,617	10.0
Schema (Rastogi et al., 2020)	22,825	463,284	20.3
PersonaChat (Zhang et al., 2018)	18,876	250,634	13.3
MWOZ (Budzianowski et al., 2018)	10,420	71,410	6.9
MSR-E2E (Li et al., 2018)	10,087	74,686	7.4
TopicChat (Gopalakrishnan et al., 2019)	10,784	235,434	21.8
ABCD (Chen et al., 2021)	8,034	64,500	8.0
ChitChat (Myers et al., 2020)	7,168	258,145	36
SMD (Eric et al., 2017)	3,031	15,928	5.3
Frames (El Asri et al., 2017)	1,369	19,986	14.6
WOZ (Mrkšić et al., 2017)	1,200	7,624	6.4
CCPE-M (Radlinski et al., 2019)	502	12,000	24.0
CamRest676 (Wen et al., 2017)	676	2,744	4.1
TOTAL	17,308,174	36,965,049	2.1

Table 7: Data statistics of the training corpus for scaling.