FutureTOD: Teaching Future Knowledge to Pre-trained Language Model for Task-Oriented Dialogue

Weihao Zeng[1*], **Keqing He**[2*], **Yejie Wang**[1], **Chen Zeng**[1]
Jingang Wang[2], **Yunsen Xian**[2], **Weiran Xu**[1*]
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]Meituan, Beijing, China
{zengwh,wangyejie,chenzeng,xuweiran}@bupt.edu.cn
{hekeqing,wangjingang,xianyunsen}@meituan.com

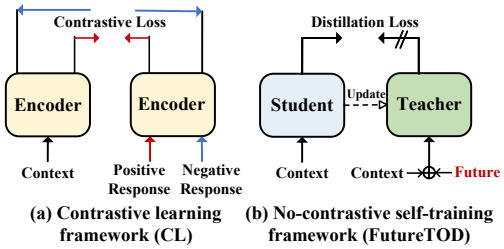
Abstract

Pre-trained language models based on general text enable huge success in the NLP scenario. But the intrinsical difference of linguistic patterns between general text and task-oriented dialogues makes existing pre-trained language models less useful in practice. Current dialogue pre-training methods rely on a contrastive framework and face the challenges of both selecting true positives and hard negatives. In this paper, we propose a novel dialogue pre-training model, FutureTOD, which distills future knowledge to the representation of the previous dialogue context using a self-training framework. Our intuition is that a good dialogue representation both learns local context information and predicts future information. Extensive experiments on diverse downstream dialogue tasks demonstrate the effectiveness of our model, especially the generalization, robustness, and learning discriminative dialogue representations capabilities. [1]

1 Introduction

Pre-trained language models (Devlin et al., 2019; Liu et al., 2019) based on a massive scale of general text corpora (Zhu et al., 2015) have been commonly used in many NLP applications. Finetuning models on these PLMs significantly improves the performance of various downstream tasks, especially natural language understanding. Despite their success, directly applying them to conversational corpora is proved to be suboptimal due to the large linguistic gap between conversations and plain text (Rashkin et al., 2019; Wolf et al., 2019). Therefore, it's vital to explore dialogue-specific pre-trained models for solving various downstream dialogue tasks.

Early pre-trained dialogue language models use chit-chat corpora from social media, such as Twitter or Reddit, aiming at retrieval (Henderson et al., 2019) and dialogue response generation (Zhang et al., 2020). These open-domain dialogues are usually short, noisy, and without specific chatting goals. Further, a more practical scenario, task-oriented dialogue (TOD), is attracting more attention. TOD has explicit goals (e.g. restaurant reservation) and many conversational interactions like belief states and database information, making language understanding and policy learning more complex than those chit-chat scenarios. Each TOD dataset is usually small because collecting and labeling such data are time-consuming. Therefore, in this paper, we focus on unsupervised dialogue pre-training for task-oriented dialogues.



(a) Contrastive learning framework (CL) (b) No-contrastive self-training framework (FutureTOD)

Figure 1: Comparison of different dialogue pre-training paradigms. The contrastive models learn context representations by pulling together positive pairs and pushing apart negative pairs. In contrast, our FutureTOD employs a self-training framework to distill future knowledge to context representations and dismiss the requirements of contrastive pairs.

Previous TOD pre-training methods usually follow a contrastive learning (CL) framework (Chen et al., 2020; He et al., 2020) as shown in Figure 1(a). CL aims to pull together semantically similar (positive) pairs and push apart semantically dissimilar (negative) pairs. SimCSE (Gao et al., 2021) employs Dropout (Srivastava et al., 2014) augmentation to construct positive pairs by passing a sentence through the encoder twice, resulting in superior performance for learning plain text representations. However, it performs poorly in the

*The first two authors contribute equally. Weiran Xu is the corresponding author.

[1]Our code, models and other related resources are publicly available at https://github.com/Zeng-WH/FutureTOD

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6532–6546

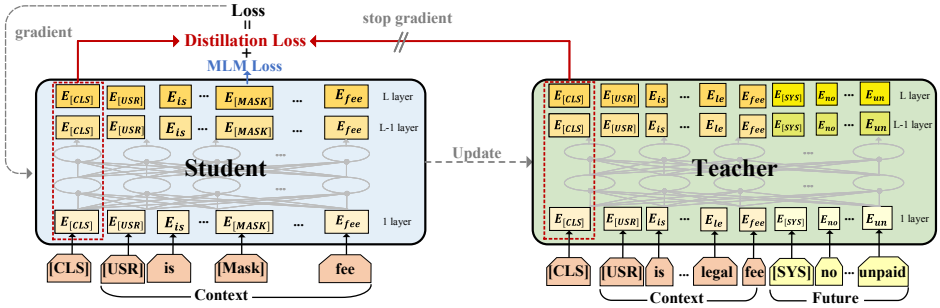


Figure 2: Overall architecture of FutureTOD. For brevity, we show only one system response utterance as future.

dialogue domain because of ignoring the intrinsic properties of dialogue data (Zhou et al., 2022). TOD-BERT (Wu et al., 2020) takes the dialogue context[2] and next response as a positive pair thus achieving promising performance on the response selection task. However, there is a large discrepancy in both semantics and data statistics between each response and its context [3], which reduces its generalization ability to other dialogue tasks. Further, DSE (Zhou et al., 2022) learns from dialogues by taking consecutive utterances of the same dialogue as positive pairs. But the assumption that consecutive utterances represent similar semantics fails sometimes when answers are general and ubiquitous. Along with the issues of choosing positive pairs, these models regard other instances in the same batch as negative samples, which also induces potential noise to contrastive learning (Arora et al., 2019), such as false negatives (Huynh et al., 2022; Chen et al., 2022) and relying on a large batch size (He et al., 2020). Overall, these contrastive methods face the challenges of both selecting true positive pairs and negative pairs that we aim to solve using a new non-contrastive pre-training framework.

In this paper, we propose a novel dialogue pre-training model, FutureTOD, which distills future knowledge to the representation of the previous dialogue context using future utterances based on a standard Transformer architecture BERT (Devlin et al., 2019). We argue that a good dialogue representation both learns local context information and predicts future knowledge. Instead of existing contrastive works, we employ a self-training framework and dismiss the requirements of contrastive pairs. As shown in Figure 1(b), we first use a student model to construct the dialogue representation of an input dialogue context. Next, we concatenate the context and following utterances and get its full representation using a teacher model. Our goal is to align the original context representation with the full representation containing future knowledge. The weights of the teacher are updated by the student periodically (He et al., 2020; Baevski et al., 2022; Liu et al., 2022). We evaluate FutureTOD on various task-oriented dialogue tasks, including intent classification, out-of-domain detection, dialogue state tracking, dialogue act prediction, and response selection. Experiment results demonstrate that FutureTOD significantly outperforms TOD-BERT, DSE, and other strong baselines in all the scenarios. We also observe FutureTOD has stronger capabilities on generalization, robustness and learning discriminative representations.

Our contributions are: (1) We propose a novel TOD dialogue pre-training model, FutureTOD, which distills future knowledge to dialogue representations. To the best of our knowledge, we are the first to use a non-contrastive self-training framework and knowledge distillation for dialogue pre-training. (2) Our model achieves consistent improvements on diverse downstream dialogue tasks over strong baselines. Extensive analyses prove the generalization, robustness, and learning discriminative dialogue representations capabilities.

2 Model

2.1 Overall Architecture

The overall architecture of FutureTOD is shown in Figure 2. We adopt BERT-base-uncased[4] as our backbone following TOD-BERT (Wu et al., 2020). We first add two special role tokens [USR]

[2]Throughout this paper, we denote a system turn including all the system sentences as the response (utterance), and all the history turns as the dialogue context.

[3]In the implementation of TOD-BERT, the context is often the concatenation of 5 to 15 utterances but the response is only a single utterance.

[4]https://huggingface.co/bert-base-uncased

or [SYS] to the prefix of each utterance and concatenate all the utterances in the same dialogue into one flat sequence. Then we split each dialogue at a randomly selected turn t to get the context and future sequences. We encode the context using a student model and obtain the output of [CLS] as the original dialogue representation. Next, we construct training targets by encoding the context and future using a teacher model. Both the student and teacher are the same BERT but the weights of the teacher are updated by the student periodically. The learning goal is to align the original context representation with the full representation containing future knowledge. We assume a good dialogue representation can't only capture local context information but also predict future knowledge.

2.2 Learning Future Knowledge

Notation We use the collected datasets by TOD-BERT (Wu et al., 2020) as our pre-training corpus. For each dialogue, we first transform it into a token sequence. Following previous work (Wu et al., 2020; Zhou et al., 2022), we add two special role tokens [USR] or [SYS] to the prefix of each utterance and concatenate all the utterances into one flat sequence $D = \{U_1, S_1, \ldots, U_n, S_n\}$. U_1 and S_1 denotes the user utterance and system utterance, respectively. n is the turn number of the dialogue.

Learning Framework Different from existing contrastive methods, we employ a self-training (van Engelen and Hoos, 2019; Grill et al., 2020) framework to distill future knowledge to the representation of the dialogue context using future utterances. The advantages are two-fold: (1) Our self-training framework doesn't require contrastive pairs thus alleviating the noise of selecting positive and negative samples. (2) Learning future knowledge encourages the model to align representations in the same latent space instead of pulling together representations of context and response belonging to different distributions. We first split each dialogue at a randomly selected turn t, so we get the context $C = \{U_1, S_1, \ldots, U_t\}$ and the future $F = \{S_t, U_{t+1}, S_{t+1}, \ldots, U_n, S_n\}$. Then we use a student model to encode the context and a teacher model to encode the context with the future. We denote the [CLS] output of the student model as h_S and the teacher as h_T. We hope the student model can capture future information while modeling the local semantics. So we design a distillation loss $\mathcal{L}_{dis}$ by minimizing the discrepancy between h_S and h_T:

$$\mathcal{L}_{dis} = \|h_S - h_T\|_2 \tag{1}$$

To explore different granularity of future information, we randomly select a ratio of future utterances from one utterance S_t to the whole utterances $\{S_t, U_{t+1}, S_{t+1}, \ldots, U_n, S_n\}$. Besides, we find performing distillation loss on multiple layers rather than only the top layer also gives consistent improvements (see Section 4.1). So, the final distillation loss $\mathcal{L}_{dis}$ is:

$$\mathcal{L}_{dis} = \sum_{l=1}^{L} (\left\| h_S^l - h_T^l \right\|_2) \tag{2}$$

where l is the l-th layer of BERT-base. We also try to apply normalization to h_S and h_T and other distillation objectives but do not observe significant change. Along with $\mathcal{L}_{dis}$, we also keep the traditional masked language modeling (MLM) (Devlin et al., 2019) loss $L_{mlm} = -\sum_{m=1}^{M} \log P(x_m)$ following Wu et al. (2020), where M is the total number of masked tokens and $P(x_m)$ is the predicted probability of the token x_m over the vocabulary size. Note that we only perform MLM on the student model. Therefore, the total loss is:

$$\mathcal{L} = \mathcal{L}_{dis} + \mathcal{L}_{mlm} \tag{3}$$

We simply sum them up and achieve the best performance in our experiments.

Parameter Updating We employ a simple algorithm to optimize the parameters of the student and teacher models iteratively. (1) **Stage 1**: We first use Eq 3 to perform gradient updating to optimize the student model and keep the teacher model fixed. We denote the interval as E epochs.[5] (2) **Stage 2**: After Stage 1, we directly assign student parameters to the teacher. The process of our method is summarized in Algorithm 1.

3 Experiment

3.1 Pre-training Setting

Pre-training Corpus We use the corpus collected by Wu et al. (2020), including 9 publicly available task-oriented datasets: MetaLWOZ (Lee et al., 2019), Schema (Rastogi et al., 2020), Taskmaster (Byrne et al., 2019), MWOZ (Budzianowski et al., 2018), MSR-E2E (Li et al., 2018), SMD (Eric et al., 2017), Frames (Asri et al., 2017), WOZ (Mrksic et al., 2017), CamRest676 (Rojas-Barahona et al., 2017). We show the full statistics in Appendix A.

[5] We empirically find $E = 10$ is the best. Please see a more detailed analysis in Section 4.1.

Algorithm 1 FutureTOD

Initialization: Teacher T, Student S, Interval E, Total Epoch M
Input: Context C, Future F
for m in [1, M] **do**
  Using S to get the output h_S of C
  Using T to get the output h_T of $C + F$
  Calculating the distillation loss L_{dis} in Equation 2
  Calculating the MLM loss L_{mlm}
  Using $L = L_{dis} + L_{mlm}$ to update S
  if m % E == 0 **then**
    Assigning S parameters to the T
  end if
end for
Output: S

Baselines We compare FutureTOD with other strong baselines. BERT (Devlin et al., 2019) and BERT-mlm denotes the original BERT-base-uncased pre-trained on a large text corpus and continual pre-trained BERT using MLM on our pre-training dialogue corpus, respectively. DialoGPT (Zhang et al., 2020) is a dialogue generation model via a language modeling target. SimCSE (Gao et al., 2021) uses Dropout to construct positive pairs and is further pre-trained on the same TOD corpus. TOD-BERT (Wu et al., 2020) uses a contrastive response selection objective by treating a response utterance and its dialogue context as positive pair. DSE (Zhou et al., 2022) takes consecutive utterances of the same dialogue as positive pairs.[6] Note that we focus on the unsupervised TOD pre-training, so we don't compare supervised methods using labeled NLI datasets (Williams et al., 2018) or dialogue act labels (He et al., 2022b).

Pre-trainging Details We train FutureTOD with a batch size of 32 and a maximum input length set of 512, respectively. Both the teacher and student models are initialized by BERT-base-uncased. Adam optimizer and a linear learning rate scheduler are employed for optimization with an initial learning rate of 5e-5 and a dropout ratio of 0.2. The mask ratio, teacher's update frequency, and the number of layers representations are set to 15%, 10 epoch, and 12 respectively. Experiments take 3 days with an early-stopped strategy based on perplexity scores of a held-out development conducted on eight NVIDIA Tesla A100 GPUs. The average length of context and response are 86.04 and 48.10 tokens respectively. The average number of utterances in context and response are 5.95 and 3.48 respectively. We use the pre-trained BERT-MLM and pre-trained TOD-BERT released by the original paper (Wu et al., 2020), and pre-trained DSE model released by Zhou et al. (2022) respectively. We use Dropout to construct positive pairs to re-implement SimCSE (Gao et al., 2021). For a fair comparison, we augment every single utterance obtained through Dropout on our pre-training corpora.

[6]We choose the unsupervised version of DSE in the original paper as our baseline for fair comparison.

3.2 Finetuning Setting

We finetune these pre-trained LMs on the following four core downstream tasks in a task-oriented system: intent recognition, dialogue state tracking, dialogue act prediction, and response selection. Following Wu et al. (2020), we only use the LMs and avoid adding too many additional components except a classification head. We use the representation of the [CLS] token as the utterance representation here. Additionally, we provide the performance of the mean pooling in Appendix D. For fair comparison, we use the same architecture for all the baselines. Along with the full data setting, we also randomly sample a few labeled training examples as the few-shot learning settings. More hyperparameters details can be seen in Appendix B.

Intent Recognition is a multi-class classification task, where the model predicts one intent label given an input sentence. We use the [CLS] embeddings as the dialogue representation and a softmax classification head. The model is trained with cross-entropy loss. We use OOS (Larson et al., 2019) intent dataset, which covers 151 intent classes over ten domains, including 150 in-domain intents and one out-of-domain (OOD) intent. We treat the OOD intent as an additional class following TOD-BERT. We report classification accuracy and recall.

Dialogue State Tracking is regarded as a multi-class classification task based on a pre-defined ontology. We use dialogue history as input and predict slot values for each (domain, slot) pair at each dialogue turn. The model is trained with cross-entropy loss summed over all the pairs. We use a widely-used TOD dataset MWOZ 2.1 (Budzianowski et al., 2018) across seven different domains. We report joint goal accuracy and slot accuracy. The former

	Model	Acc (all)	Acc (in)	Acc (out)	Recall (out)
1-Shot	BERT	29.3%	35.7%	81.3%	0.4%
	BERT-mlm	38.9%	47.4%	81.6%	0.5%
	SimCSE	29.9%	36.4%	81.7%	0.6%
	TOD-BERT	42.5%	52.0%	81.7%	0.1%
	DSE	42.3%	51.7%	81.8%	0.4%
	FutureTOD	**43.1%***	**52.2%**	**81.8%**	**2.1%***
10-Shot	BERT	75.5%	88.6%	84.7%	16.5%
	BERT-mlm	76.6%	90.5%	84.3%	14.0%
	SimCSE	74.5%	88.9%	83.5%	9.6%
	TOD-BERT	77.3%	**91.0%**	84.5%	15.3%
	DSE	77.8%	90.8%	85.2%	19.1%
	FutureTOD	**78.1%**	90.8%	**85.5%***	**20.5%***
Full (100-shot)	BERT	84.9%	95.8%	88.1%	35.6%
	DialoGPT	83.9%	95.5%	87.6%	32.1%
	BERT-mlm	85.9%	96.1%	89.5%	46.3%
	SimCSE	82.3%	94.7%	86.6%	26.6%
	TOD-BERT	86.6%	**96.2%**	89.9%	43.6%
	DSE	84.3%	95.8%	87.7%	32.5%
	FutureTOD	**87.2%***	96.0%	**90.0%**	**47.6%***

Table 1: Intent recognition results on the OOS dataset. Acc(all), Acc(in), Acc(out) denotes the overall accuracy, in-domain intent accuracy and out-of-domain intent accuracy. The numbers with * are significant using t-test with $p < 0.01$.

considers true if and only if all the predicted values exactly match its ground truth values at each dialogue turn while the latter individually compares each (domain, slot, value) triplet to its ground truth label. Joint goal accuracy is the main metric.

Dialogue Act Prediction is a multi-label classification task where the model takes dialogue history as input and predicts the system actions. The model is trained with binary cross-entropy loss summed over all the actions. For prediction, we set the threshold to 0.5. We use two datasets MWOZ (Budzianowski et al., 2018) and DSTC2 (Henderson et al., 2014). Following Wu et al. (2020), we use the same data preprocessing to uniform the original dialogue acts to a general format. We report micro-F1 and macro-F1 scores for the dialogue act prediction task.

Response Selection is a ranking task where the model selects the most relevant response from a candidate pool given an input dialogue history. We use a shared pre-trained LM to encode the dialogue and each response respectively and compute its cosine similarity score. We randomly sample several system responses from the corpus as negative samples. In our experiments, we set batch size equals to 25 for all the models. We also use MWOZ and DSTC2 as our evaluation datasets. We use k-to-100 accuracy as the metric. For each history, we combine its ground-truth response with 99 randomly sampled responses and rank these 100 responses based on their similarities with the query in the embedding space. The k-to-100 accuracy represents the ratio of the ground-truth response being ranked at the top-k.

3.3 Main Results

Intent Recognition We evaluate our FutureTOD on the intent recognition dataset OOS, including in-domain (IND) and out-of-domain (OOD) in Table 1. We find FutureTOD outperforms all the baselines on 10 of 12 metrics, especially with significant improvements in overall accuracy and OOD metrics. SimCSE (82.3% Acc(all)) is even worse than the original BERT (84.9% Acc(all)) in the full setting. Moreover, the 1.5 drop of Acc(out) is more significant than 1.1 of Acc(in), demonstrating that SimCSE ignores intrinsic dialogue structures and fails to model the relations between each utterance in the same dialogue. We also find TOD-BERT achieves comparable performance on Acc(in) except Recall(out), indicating the robustness of our method. Surprisingly, a recent strong baseline DSE performs poorly in the full setting. We argue the assumption that consecutive utterances represent similar semantics may fail in practical dialogues. Generally, FutureTOD achieves comparable or higher performance on in-domain intent accuracy, but significant improvements on out-of-domain accuracy, which proves the robustness and generalization ability of our method.

Dialogue State Tracking Table 2 displays the results of dialogue state tracking on MWOZ 2.1. Our FutureTOD achieves state-of-the-art results on 9 of 10 metrics. We find our method obtains significant improvements on Joint Acc than Slot Acc, showing the superiority of modeling overall dialogue context. Although these baselines achieve fair results on each (domain, slot, value) triplet, we observe they tend to overfit to the easy slot value pairs with high accuracy but fail to recognize hard ones, leading to poor overall joint goal accuracy. For example, FutureTOD outperforms DSE by 0.1% on Slot Acc but 0.5% on Joint Acc. All the results show the effectiveness of our method.

Dialogue Act Prediction Table 3 shows the results of dialogue act prediction on MWOZ and DSTC2. Our FutureTOD achieves state-of-the-art results on all the metrics. We find our method obtains comparable performance only using 10% data than the baselines using 100% data, which verifies the superior few-shot learning capability. We

Model	1% Data		5% Data		10% Data		25% Data		Full Data	
	Joint Acc	**Slot Acc**	**Joint Acc**	**Slot Acc**	**Joint Acc**	**Slot Acc**	**Joint Acc**	**Slot Acc**	**Joint Acc**	**Slot Acc**
BERT	6.4%	84.4%	19.6%	92.0%	32.9%	94.7%	40.8%	95.8%	45.6%	96.6%
BERT-mlm	9.9%	**86.6%**	28.1%	93.9%	39.5%	95.6%	44.0%	96.4%	47.7%	96.8%
SimCSE	7.4%	84.8%	21.1%	91.6%	35.6%	95.0%	43.8%	96.3%	48.0%	96.8%
TOD-BERT	8.0%	85.3%	28.6%	93.8%	37.0%	95.2%	44.3%	96.3%	48.0%	96.9%
DSE	9.8%	86.3%	23.8%	93.0%	37.8%	95.5%	43.4%	96.3%	49.9%	97.0%
FutureTOD	**9.9%**	85.5%	**29.1%***	**94.1%***	**40.7%***	**95.8%**	**45.7%***	**96.5%**	**50.4%***	**97.1%**

Table 2: Dialogue state tracking results on MWOZ 2.1. We report joint goal accuracy (Joint Acc) and slot accuracy (Slot Acc) for the full data and few-shot settings. The numbers with * are significant using t-test with $p < 0.01$.

	Model	MWOZ		DSTC2	
		micro-F1	macro-F1	micro-F1	macro-F1
1% Data	BERT	84.0%	66.7%	77.1%	25.8%
	BERT-mlm	87.5%	73.3%	79.6%	26.4%
	SimCSE	81.0%	62.1%	78.9%	27.3%
	TOD-BERT	86.9%	72.4%	82.9%	28.0%
	DSE	82.9%	65.1%	72.4%	21.4%
	FutureTOD	**87.9%***	**75.0%***	**83.7%***	**31.0%***
10% Data	BERT	89.7%	78.4%	88.2%	34.8%
	BERT-mlm	90.1%	78.9%	91.8%	39.4%
	SimCSE	89.6%	77.8%	92.3%	40.5%
	TOD-BERT	90.2%	79.6%	90.6%	38.8%
	DSE	89.9%	79.4%	91.1%	39.0%
	FutureTOD	**91.0%***	**80.5%***	**93.6%***	**40.9%**
Full Data	BERT	91.4%	79.7%	92.3%	40.1%
	DialoGPT	91.2%	79.7%	93.8%	42.1%
	BERT-mlm	91.7%	79.9%	90.9%	39.9%
	SimCSE	91.6%	80.3%	91.5%	39.6%
	TOD-BERT	91.7%	80.6%	93.8%	41.3%
	DSE	91.7%	81.3%	92.6%	40.2%
	FutureTOD	**92.0%**	**81.9%***	**94.6%***	**44.6%***

Table 3: Dialogue act prediction results on MWOZ and DSTC2. The numbers with * are significant using t-test with $p < 0.01$.

	Model	MWOZ		DSTC2	
		1-to-100	3-to-100	1-to-100	3-to-100
1% Data	BERT	7.8%	20.5%	3.7%	9.6%
	BERT-mlm	13.0%	34.6%	12.5%	24.9%
	SimCSE	17.2%	32.6%	27.6%	46.4%
	TOD-BERT	-	-	37.5%	55.9%
	DSE	7.9%	21.2%	2.4%	6.1%
	FutureTOD	**35.8%***	**53.5%***	**39.5%***	**64.0%***
10% Data	BERT	20.9%	45.4%	8.9%	21.4%
	BERT-mlm	22.3%	48.7%	19.0%	33.8%
	SimCSE	37.2%	60.6%	42.0%	63.5%
	TOD-BERT	-	-	49.7%	66.6%
	DSE	24.8%	49.4%	42.0%	59.7%
	FutureTOD	**50.0%***	**72.8%***	**51.3%***	**70.0%***
Full Data	BERT	47.5%	75.5%	46.6%	62.1%
	DialoGPT	35.7%	64.1%	39.8%	57.1%
	BERT-mlm	48.1%	74.3%	50.0%	65.1%
	SimCSE	64.2%	85.4%	55.6%	70.5%
	TOD-BERT	65.8%	87.0%	56.8%	70.6%
	DSE	63.3%	85.3%	58.3%	72.0%
	FutureTOD	**68.5%***	**87.9%***	**58.4%**	**72.6%***

Table 4: Response selection evaluation results on MWOZ and DSTC2 for 1%, 10% and full data setting. We report 1-to-100 and 3-to-100 accuracy, which represents the ratio of the ground-truth response being ranked at the top-1 or top-3 given 100 candidates. The numbers with * are significant using t-test with $p < 0.01$.

find DSE performs poorly in the 1% data setting because the original DSE uses one utterance as the query and lacks the ability of modeling long context. In contrast, our model achieves consistent performance in all the settings, showing better generalization ability than previous baselines.

Response Selection Table 4 displays the results of response selection on MWOZ and DSTC2. Our FutureTOD achieves state-of-the-art results on all the metrics. Besides, we find the improvements in the 1% data setting are more significant than the full data. Note that TOD-BERT uses the response contrastive learning as the pre-training objective on full MWOZ training data so we don't report its results of few-shot learning. However, our method still significantly outperforms TOD-BERT on DSTC2 without using response selection loss. It proves FutureTOD learns generalized dialogue representations by distilling future knowledge to pre-trained models and performs well on downstream tasks.

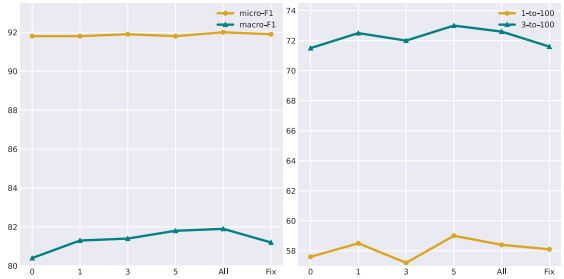
Overall, FutureTOD achieves state-of-the-art results for most of the downstream tasks while existing dialogue pre-trained models fail in specific tasks. The results demonstrate our pre-training method has strong generalization capability for diverse dialogue tasks. The results on out-of-domain intent recognization also prove its robustness.

4 Qualitative Analysis

4.1 Hyper-parameter Analysis

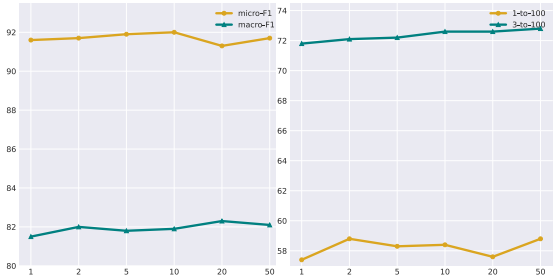
Effect of Max Future Length We randomly select a part of future utterances ranging from 1 to the max future length P. To explore the effect of different max future lengths, we set the P to 1, 3, 5, and All respectively. [7] If the $P = All$, we can randomly select any length of utterances from the whole future utterances. For comparison, we also add a baseline $P = Fix$ which must use the whole future utterances together. For example, if we have 5 future ut-

[7] If the real length of total future utterances is lower than the given max limit, we just randomly select from the whole future.



(a) MWOZ (b) DSTC2

Figure 3: Ablation study of max future lengths. We report the results of dialogue act prediction on MWOZ and response selection on DSTC2. The X-asix and Y-asix denotes the max future length and performance.



(a) MWOZ (b) DSTC2

Figure 4: Ablation study of the teacher's update frequency. We conduct dialogue act prediction on MWOZ and response selection on DSTC2. The X-asix and Y-asix denotes update frequency and performance.

terances $F = \{S_t, U_{t+1}, S_{t+1}, U_{t+2}, S_{t+2}\}$. When $P = 3$, we can select any length no longer than 3, such as $\{S_t\}$ or $\{S_t, U_{t+1}, S_{t+1}\}$; When $P = All$, we can select any length of future from the 5 utterances, that is $\{S_t\}$ or $\{S_t, U_{t+1}, S_{t+1}\}$ or F; When $P = Fix$, we can only select F. Figure 3 shows that FutureTOD generally gets improvements with increasing P. We argue that more future turns make the model learn comprehensive knowledge. We also observe that directly using all the future utterances like $P = Fix$ can't bring further improvements because diverse future knowledge with different granularity also makes an effect. An intuitive explanation is that too long future utterances possibly cause bias to a short dialogue context. Assuming a context only contains a single utterance but we always use ten, even more, future utterances to distill knowledge, the representation of the context will overfit to the future. Randomly selecting future information plays a role similar to Dropout (Srivastava et al., 2014). We leave more complicated selection strategies to future work, such as adaptively selecting the future for different lengths of context. We also conducted experiments using a teacher model that only encodes the future. However, the model's performance is poor. For detailed analysis, please refer to the Appendix C

Effect of Frequency of Updating Teacher FutureTOD updates the teacher model using the student parameters every E epoch. Figure 4 shows the effect of updating frequency E. We find $E = 10$ gets decent performance in general. We assume too small E makes the teacher tightly close to the student and prone to collapse while too large E can't produce a high-quality teacher model as learning signals and make the training slow. We also try other updating strategies such as momenta updating (He et al., 2020) and non-constant E but don't observe improvements. The simple strategy of updating every E epoch is simple and robust.

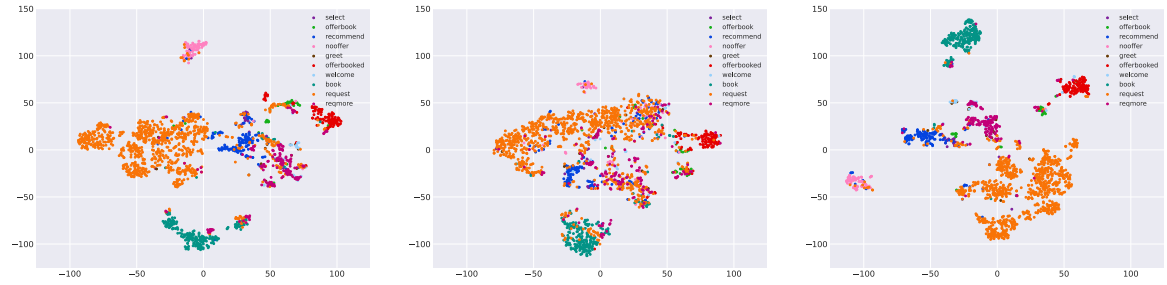
Top-K Layer	MWOZ		DSTC2	
	micro-F1	macro-F1	1-to-100	3-to-100
1	91.63%	80.46%	58.08%	72.11%
3	91.60%	80.49%	58.40%	72.16%
6	91.75%	81.02%	58.20%	**72.80%**
9	91.72%	80.89%	**58.51%**	72.79%
12	**91.95%**	**81.92%**	58.41%	72.60%

Table 5: Ablation study of using top-K layer representations for distillation. For example, $K = 3$ denotes we use the top 3 layers of BERT-base to compute Eq 2.

Effect of Distillation Layers We use the different top layers for the distillation loss Eq 3 in Table 5. We find adding more layers for distilling future knowledge can significantly improve performance. It indicates that different types of features extracted at different layers enhance learning different granularity of future information and improve downstream tasks.

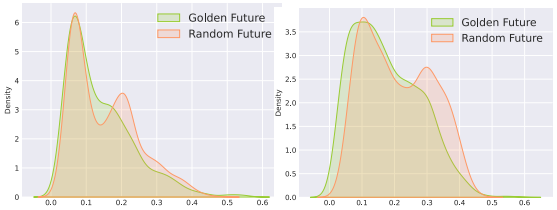
4.2 Visualization

Figure 5 shows the visualization of the system response representations of TOD-BERT, DSE and FutureTOD given the same input from the MWOZ test set. We use a pre-trained model to get [CLS] features and perform dimension reduction using the t-distributed stochastic neighbor embedding (tSNE). Different colors represent different dialogue act labels of the responses. We observe that FutureTOD builds compact and clearly separable dialogue representations for different clusters, which help distinguish semantically similar dialogues.



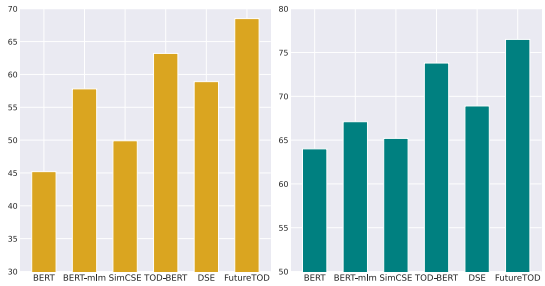
(a) TOD-BERT (b) DSE (c) FutureTOD

Figure 5: The tSNE visualization of TOD-BERT, DSE and FutureTOD representations of system responses in the MWOZ test set. Different colors represent different dialogue acts.



(a) MWOZ (b) DSTC2

Figure 6: Distance distribution curves of golden and random future. The X-axis denotes the MSE distance of representations between the dialogue history and the concatenation of history and golden or random response. The Y-axis denotes the ratio.



(a) MWOZ (b) DSTC2

Figure 7: The ratio of the test dialogue history where its distance between history and (history, golden response) is smaller than the one between history and (history, random response). Larger numbers denote better results.

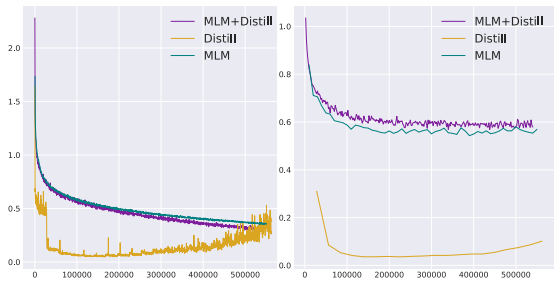
4.3 Understanding Future Knowledge

To understand whether our FutureTOD can capture future knowledge, we perform a qualitative analysis to exhibit the capability of predicting future information in Figure 6. For each dialogue history, we combine its golden response with 99 randomly sampled responses. Then we compute the mean square error (MSE) distance between the representations of the dialogue history and the concatenation of history and response using a pre-trained FutureTOD model. For these randomly sampled responses, we report the average distance. Figure 6 displays the distance distribution curves of golden and random future in the test set. The area under the shadow represents the ability of the model to predict the future. We find FutureTOD obtains similar representations corresponding to the golden future response. We also compute the average distance of all the test dialogues. We observe FutureTOD gets 1.449 of golden responses, smaller than 1.503 of random responses on MWOZ. Similar results are shown on DSTC2. They prove the effectiveness of FutureTOD capturing future knowledge.

Besides, we compare different pre-trained models in predicting future information in Figure 7. For each dialogue history in the test set, we compute the MSE distances between representations of dialogue history with/without golden or random responses. We assume the distances of golden responses are smaller than those of random responses. Therefore, we display the ratio of the test dialogue history where its distance of golden response is smaller than one of random response. As Figure 7 shows, we find FutureTOD obtains the highest ratio than the others, demonstrating the stronger capability of capturing future knowledge.

4.4 Learning Process

Figure 8 displays the training and evaluation learning curves in the pre-training stage. We show three pre-training objectives: MLM, Distill, and MLM+Distill(FutureTOD). We find that only Distill loss leads to an unstable learning process and can't converge. We argue that adding random masks to the input sequence of the student model makes the architecture asymmetric between the



(a) Training curves (b) Evaluation curves

Figure 8: Training and evaluation curves of different pre-training objectives. We scale up MLM loss by 50 times to display the three curves in the same figure.

student and teacher models, which is beneficial to preventing collapse. We also observe that adding another projection layer to the teacher model (Grill et al., 2020) or momentum updating (He et al., 2020) can't bring further improvements.

5 Related Work

Self-Supervised Learning Self-supervised learning (SSL) has been a very active area of research in CV, NLP, and speech. Contrastive methods (Chen et al., 2020; He et al., 2020) in computer vision achieve huge success in ImageNet. Further, Wu et al. (2020); Gao et al. (2021); Zhou et al. (2022) in NLP introduce contrastive methods to unsupervised sentence or dialogue representation learning. However, these methods suffer from large batch size (He et al., 2020), easy negatives (Wang and Liu, 2021), and false negatives (Huynh et al., 2022). Besides, carefully designing appropriate augmentation methods (Fang et al., 2020; Gao et al., 2021) is also challenging, especially in NLP. Another line of SSL is masked image/language/speech modeling. The most prominent model is BERT (Devlin et al., 2019) which randomly masks some of the input tokens to recover from the remaining input. Vision methods follow similar ideas and predict visual tokens (Dong et al., 2021) or input pixels (He et al., 2022a). Grill et al. (2020); Baevski et al. (2022) use a momentum encoder to bridge the gap between different augmentation or masked views. Different from these works, we use future utterances to distill knowledge to the representation of the previous dialogue context without any augmentation.

Dialogue Pre-trained Language Models Zhang et al. (2020) adopts the pre-trained GPT-2 model (Radford et al., 2019) on Reddit data to perform open-domain dialogue response generation. Gao et al. (2021); Wu et al. (2020); Zhou et al. (2022) adopt contrastive learning to learn text or TOD dialogue representations. They use Dropout (Srivastava et al., 2014) augmentation, context-response pair, and consecutive utterances to construct positive pairs, respectively. Henderson et al. (2020); Liu et al. (2021) use the similar idea to learn dialogue representations mainly for dialogue retrieval or response selection. Apart from these unsupervised methods, Zhou et al. (2022); He et al. (2022b) use labeled dialogue data to perform supervised or semi-supervised pre-training. They usually use dialogue acts or dialogue NLI labels (Williams et al., 2018). Since we focus on unsupervised pre-training in this paper, we don't compare these models and leave it to future work.

6 Conclusion

We propose a novel dialogue pre-training model, FutureTOD, which distills future knowledge to dialogue representations. Instead of existing contrastive works, we employ a simple self-training framework to learn from each other and dismiss the requirements of contrastive pairs. We perform comprehensive experiments on various task-oriented dialogue tasks, including intent classification, out-of-domain detection, dialogue state tracking, dialogue act prediction, and response selection. FutureTOD significantly outperforms TOD-BERT, DSE, and other strong baselines in all the scenarios. FutureTOD is of excellent performance and easy-to-deploy without modifying any model architecture.

Acknowledgements

We thank all anonymous reviewers for their helpful comments and suggestions. We are also grateful to the track organizers for their valuable work. This work was partially supported by National Key R&D Program of China No. 2019YFF0303300 and Subject II No. 2019YFF0303302, DOCOMO Beijing Communications Laboratories Co., Ltd, MoE-CMCC "Artifical Intelligence" Project No. MCM20190701. Jingang Wang is funded by Beijing Nova Program(Grant NO. 20220484098)

Limitations

Although FutureTOD achieves significant improvements over existing baselines, there are some directions to explore for future work: (1) In this paper, FutureTOD doesn't use any data augmentation strategies to enhance representations. We believe

existing augmentation methods will benefit further improving performance. (2) We design a simple technique of constructing the teacher. More complicated methods should be considered, such as multi-teacher and large teacher. (3) FutureTOD in this paper cares about dialogue understanding tasks like intent detection, dialogue state tracking, etc. We hope to extend the similar idea to the generative dialogue pre-trained models and larger TOD corpus. Besides, exploiting limited dialogue labels is also valuable to explore.

Ethics Statement

The datasets used in this paper are all public and have been checked before use to not include any information that names or uniquely identifies individual people or offensive content. However, since the datasets come from the Internet, potential bias may still be introduced. This paper does not contain any data collection or release, so there are no privacy issues. Our model is pre-trained on GPU, which may cause an environmental impact. This paper does not involve human annotation or research with human subjects.

References

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*.

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

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *EMNLP*.

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

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

Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. 2022. Incremental false negative detection for contrastive learning. *ArXiv*, abs/2106.03719.

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

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. 2021. Peco: Perceptual codebook for bert pre-training of vision transformers. *ArXiv*, abs/2111.12710.

Mihail Eric, Lakshmi. Krishnan, François Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. *ArXiv*, abs/1705.05414.

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

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.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. 2022a. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.

Wanwei He, Yinpei Dai, Binyuan Hui, Min Yang, Zhen Cao, Jianbo Dong, Fei Huang, Luo Si, and Yongbin Li. 2022b. Space-2: Tree-structured semi-supervised

contrastive pre-training for task-oriented dialog understanding. In *COLING*.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkvsi'c, Pei hao Su, Tsung-Hsien, and Ivan Vulic. 2020. Convert: Efficient and accurate conversational representations from transformers. *ArXiv*, abs/1911.03688.

Matthew Henderson, Blaise Thomson, and J. Williams. 2014. The second dialog state tracking challenge. In *SIGDIAL Conference*.

Matthew Henderson, Ivan Vulic, Daniel Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios P. Spithourakis, Tsung-Hsien Wen, Nikola Mrksic, and Pei hao Su. 2019. Training neural response selection for task-oriented dialogue systems. *ArXiv*, abs/1906.01543.

Tri Huynh, Simon Kornblith, Matthew R. Walter, Michael Maire, and Maryam Khademi. 2022. Boosting contrastive self-supervised learning with false negative cancellation. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 986–996.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. 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 1311–1316, Hong Kong, China. Association for Computational Linguistics.

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

Xiujun Li, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *ArXiv*, abs/1807.11125.

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021. Dialoguecse: Dialogue-based contrastive learning of sentence embeddings. *ArXiv*, abs/2109.12599.

Xingbin Liu, Jinghao Zhou, Tao Kong, Xianming Lin, and Rongrong Ji. 2022. Exploring target representations for masked autoencoders. *ArXiv*, abs/2209.03917.

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

Nikola Mrksic, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve J. Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *ACL*.

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *AAAI*.

Lina Maria Rojas-Barahona, Milica Gavsic, Nikola Mrksic, Pei hao Su, Stefan Ultes, Tsung-Hsien Wen, Steve J. Young, and David Vandyke. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *EACL*.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15:1929–1958.

Jesper E. van Engelen and Holger H. Hoos. 2019. A survey on semi-supervised learning. *Machine Learning*, 109:373–440.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*.

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

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 *EMNLP*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B. Dolan. 2020. Dialogpt : Large-scale generative pre-training for conversational response generation. In *ACL*.

Zhihan Zhou, Dejiao Zhang, Wei Xiao, Nicholas Dingwall, Xiaofei Ma, Andrew O. Arnold, and Bing Xiang. 2022. Learning dialogue representations from consecutive utterances. In *NAACL*.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

A Pre-training Data statistics

We use the corpus collected by Wu et al. (2020), including 9 publicly available task-oriented datasets: MetaLWOZ (Lee et al., 2019), Schema (Rastogi et al., 2020), Taskmaster (Byrne et al., 2019), MWOZ (Budzianowski et al., 2018), MSR-E2E (Li et al., 2018), SMD (Eric et al., 2017), Frames (Asri et al., 2017), WOZ (Mrksic et al., 2017), CamRest676 (Rojas-Barahona et al., 2017). The full statistics in Table 6. These existing datasets are open-source and have no ethical concerns.

B Finetuning Details

For BERT-mlm and TOD-BERT, we use the results reported by TOD-BERT (Wu et al., 2020) directly. We use the same hyperparameters for all the downstream tasks except the batch size and learning rate. We finetune all downstream tasks for 50 epochs with an early-stopped strategy evaluated on the validation set every 50 steps with patience set to 10. We respectively set batch size to 8, 25, 16 and 100 for intent recognition, dialogue state tracking, dialogue act prediction, and response selection. We choose the best learning rate from {2e-5, 5e-5, 7e-5, 1e-4, 2e-4} using grid search. We used the last layer's hidden states of the pre-trained model for downstream tasks. We also experimented with using hidden states from all layers, but find no significant change in performance.

C Only the Future

We use a student model to encode the context and a teacher model to encode both the context and the future in our method. We also conducted experiments using the teacher model without the context, but only with the future. However, as shown in Table 7, the latter model did not perform well. For example, in response selection, the top-1 accuracy decreased from 58.4% to 56.3%, and the top-3 accuracy decreased from 72.6% to 70.6%. In dialogue act prediction, the micro-F1 decreased from 92.0% to 90.9%, and the macro-F1 decreased from 81.9% to 81.3%. We analyzed that this is due to the model collapse caused by directly aligning context and response without negative samples like TOD-BERT.

Name	# Dialogue	# Utterance	Avg. Turn	# Domain
MetaLWOZ	37,884	432,036	11.4	47
Schema	22,825	463,284	20.3	17
Taskmaster	13,215	303,066	22.9	6
MWOZ	10,420	71,410	6.9	7
MSR-E2E	10,087	74,686	7.4	3
SMD	3,031	15,928	5.3	3
Frames	1,369	19,986	14.6	3
WOZ	1,200	5,012	4.2	1
CamRest676	676	2,744	4.1	1

Table 6: Data statistics for our pre-training task-oriented dialogue datasets.

Task	Metric	Method	
		C $\leftrightarrow$ F	C $\leftrightarrow$ C+F
Dialogue Act Prediction	micro-F1	90.9%	92.0%
	macro-F1	81.3%	81.9%
Response Selection	1-to-100	56.3%	58.4%
	3-to-100	70.6%	72.6%

Table 7: Ablation of the Teacher Input. We report the results of dialogue act prediction on MWOZ and response selection on DSTC2. C $\leftrightarrow$ C+F denotes the teacher model that encodes both the context and the future(our default setting). C $\leftrightarrow$ F denotes the teacher model that encodes only the future, without the context.

D Different Representation Methods

By default, we use the [CLS] token's representation as the utterance representation. To explore the impact of different utterance representation methods, we compare [CLS] token representations with the mean pooling of all the token representations. Table 8 shows that our FutureTOD model achieves comparable performance using both [CLS] and mean pooling. Both methods outperform the baselines. For instance, the FutureTOD(AVG) model achieves 87.0% accuracy for the intent recognition task, while FutureTOD(CLS) achieves 87.2%. These results surpass the 86.6% accuracy achieved by TOD-BERT(CLS), demonstrating the robustness of our model across different representation methods.

Task	Metric	Model		
		TOD-BERT(CLS)	**FutureTOD(AVG)**	**FutureTOD(CLS)**
Intent Recognition	Acc(all)	86.6%	87.0%	**87.2**%
	Acc(in)	**96.2**%	95.5%	96.0%
	Acc(out)	89.9%	**90.2**%	90.0%
	Recall(out)	43.6%	**48.8**%	47.6%
Dialogue State Tracking	Joint Acc	48.0%	50.1%	**50.4**%
	Slot Acc	96.9%	**97.1**%	**97.1**%
Dialogue Act Prediction	micro-F1	93.8%	**95.1**%	94.6%
	macro-F1	41.3%	**45.9**%	44.6%
Response Selection	1-to-100	56.8%	57.7%	**58.4**%
	3-to-100	70.6%	72.5%	**72.6**%

Table 8: Ablation study of different representation methods. We report the results of intent recognition on OOS, DST on MWOZ, dialogue act prediction on DSTC2, and response selection on DSCT2. TOD-BERT (CLS) and FutureTOD (CLS) denote using CLS token representation as the utterance representation. FutureTOD (AVG) denotes using the mean pooling of all tokens within the utterance as the utterance representation.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In section 7 Limitations, we discuss the limitations of our work

☑ A2. Did you discuss any potential risks of your work?
In section 8 Ethics Statement, we discuss the pre-training corpus come from the Internet, potential bias may be introduced.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
In Abstract section and Section one Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3, the data we use is scientific artifacts

☑ B1. Did you cite the creators of artifacts you used?
Section 3.1

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In section 3. We used the existing datasets following TOD-BERT, so we use the original license same as it.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
In Section 3.1, we said we using data following Tod-BERT

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We discuss in Ethics Statement. The data have been checked before use to not include any information that names or uniquely identifies individual people or offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In Section 3 and Appendices A

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In Table 6

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

section 3 and 4

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendices B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
In section 3 and 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We use huggingface, we mention in Section 2

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.