SPACE-2: Tree-Structured Semi-Supervised Contrastive Pre-training for Task-Oriented Dialog Understanding

Wanwei He1,2*, Yinpei Dai3, Binyuan Hui1,2, Min Yang1, Zheng Cao3, Jianbo Dong3, Fei Huang4, Luo Si3, Yongbin Li3*

1Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
2University of Chinese Academy of Sciences, China
3DAMO Academy, Alibaba Group

{ww.he, min.yang}@siat.ac.cn
{yinpei.dyp, f.huang, luo.si, shuide.lyb}@alibaba-inc.com

Abstract

Pre-training methods with contrastive learning objectives have shown remarkable success in dialog understanding tasks. However, current contrastive learning solely considers the self-augmented dialog samples as positive samples and treats all other dialog samples as negative ones, which enforces dissimilar representations even for dialogs that are semantically related. In this paper, we propose SPACE-2, a tree-structured pre-trained conversation model, which learns dialog representations from limited labeled dialogs and large-scale unlabeled dialog corpora via semi-supervised contrastive pre-training. Concretely, we first define a general semantic tree structure (STS) to unify the inconsistent annotation schema across different dialog datasets, so that the rich structural information stored in all labeled data can be exploited. Then we propose a novel multi-view score function to increase the relevance of all possible dialogs that share similar STSs and only push away other completely different dialogs during supervised contrastive pre-training. To fully exploit unlabeled dialogs, a basic self-supervised contrastive loss is also added to refine the learned representations. Experiments show that our method can achieve new state-of-the-art results on the DialoGLUE benchmark consisting of seven datasets and four popular dialog understanding tasks. For reproducibility, we release the code and data at https://github.com/AlibabaResearch/DAMO-ConvAI/tree/main/space-2.

1 Introduction

Task-oriented dialog (TOD) systems aim to help users to accomplish specific tasks through natural language interactions, e.g., restaurant booking, hotel reserving, and movie searching (Young et al., 2013). In order to fulfill a user goal, the dialog systems must be capable of extracting structured semantics from the dialog utterances, which is often referred to as dialog understanding. Common tasks of dialog understanding include: 1) intent prediction, where the system classifies the user intention given the current utterance, 2) slot filling, where the system extracts the value spans for pre-defined slots from user utterances, and 3) dialog state tracking, where slot-value pairs are predicted and updated according to multi-turn dialog history.

Recent methods address the dialog understanding problems mainly by task-adaptive training with pre-trained language models (PLMs) (Henderson et al., 2020; Mehri et al., 2020; Zhang et al., 2022). By fine-tuning the entire parameters of PLMs on dialog corpora, the versatile knowledge acquired from large-scale corpora can be adapted to handle various dialog understanding tasks. For example, TOD-BERT (Wu et al., 2020) continually pre-trained a BERT on large-scale task-oriented dialog corpora with similar objectives to grasp more knowledge in conversations and achieved good results on a wide range of tasks. To further enhance the representations of dialogs, contrastive learning is widely used for dialog pre-training via pulling semantically similar sentences together and pushing apart dis-similar ones, which has shown strong ability in few-shot dialog understanding (Zhang et al., 2020, 2021a).

Although contrastive learning (CL) is effective in pre-training dialog models, it still suffers from several problems. First, the current practice of CL often neglects the semantic structures of dialog examples and only utilizes shallow self-supervised information. Figure 1 illustrates the semantic frames of two dialog examples, which indicate the hierarchical structures with semantic meanings in terms of intentions and slots. The semantic structure is often summarized as the turn-level structural annotation and is prominent in multi-domain task-oriented dialogs. Exploiting such structure during
task-adaptive pre-training has been proved to be effective in dialog understanding tasks (Yu et al., 2020). However, in conventional CL, only the examples with the exact same annotations are deemed as positive samples, while all other examples are considered as negative ones (Gao et al., 2021), ignoring the fact that different examples may share similar semantics to some extent as shown in Figure 1. Second, in previous dialog pre-training methods, no annotations have been exploited to learn better pre-trained representations. But in computer vision, there have already been many works demonstrating that labeled data can accelerate the pre-training procedure as well as improve the model performance (Assran et al., 2020; Khosla et al., 2020; Dai et al., 2021b). Therefore, we argue that it is crucial to combine both the labeled and unlabeled dialog data to learn more powerful pre-trained dialog models.

Built on these motivations, we propose a novel tree-structured pre-trained conversation model (SPACE-2), which fully exploits the structural information in labeled data to improve the dialog pre-training on large-scale unlabeled corpora via semi-supervised contrastive learning. To be specific, we first define a general semantic tree structure (STS), which is suitable for TOD and compatible with the discrepancy of annotation schema in varied dialog datasets. Then we propose a novel multi-view scoring function to measure the similarity among different labeled dialogs. Given two dialogs with STSs, the scoring function is able to calculate their relations hierarchically and aggregate the output multiple scores from either global or local aspects. In this way, all labeled data are viewed as positive samples with soft scores instead of hard scores (0 or 1) in traditional CL. Thus, more subtle semantic structures in training samples can be taken into account, and dialog representations can be learned in a semi-supervised manner.

Our contributions are summarized as follows:

- We explore tree-structured semi-supervised contrastive pre-training for task-oriented dialog understanding. To the best of our knowledge, this is the first study to inject structural information and exploit labeled semantic frames in a pre-trained conversation model.

- We propose a multi-view score function for similarity measurement among different labeled dialogs, which could calculate relations hierarchically and aggregate multiple judgements from either global or local aspects.

- We conduct extensive experiments on the DialogGLUE benchmark with seven different datasets. Empirical results show that our SPACE-2 consistently performs better than strong competitors on four dialog understanding tasks (i.e., intent prediction, slot filling, semantic parsing, and dialog state tracking).

2 Related Work

2.1 Pre-trained Models for Dialog Understanding

Recent advances in pre-trained language models (PLMs) have spurred success in natural language understanding in task-oriented dialog (TOD) systems. Since general-purpose PLMs such as BERT (Devlin et al., 2019) are not able to capture dialog-oriented knowledge in TOD, such as explicit user intent, many studies proposed to perform pre-training on dialog corpora. For example, ConvRT (Henderson et al., 2020) explored the pre-trained language transformers for TOD via learning a dialog response selection loss. ConvBERT (Mehri et al., 2020) fine-tuned BERT on a large open-domain dialog corpus using a masked language modeling objective so that more semantically meaningful representations of dialogs can be learned. TOD-BERT (Wu et al., 2020) incorporated role tokens in the input and employed a contrastive objective for response selection. Zhang et al. (2021b) proposed to use masked utterance modeling and response contrastive loss to produce DialogueBERT and showed good performance on intent, emotion, and entity recognition. Apart from training with common objectives, some work (Yu et al., 2020; He et al., 2020b; Hui et al., 2021; Dai et al., 2021a) also
proposed to learn relations between schema and keywords of utterances during pre-training, which benefits in certain downstream tasks that require parsing semantics. The work of (Sun et al., 2019; Liu et al., 2020a) tried to integrate structural information of knowledge graphs to enhance language representations. Unlike these methods, we are the first to inject the semantic tree structure into pre-training models for dialog understanding.

### 2.2 Contrastive Learning

Contrastive learning (CL) is one of the pre-training methods that leverage large-scale unlabeled data to learn meaningful sentence representations. The key idea of CL is to narrow the distance between two semantically similar sentence representations, meanwhile, push away the representations of dissimilar sentences (Gao et al., 2021; Wu et al., 2021; Yan et al., 2021). There are also several studies focused on applying CL to dialog understanding tasks. For example, Zhang et al. (2021a) used CL as self-supervised pre-training and adapted the pretrained models on few-shot intent prediction. The studies (Mehri and Eric, 2021; Vulić et al., 2021) utilized supervised CL at the fine-tuning stages for intent prediction, where the samples from the same class are all regarded as positives. He et al. (2020a) proposed to use adversarial samples as difficult negative instances in CL for zero-shot cross-domain slot-filling. Different from previous work, we use CL in both pre-training and fine-tuning stages. In addition, we train dialog models with both labeled and unlabeled data in a semi-supervised manner.

### 2.3 Semi-Supervised Contrastive Learning

Semi-supervised contrastive learning has proved to be an effective pre-training method in many research field, such as image recognition (Yuan et al., 2021; Park et al., 2021; Li et al., 2021; Kim et al., 2021a), image segmentation (Alonso et al., 2021; Zhou et al., 2021) and speech recognition (Xiao et al., 2021; Inoue and Goto, 2020). It can combine both labeled and unlabeled data together to train effective neural models for downstream tasks. He et al. (2022b) proposed to use consistency regularization loss to learn dialog policy from labeled and unlabeled dialog corpora via a semi-supervised manner. However, leveraging such a semi-supervised contrastive learning paradigm in task-oriented dialog understanding remains unexplored. In this paper, we explore semantic structures in TOD, and propose a new contrastive learning approach that calculates semantic tree-structure similarity among all possible labeled dialog data.

### 3 Pre-training Data

In this section, we first describe the dialog datasets (He et al., 2022a) used for pre-training, including a labeled dialog dataset (AnPreDial) and a large-scale unlabeled dialog corpus (UnPreDial). Then, we elaborate on the semantic tree structure (STS) for the task-oriented dialog data.

#### 3.1 Dataset Description

To provide sufficient high-quality dialog corpora to train our pre-trained conversation model, we use a labeled dialog dataset called AnPreDial, which contains 32 existing labeled TOD datasets, ranging from single-turn question answering to multi-turn dialogs. We also use a large-scale unlabeled dialog corpus called UnPreDial with 21 dialog datasets, ranging from open-domain online forums to document-grounded dialogs. The statistics of our pre-trained dialog corpora are shown in Table 1 and more details can be found in Appendix A.

#### 3.2 Semantic Tree Structure

Task-oriented dialogs often contain several user goals, and the system needs to help fulfill these goals through interacting with users. To this end, the user utterances and system responses are usually composed of rich semantics such as intents and slots. In different datasets, the annotation taxonomy used to describe the semantic scopes varies. For example, in some intent recognition datasets such as BANKING77 (Casanueva et al., 2020) and CLINIC150 (Larson et al., 2019), they only possess a pre-defined intent set and formulate the task as a classification problem. In other datasets like Taskmaster (Byrne et al., 2019a), only the slot labels are provided; for instance, given a user utterance “I need a table for tonight at 7 pm for 8 people”, the label is “time.reservation=7 pm for 8 people”.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>AnPreDial</th>
<th>UnPreDial</th>
</tr>
</thead>
<tbody>
<tr>
<td># Datasets</td>
<td>32</td>
<td>21</td>
</tr>
<tr>
<td># Dialogs</td>
<td>459,465</td>
<td>3,217,058</td>
</tr>
<tr>
<td># Turns</td>
<td>3,366,479</td>
<td>19,578,028</td>
</tr>
<tr>
<td>Avg. tokens per turn</td>
<td>13.9</td>
<td>14.5</td>
</tr>
<tr>
<td>Avg. tokens per dialog</td>
<td>101.8</td>
<td>88.2</td>
</tr>
<tr>
<td>Total unique tokens</td>
<td>46.8M</td>
<td>283.7M</td>
</tr>
</tbody>
</table>

Table 1: Statistics of our pre-trained dialog dataset.
As illustrated in Figure 3, we build our SPACE-2 model based on the bidirectional Transformer architecture (Vaswani et al., 2017). Different from the vanilla input representations as in BERT (Devlin et al., 2019), we set our input embeddings consisting of four elements: tokens, roles, turns, and positions. Role embeddings are used to segment

which role the current token belongs to either user or system. Turn embeddings are assigned to each token according to its turn number in the dialog. Position embeddings are assigned to each token according to its relative position within its belonging sentence. Appendix B shows more details.

4.2 Tree-Structured Semi-Supervised
Contrastive Learning Paradigm

We aim to leverage semi-supervised pre-training to learn better pre-trained representations from both the labeled and unlabeled data. Concretely, we adopt a tree-structured supervised contrastive objective on the labeled dataset AnPreDial, while a tree-structured self-supervised contrastive objective on the unlabeled dataset UnPreDial.

4.2.1 Tree-Structured Supervised CL

For labeled data, we treat every dialog sample as a positive sample assigned with similarity scores to each other. Between two semantic tree structures (STSs), we define a score function to calculate similarity scores from different views. Let $D, I, S, V$ denote the sets of node in domain, intent, slot and value layer. For the sample from MULTIWOZ dataset in Figure 2, its $D = \{\text{restaurant}\}$, $I = \{\text{inform}, \text{name}\}$, $S = \{\text{food area}\}$, and $V = \{\text{indian south}\}$.
where we use \(s_k\) and \(S_{all}\) for two different STSs: \(s_k \in S_{all}^i, s_k \in S_{all}^j\). Is the function for the Jaccard similarity coefficient, which takes the ratio of Intersection over Union.

Given the dialog context, our model output a pooled representation at the [CLS] position as the sentence embedding \(z\) of the whole context. Thus for any sample pair \(i, j\), the output sentence embeddings are denoted as \(z_i, z_j\). We use \(f_{k, i,j}\) to denote the \(k\)-th similarity score \(f_k(T_i, T_j)\) for simplicity. Suppose the size of the current batch is \(N\), we follow the common practice in Gao et al. (2021) to duplicate the data via dropout-based data augmentation. Then we acquire a \(2N\)-length new batch. Let \(I = \{1 \ldots 2N\}\) be the index set of the new batch, and \(C(i) \equiv I / \{i\}\). Since we compute multiple scores \(f_{k, i,j}\) for each STS pair, different types of scoring functions can be used to construct the CL loss. In this paper, we propose both single-view and multi-view scoring functions to build the tree-structured supervised CL objectives as follows and find that the multi-view method performs better.

**Single-View Scoring Objective.** Similar to the common practice in the current CL, we simply average the \(K\) scores into a single value to weigh the semantic similarity among samples:

\[
L_{sup}^{single} = - \sum_{i \in I} \sum_{j \in C(i)} \left( \frac{1}{K} \sum_{k=1}^{K} f_{k, i,j} \right) \times \log \frac{\exp (\sigma(z_i) \cdot \sigma(z_j) / \tau)}{\sum_{l \in C(i)} \exp (\sigma(z_i) \cdot \sigma(z_l) / \tau)}
\]

where \(\sigma(\cdot)\) is a normalized linear mapping to transform \(z\) to a unit vector. \(W\) is a learnable square weight matrix and \(b\) is a bias vector. \(\sigma(\cdot)\) is the normalization operation and \(\tau \in [0, 1]\) is a temperature hyper-parameter.

**Multi-View Scoring Objective.** Figure 4 demonstrates our multi-view scoring objective. Instead of summing up as one score, we can separately utilize multiple scores to better exploit sub-structures from different semantic spaces (Zhu et al., 2019) as the following loss:

\[
L_{sup}^{multi} = - \sum_{i \in I} \sum_{j \in C(i)} \sum_{k=1}^{K} \left( \frac{f_{k, i,j}}{\sum_{m \in C(i)} f_{k, i,m}} \right) \times \log \frac{\exp (\sigma_k(z_i) \cdot \sigma_k(z_j) / \tau)}{\sum_{l \in C(i)} \exp (\sigma_k(z_i) \cdot \sigma_k(z_l) / \tau)}
\]

\[
\sigma_k(z) = \text{Norm}(W_k z + b_k)
\]

where \(\sigma_k(\cdot)\) denotes linear mappings with different parameters. As illustrated in Figure 5, our methods are quite distinct from vanilla supervised CL (Vulić et al., 2021), where only augmented data or data from the same label are considered as positive samples \((f_{i,j} = 1)\), and all other data are considered as negative samples \((f_{i,j} = 0)\). Therefore, our method can be viewed as a generalized case where all labeled data are regarded as positive samples assigned with soft scores.

4.2.2 Tree-Structured Semi-Supervised CL

For unlabeled data, since there are no available labels, we adopt a tree-structured self-supervised contrastive objective in a similar way in Gao et al. (2021), where only the augmented data by dropout is deemed as a positive sample. In particular, under the multi-view setting, the objective \(L_{sup}^{multi}\) still shares the same \(K\) sub-spaces as \(L_{sup}^{multi}\) as follows:

\[
L_{sup}^{single} = - \sum_{i \in I} \sum_{j \in C(i)} \left( 1(j = i^+) \times \log \frac{\exp (\sigma(z_i) \cdot \sigma(z_j) / \tau)}{\sum_{l \in C(i)} \exp (\sigma(z_i) \cdot \sigma(z_l) / \tau)} \right)
\]

\[
L_{sup}^{multi} = - \sum_{i \in I} \sum_{j \in C(i)} \sum_{k=1}^{K} \left( 1(j = i^+) \times \log \frac{\exp (\sigma_k(z_i) \cdot \sigma_k(z_j) / \tau)}{\sum_{l \in C(i)} \exp (\sigma_k(z_i) \cdot \sigma_k(z_l) / \tau)} \right)
\]
where $1(j = i^+)$ means it only takes 1 when the data sample $j$ is the augmented data $i^+$ of data sample $i$, otherwise 0.

In order to maintain the language modeling ability and extract the slot values better, we also apply the spanMLM loss $L_{slm}$ (Joshi et al., 2020) to the output representations on top of the dialog context. Therefore, our final loss is computed as:

$$L^*_{total} = L^*_{sup} + L^*_{self} + L_{slm}$$

where * can be either single or multi.

5 Experimental Setup

5.1 Task Formulation

DialogGLUE benchmark consists of seven different datasets, including four tasks as follows:

**Intent Prediction.** This is a classification task for models to predict which intent class the user conveys in the utterance. There are three datasets accordingly: 1) BANKING77 (Casanueva et al., 2020), a single turn dataset in the banking domain with 77 intent labels and 13k utterances; 2) CLINIC150 (Larson et al., 2019), a multi-domain dataset that contains 23.7k utterances, 150 intents, 10 domains; and 3) HWU64 (Liu et al., 2021) includes 25.7k utterances for 64 intents spanning 21 domains. To conduct the downstream task, a linear layer neural network is added to the pooled output $z$ to classify. The evaluation metric is accuracy.

**Slot Filling.** This is a sequential labeling task in which a model tries to predict the IOB tags upon the sequence of dialog context. There are two datasets: 1) REST8K (Coope et al., 2020), which contains 8.2k utterances and 5 slots (date, time, people, first name, last name) from a restaurant domain; 2) a transformed DSTC8 (Rastogi et al., 2020) dataset comprising of 5.5k utterances and slot annotations in 4 domains (buses, events, homes, rental cars). Both datasets are single-turn formed and every token in the utterance is predicted as either being the beginning of a slot value (B-), inside a slot value (I-) or not belonging to a slot value (O). A linear layer neural network is added to the representation of the final output layer to predict correct tags (e.g., “B-time”, “I-people”). The evaluation metric is the macro-averaged F1 score.

**Semantic Parsing.** The original TOP (Gupta et al., 2018) dataset targeting at hierarchical parsing is transformed into a joint task of intent prediction and slot-filling, which contains 44k utterances. The metric is the exact-match as in Mehri et al. (2020).

**Dialog State Tracking.** This task is multi-turn slot-filling, where the model interprets the user meanings during the course of a dialog to maintain a dialog state in terms of slot-value pairs. MULTiWOZ 2.1 (Eric et al., 2020) is the chosen dataset version with 10k dialogs. Following Mehri et al. (2020), we report joint goal accuracy and use TripPy (Heck et al., 2020) as the downstream model.

5.2 Baselines

Our compared baselines can be divided into two categories. The first type is the pre-trained model adapted for general dialog understanding tasks in DialogGLUE. We choose BERT (Devlin et al., 2019), ConvBERT (Mehri et al., 2020) and its variants, and TOD-BERT (Wu et al., 2020) as our baselines. The second type is the pre-trained model designed only for specific dialog tasks. In this paper, we choose Example+Observer (Mehri and Eric, 2021), DNNC (Zhang et al., 2020), CPFT (Zhang et al., 2021a); and ConvFit (Vulić et al., 2021) for the intent prediction task. We also choose SpanConveRT (Coope et al., 2020), ConVEx (Hender-
Table 2: Total results of seven datasets from the DialoGLUE benchmark. † denotes original results from DialoGLUE. adapt denotes re-implementation via adaptively pre-training on our corpora. ‘w/o’ denotes ablation study in Sec 6.4.

5.3 Settings

In our experiment, we initialize SPACE-2 with ConvBERT. The input length is bound to 256 and the batch size is 128. AdamW is used for optimization with an initial learning rate of 1e-5. The dropout rate is 0.2. For semi-supervised pre-training, we combine the labeled and unlabeled data to form a batch in a 1:1 ratio. For each few-shot experiment, we exclude the training data in the AnPreDial accordingly to avoid unfair data use. For instance, if the target few-shot task is HWU64, we only use the few-shot training data of HWU64 and all other datasets as pre-training resources, so no extra training data (dev&test) in HWU64 is used. For all downstream tasks, we average 5 seeds for final results.

6 Experimental Results

6.1 Overall Performance on DialoGLUE

We follow the original settings in Mehri et al. (2020) and obtain complete results for our model with single-view scoring objectives, denoted as SPACE-2\textsuperscript{single}, and with multi-view scoring objectives, denoted as SPACE-2\textsuperscript{multi}, on all seven datasets of DialoGLUE. As shown in Table 2, our SPACE-2\textsuperscript{multi} model achieves state-of-the-art performance across all tasks, improving the average score of previous best results by 1.39% (86.89% → 88.10%) and 5.66% (73.80% → 77.98%) on full-data training and 10-shot training settings, respectively. This indicates that our semi-supervised contrastive learning can acquire better pre-trained representations to adapt efficiently to downstream dialog tasks by utilizing both labeled and unlabeled data. Although the superior results of SPACE-2\textsuperscript{single} over other baselines, it is worse than SPACE-2\textsuperscript{multi}, suggesting that multi-view scoring is crucial to improve the overall performance. To eliminate the discrepancy brought by
the pre-training dialog corpora, we choose two similar models, TOD-BERT and ConvBERT+MLM, to continually pre-train on AnPreDial and UnPreDial for fairness, but they are still not comparable to SPACE-2.

6.2 More Comparison on Intent Prediction

The experimental results of three intent prediction datasets are shown in Table 3. As we can see, SPACE-2\textsubscript{multi} achieves SOTA results on all datasets under all settings. Specifically, it outperforms the previous best model ConvFiT by 0.93%, 2.33%, 3.53% on BANKING\textsubscript{77}, CLINIC\textsubscript{150}, HWU\textsubscript{64} under the few-data training setting, respectively. It also improves ConvFiT by 0.61%, 0.46%, and 1.91% on the above three datasets under the full-shot training setting. The improvements indicate that our method has a better ability to discriminate similar intents via tree-structure enhanced contrastive learning than vanilla pre-training on ConvFiT.

6.3 More Comparison on Slot-Filling

We also conduct slot-filling experiments, particularly in the few-shot setting, to probe whether our model is also suitable for slot extraction. We follow the same data split and settings in GenSF (Mehri and Eskenazi, 2021) for a fair comparison. As shown in Table 4, SPACE-2\textsubscript{multi} achieves state-of-the-art results across all experimental settings on DSTC8 and REST8K. Note that in Mehri and Eskenazi (2021), DSTC8 is evaluated in every single domain (Bus, Event, Home, RentalCar) and REST8K is evaluated with varied proportions (1/8, 1/16, 1/32) of the training set. Due to the limited space, we only show the sub-spaces of D, I, S, V here. As we can see, The hidden representations of SPACE-2\textsubscript{multi} are able to differentiate the similar and dissimilar parts in different semantic sub-spaces. The learned latent sub-space is highly correlated with the dialog annotations of domain, intent, slot or value, which confirms our assumption.

6.4 Ablation Study

To figure out which factor contributes mainly to our semi-supervised pre-training paradigm, we conduct ablation studies on DialoGLUE as shown in the bottom part of each setting in Table 2. We first investigate the effect of the tree-structured supervised contrastive learning by replacing it with the traditional supervised contrastive learning (denote as ‘w/o tree’). As we can see, vanilla supervised contrastive learning only considers self-augmented data or data with the exact same labels as positive samples, so it can not capture a more detailed semantic structure than our tree-structured method and degrade from 88.10% to 87.57% in an average score. We also try to remove the label information from AnPreDial and treat all pre-training data as unlabeled data (denote as ‘w/o annotation’), or remove the unlabeled data (denote as ‘w/o UnPreDial’) to see the impact. It shows that both the labeled data and unlabeled data are crucial in our semi-supervised pre-training scheme but labeled data contributes more to the overall performance.

6.5 Visualization Analysis

In this part, we investigate why explicitly decomposing the semantic structure into sub-structures for different views can benefit dialog understanding. Figure 6 illustrates the 2D t-SNE visualization of the output unit vectors $\sigma_k(z)$ for test dialog samples from the MULTIWOZ dataset. Due to the limited space, we only show the sub-spaces of D, I, S, V here. As we can see, The hidden representations of SPACE-2\textsubscript{multi} are able to differentiate the similar and dissimilar parts in different semantic sub-spaces. The learned latent sub-space is highly correlated with the dialog annotations of domain, intent, slot or value, which confirms our assumption.

7 Conclusion

In this work, we propose a new pre-trained conversation model named SPACE-2, which learns dialog representations from both labeled and unlabeled corpora via tree-structured semi-supervised contrastive learning (CL). To be specific, we utilize a semantic tree structure (STS) to unify the incon-
consistent annotation schema for all datasets. Then, we propose a novel multi-view scoring function to compute the semantic similarity based on STS's sub-structures for labeled dialogs. Different from traditional CL, our method increases the relevance of representations for all similarly labeled dialogs and only pushes away completely different dialogs. A self-supervised CL loss is also integrated to exploit unlabeled dialogs. Extensive experiments on the DialoGLUE benchmark with seven datasets demonstrate that our model achieves the best results on all dialog understanding tasks, under the full-data and the few-shot settings.

Acknowledgement

This work was partially supported by National Natural Science Foundation of China (No. 61906185), Youth Innovation Promotion Association of CAS China (No. 2020357), Shenzhen Science and Technology Innovation Program (No. KQTD20190929172835662), Shenzhen Basic Research Foundation (No. JCYJ20210324115614039 and No. JCYJ20201009113441941). This work was supported by Alibaba Group through Alibaba Research Intern Program.

References


Sai Zhang, Yuwei Hu, Yuchuan Wu, Jiawan Wu, Yongbin Li, Jian Sun, Caixia Yuan, and Xiaojie Wang. 2022. A slot is not built in one utterance: Spoken language dialogs with sub-slots. In *Findings of


A Details for Pre-training Dialog Corpora

All dialog corpora we used in our experiments are given in Table 5 and Table 6. There are totally 32 labeled datasets and 21 unlabeled datasets. For the labeled data, each dialog sample is unified into the same data format with tree-structured labels. For the unlabeled data, each sample is performed careful cleaning due to the high noise. The processing methods include: (1) removing the instances where there is a URL in utterances. (2) removing the instances containing word repetitions of at least three words. (3) removing non-English sentences. (4) removing sentences containing special markers such as “[” or “]”, as this could be markup. (5) removing offensive language. (6) replacing the non-unicode characters like emojis.

B Input Representations

The input embeddings consist of four elements: tokens, roles, turns, and positions. Role embeddings, $E_u$, $E_s$, are like segmentation embeddings in BERT and are used to differentiate which role the current token belongs to either user or system. Turn embeddings are assigned to each token according to its adverse turn number in the input dialog data. Position embeddings are assigned to each token according to its relative position within its belonging sentence. For the tokens, we use special tokens [BOU], [EOU] to bound the user utterance, and [BOS], [EOS] to bound the system response.

C Details for Compared Baselines

C.1 Baselines for Dialog Understanding

We choose the following baselines: 1) BERT, a BERT$_{base}$ model adapted to all downstream tasks; 2) ConvBERT, a BERT continually pre-trained on a large open-domain dialogue corpus with MLM (Mehri et al., 2020); 3) ConvBERT-DG, a ConvBERT continually pre-trained on all DialoGLUE data before conducting specific downstream tasks with MLM. 4) ConvBERT+MLM, a ConvBERT continually trained on training data of specific downstream tasks; 5) ConvBERT-DG+MLM, a ConvBERT+MLM pre-trained on all DialoGLUE data before conducting downstream tasks; and 6) TOD-BERT, a BERT pre-trained on nine existing task-oriented dialogue corpora using MLM and response selection loss (Wu et al., 2020).

C.2 Baselines for Specific Tasks

In this paper, we focus on intent prediction and slot-filling in particular since there are adequate pre-trained baselines specifically. For the intent prediction task, we choose: 1) Example+Observer, a ConvBERT+MLM that uses example-driven training based on similarity matching and observers for transformer attentions (Mehri and Eric, 2021); 2) USE+ConveRT, a dual sentence encoder pre-trained on 654M Reddit dialogs (Casanueva et al., 2020); 3) DNNC, a discriminative nearest-neighbor model which finds the best-matched sample via similarity matching (Zhang et al., 2020); 4) CPFT, which uses self-supervised contrastive learning to perform pre-training and fine-tuning (Zhang et al., 2021a); and 5) ConvFit, which uses regular supervised contrastive learning on labeled data after pre-training on chatting data (Vulić et al., 2021). For the slot-filling task, we choose: 1) Span-ConveRT, a slot-filling model pre-trained on collected data (Coope et al., 2020); 2) ConVEx, a span-ConveRT pre-trained with the pairwise cloze task on Reddit for few-shot slot labeling (Henderson and Vulić, 2021); and 3) GenSF, a generative pre-trained slot-filling model (Mehri and Eskenazi, 2021).
Table 5: Statistics for each labeled dataset in AnPreDial.
<table>
<thead>
<tr>
<th>Name</th>
<th># Dialog</th>
<th># Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>MulDoGo_un(Peskov et al., 2019)</td>
<td>63,404</td>
<td>1,013,985</td>
</tr>
<tr>
<td>ABCD(Chen et al., 2021)</td>
<td>10,042</td>
<td>143,855</td>
</tr>
<tr>
<td>AirDialog(Wei et al., 2018)</td>
<td>361,822</td>
<td>4,645,475</td>
</tr>
<tr>
<td>CCPE(Radlinski et al., 2019)</td>
<td>502</td>
<td>11,971</td>
</tr>
<tr>
<td>MetalWOZ(Shalyminov et al., 2020)</td>
<td>40,203</td>
<td>458,237</td>
</tr>
<tr>
<td>CMU_DoG(Zhou et al., 2018)</td>
<td>4,221</td>
<td>134,197</td>
</tr>
<tr>
<td>ConvQuestions(Kacupaj et al., 2021)</td>
<td>11,200</td>
<td>112,000</td>
</tr>
<tr>
<td>CoQA(Reddy et al., 2019)</td>
<td>7,699</td>
<td>233,260</td>
</tr>
<tr>
<td>CoSQL(Yu et al., 2019)</td>
<td>2,458</td>
<td>25,433</td>
</tr>
<tr>
<td>doc2dial(Feng et al., 2020)</td>
<td>688</td>
<td>52,688</td>
</tr>
<tr>
<td>DSTC10-track2(Kim et al., 2021b)</td>
<td>2,919</td>
<td>55,377</td>
</tr>
<tr>
<td>DSTC10-track3(Kottur et al., 2021)</td>
<td>11,244</td>
<td>110,767</td>
</tr>
<tr>
<td>MedicalDialog(Zeng et al., 2020)</td>
<td>482</td>
<td>981</td>
</tr>
<tr>
<td>Self-Dialog(Fainberg et al., 2018)</td>
<td>24,165</td>
<td>372,720</td>
</tr>
<tr>
<td>WOW(Dinan et al., 2018)</td>
<td>22,311</td>
<td>201,999</td>
</tr>
<tr>
<td>TopicChat(Gopalakrishnan et al., 2019)</td>
<td>8,628</td>
<td>188,378</td>
</tr>
<tr>
<td>Persona-Chat(Zhang et al., 2018)</td>
<td>11,087</td>
<td>81,032</td>
</tr>
<tr>
<td>MMD(Saha et al., 2018a)</td>
<td>1,506,129</td>
<td>5,477,066</td>
</tr>
<tr>
<td>CSQA(Saha et al., 2018b)</td>
<td>197,001</td>
<td>4,153,092</td>
</tr>
<tr>
<td>AmazonQA(Gupta et al., 2019)</td>
<td>923,685</td>
<td>1,847,370</td>
</tr>
<tr>
<td>ChitChat(Will et al., 2020)</td>
<td>7,168</td>
<td>258,145</td>
</tr>
<tr>
<td>Total</td>
<td>3,217,058</td>
<td>19,578,028</td>
</tr>
</tbody>
</table>

Table 6: Statistics for each unlabeled dataset in UnPreDial.