BootTOD: Bootstrap Task-oriented Dialogue Representations by Aligning Diverse Responses

Weihao Zeng[1], **Keqing He**[2], **Yejie Wang**[1], **Dayuan Fu**[1], **Weiran Xu**[1]
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]Meituan, Beijing, China
{zengwh,wangyejie,fdy,xuweiran}@bupt.edu.cn
hekeqing@meituan.com

Abstract

Pre-trained language models have been successful in many scenarios. However, their usefulness in task-oriented dialogues is limited due to the intrinsic linguistic differences between general text and task-oriented dialogues. Current task-oriented dialogue pre-training methods rely on a contrastive framework, which faces challenges such as selecting true positives and hard negatives, as well as lacking diversity. In this paper, we propose a novel dialogue pre-training model called BootTOD. It learns task-oriented dialogue representations via a self-bootstrapping framework. Unlike contrastive counterparts, BootTOD aligns context and context+response representations and dismisses the requirements of contrastive pairs. BootTOD also uses multiple appropriate response targets to model the intrinsic one-to-many diversity of human conversations. Experimental results show that BootTOD outperforms strong TOD baselines on diverse downstream dialogue tasks.

Keywords: Task-Oriented Dialogues, Self-BootStrapping, Dialogue Pretraining

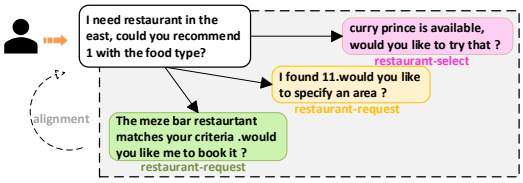


Figure 1: The same context may have multiple appropriate responses in a task-oriented dialogue.

1. Instroduction

Previous unsupervised pre-training models for Task-Oriented Dialogues have employed contrastive learning (CL) framework (Chen et al., 2020; He et al., 2020), with the goal of bringing semantically similar (positive) pairs closer together while separating semantically dissimilar (negative) pairs. TOD-BERT (Wu et al., 2020) employs dialogue history and corresponding responses as positive pairs, achieving excellent performance on response selection tasks but only marginal improvements on other dialogue tasks. This is due to the fact that TOD-BERT selects responses from other dialogues as negatives, and these negative responses may be suitable for the current context, resulting in false negatives (Huynh et al., 2022; Chen et al., 2022). Furthermore, DSE (Zhou et al., 2022) learns from dialogues by using consecutive utterances from the same dialogue as positive pairs. However, this assumption that consecutive utterances represent similar semantics can fail when answers are general and ubiquitous.

Despite the remarkable progress of previous TOD PLMs, there are still two challenges. First, these contrastive methods suffer from selecting noisy positive and negative pairs, such as false negatives (Huynh et al., 2022; Chen et al., 2022), unreasonable assumptions (Zhou et al., 2022) and relying on a large batch size (He et al., 2020). Limited exploration has been attempted to perform dialogue pre-training using a non-contrastive framework. Second, most work ignores the one-to-many property in conversation where multiple responses can be appropriate under the same conversation context (as shown in Figure 1). PLATO (Bao et al., 2019) proposes discrete latent variables to improve utterance-level diversity in open-domain dialog generation, but none of the previous TOD pre-training methods consider such one-to-many property which is also prevalent in task-oriented dialogues. Current TOD PLMs tend to capture the most common dialog policy but ignore rarely occurred yet feasible user behaviors, resulting in duplicate and plain responses.

To solve the issues, in this paper, we propose a novel dialogue pre-training model, BootTOD, which learns task-oriented dialogue representations via a self-bootstrapping framework. Instead of contrastive counterparts, we introduce a self-bootstrapping framework to align context and context+response representations and dismiss the requirements of contrastive pairs. Besides, BootTOD aligns the context representation with multiple appropriate response targets to model the intrinsic one-to-many diversity of human conversations. Specifically, we use a BERT model to encode the dialogue context and align its representation with

Weiran Xu is the corresponding author.

LREC-COLING 2024, pages 2458–2467
20-25 May, 2024.

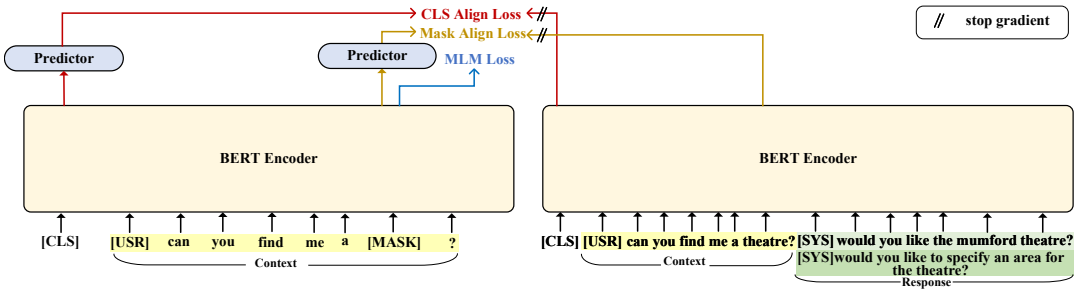


Figure 2: Overall architecture of our proposed BootTOD.

the full sequence containing context and response. We argue that a good dialogue representation both learns local context information and predicts future knowledge. Our alignment objectives contain three aspects: dialogue representation alignment using [CLS], [MASK] token representation alignment, and original MLM loss (Devlin et al., 2019). We evaluate BootTOD on various task-oriented dialogue tasks, including intent classification, dialogue state tracking, dialogue act prediction, and response selection. Results show that BootTOD achieves consistent improvements over strong TOD baselines in all the scenarios, which proves its generalization capability.

Our contributions are: (1) We propose a novel dialogue pre-training model, BootTOD, which uses a self-bootstrapping framework to align the context representation with diverse response targets. (2) Our model outperforms strong TOD baselines on diverse downstream dialogue tasks

2. Model

2.1. Overall Architecture

Figure 2 displays the overall architecture of BootTOD. Following previous work (Wu et al., 2020; Zhou et al., 2022; Zeng et al., 2023), we adopt BERT-base-uncased[1] as our backbone. We add two special role tokens [USR] 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 response. We encode the dialogue context via a predictor layer and align its representation with the full sequence containing context and response, including [CLS] alignment, [MASK] token alignment, and mask language model (MLM). We aim to make the model capture local context information and predict future knowledge.

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

2.2. Bootstrap Task-oriented Dialogue Representations

For each dialogue, we first transform it into a token sequence $D = \{U_1, S_1, \ldots, U_n, S_n\}$. U_i and S_i denote the user utterance and system utterance with a prefix of two special role tokens [USR] or [SYS], respectively. n is the turn number of the dialogue.

Compared to existing contrastive methods, we employ a self-bootstrapping framework to align context and context+response representations to learn future knowledge. The advantages are two-fold: (1) Our 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. Assuming we split each dialogue at a randomly selected turn t, the context is $C = \{U_1, S_1, \ldots, U_t\}$ and the response is $R = \{S_t, U_{t+1}, S_{t+1}, \ldots, U_n, S_n\}$. Note that in this paper, we denote a response as a multi-turn collection ending with a system utterance. We concatenate all the utterances into sequence and use a shared BERT encoder f to process the context and context+response sequences respectively. Inspired by (Chen and He, 2020; Grill et al., 2020), we use a shared predictor MLP head h to transform the representations of the context C. We hope the context representation can predict future information while modeling the local semantics. Therefore, we design three alignment objectives as follows.

Dialogue Representation Alignment Loss

$$\mathcal{L}_{cls} = \sum_{l=1}^{L} \left\| h(c_{cls}^{l}) - r_{cls}^{l} \right\|_2 \quad (1)$$

where l is the l-th layer of BERT-base and h is the predictor. c_{cls}^{l} and r_{cls}^{l} are the l-th layer [CLS] representations of context and context+response, respectively. We find perform alignment loss on multiple layers rather than only the top layer gives consistent improvements (see Section 4.2). We also try to apply normalization to c_{cls}^{l}, r_{cls}^{l} and other forms of objectives but do not observe significant

change.

Token Representation Alignment Loss Apart from the dialogue-level alignment, we also propose a token-level alignment loss to learn fine-grained token representations.

$$\mathcal{L}_{mask} = \sum_{m=1}^{M} \sum_{l=1}^{L} \left\| h(c^l_{mask,m}) - r^l_m \right\|_2 \quad (2)$$

where M is the total number of masked tokens. $c^l_{mask,m}$ is the l-th layer [MASK] token representation of context and r^l_m is the corresponding original token's l-th layer representation of context+response. Note that we only perform mask strategy to the context instead of context+response sequence, which provides more accurate contextual targets to the context representations.

Mask Language Model Loss We also keep the traditional masked language modeling (MLM) (Devlin et al., 2019) loss following Wu et al. (2020).

$$\mathcal{L}_{mlm} = -\sum_{m=1}^{M} \log P\left(x_m\right) \quad (3)$$

where $P(x_m)$ is the predicted probability of the mask token x_m over the vocabulary size.

We simply sum them up and achieve the best performance in our experiments. Inspired by (Chen and He, 2020), we employ a stop-gradient strategy to the representations of context+response as shown in Figure 2 to prevent collapsing. To explore the diversity of different response targets, we randomly select a ratio of consecutive response utterances from $R = \{S_t, U_{t+1}, S_{t+1}, \ldots, U_n, S_n\}$, such as $\{S_t\}$ and $\{S_t, U_{t+1}, S_{t+1}\}$. And the last turn of response must be a system utterance. For the same context with multiple appropriate responses, BootTOD aligns the context representation with diverse response targets by iterating over the whole dataset).

3. Experiment

3.1. Training Details

Pre-training Corpus We utilize nine task-oriented datasets that collected by Wu et al. (2020).

Baselines We compare BootTOD against several strong baselines, including BERT (Devlin et al., 2019), BERT-mlm (continual pre-training on dialogues), DialoGPT (Zhang et al., 2020), SimCSE (Gao et al., 2021), TOD-BERT (Wu et al., 2020), and DSE (Zhou et al., 2022). We focus on unsupervised TOD pre-training so we exclude comparisons with supervised methods that utilize labeled NLI datasets (Williams et al., 2018; Welleck et al., 2019) or dialogue act labels(He et al., 2022b).

Pre-trainging Details BootTOD's training uses a batch size of 48, a maximum input length of 512, and initiates with BERT-base-uncased. It's optimized with Adam, a learning rate of 5e-5, and 0.2 dropout. The mask ratio is 15%, and the predictor head has two layers plus ReLU, with dimensions of 768 and 512. After pre-training, we retain the Bert encoder parameters and remove the MLP head for subsequent fine-tuning. The 3-day pre-training involves an early-stop strategy based on perplexity, using eight NVIDIA Tesla A100 GPUs.

Finetuning Details For BERT-mlm and TOD-BERT, we directly use the results reported by TOD-BERT (Wu et al., 2020). We adopt the same hyperparameters for all downstream tasks.

	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%
	BootTOD	**44.0%***	**53.5%***	81.7%	**1.0%**
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%
	BootTOD	**78.4%***	**91.1**%	**85.6%***	**21.2%***
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%
	BootTOD	**88.2%***	96.1%	**91.1%***	**52.7%***

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 $p < 0.01$.

Model	5 % Data		10 % Data		Full Data	
	Joint Acc	Slot Acc	Joint Acc	Slot Acc	Joint Acc	Slot Acc
BERT	19.6%	92.0%	32.9%	94.7%	45.6%	96.6%
BERT-mlm	28.1%	93.9%	39.5%	95.6%	47.7%	96.8%
SimCSE	21.1%	91.6%	35.6%	95.0%	48.0%	96.8%
TOD-BERT	28.6%	93.8%	37.0%	95.2%	48.0%	96.9%
DSE	23.8%	93.0%	37.8%	95.5%	49.9%	97.0%
BootTOD	**30.3%***	**94.2%***	**40.8%***	**96.0%***	**50.7%***	**97.2%**

Table 2: Dialogue state tracking results on MWOZ 2.1. Joint Acc and Slot Acc denote the joint goal accuracy and slot accuracy. The numbers with * are significant using t-test with $p < 0.01$.

3.2. Main Results

We evaluated various pre-trained language models on four core task-oriented dialogue tasks (We detail the evaluation tasks and evaluation metrics in the Appendix A.). We conducted experiments using the whole dataset, as well as a few-shot setting. The few-shot setting here aligns with TOD-BERT (Wu et al., 2020) and FutureTOD (Zeng et al., 2023). Specifically, this involves fine-tuning using just 1% or 10% of the entire dataset, as opposed to using the full dataset for fine-tuning. The few-shot experi-

	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%
	BootTOD	**87.7%**	**73.8%***	**85.8%***	**33.9%***
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%
	BootTOD	**90.9%***	**80.7%***	**93.9%***	**42.8%**
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%
	BootTOD	**91.8%**	**82.3%***	**95.9%***	**46.5%***

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%
	BootTOD	**37.0%***	**60.5%***	**38.1%***	**61.3%***
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%
	BootTOD	**50.0%***	**72.0%***	**52.3%***	**69.6%***
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%
	BootTOD	**68.8%***	**87.6%***	**59.1%***	**72.3%**

Table 4: Response selection results on MWOZ and DSTC2. 1-to-100 and 3-to-100 denote 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 $p < 0.01$.

Model	DSTC2		MWOZ	
	micro-F1	macro-F1	1-to-100	3-to-100
BootTOD	**95.85%**	**46.53%**	68.79%	87.61%
w/o Mask Align	95.58%	46.17%	68.74%	87.70%
w/o CLS Align	95.06%	45.37%	67.11%	87.38%
w/o Stop Gradient	95.50%	46.13%	**68.86%**	**88.16%**
w/o MLP Head	95.03%	45.65%	68.34%	87.67%

Table 5: Ablation study Results. Removing the MLM will make BootTOD fail to converge, so we do not report this result.

ments were randomly sampled at least three times with different seeds.

Intent Recognition Table 1 displays the results of intent recognition on the OOS dataset (Larson et al., 2019). We observe that BootTOD outperforms all the baselines on 10 of 12 metrics, particularly with significant improvements in overall accuracy and OOD metrics. These results demonstrate the generalization ability of BootTOD across both in-domain and out-of-domain metrics.

Dialogue State Tracking Table 2 shows the results of dialogue state tracking on MWOZ 2.1. Our BootTOD achieves state-of-the-art results on all the metrics. We find SimCSE performs poorly in the 5% data setting because it ignores the intrinsic properties of dialogue data and can not model overall dialogue well with few data. Our method achieves a greater improvement on joint accuracy than on slot accuracy, indicating the strength of understanding the overall dialogue context. We also find that these baselines overfit to the easy slot values, but can't predict the hard ones, resulting in comparable slot accuracy but poor joint accuracy. For example, BootTOD outperforms TOD-BERT by 0.3% on Slot Acc but 2.7% on Joint Acc in the full data setting, which indicates the superiority of dialogue modeling.

Dialogue Act Prediction Table 3 shows the results of dialogue act prediction on MWOZ and DSTC2. Our BootTOD 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.

Response Selection Table 4 displays the results of response selection on MWOZ and DSTC2.[2] Our BootTOD achieves state-of-the-art results on all the metrics. We find DSE performs poorly in the 1% data setting and even worse than BERT on DSTC2. It shows the assumption that consecutive utterances represent similar semantics fails in practical dialogue scenarios. Although TOD-BERT is pre-trained with a response contrastive objective, our method still outperforms it on DSTC2 significantly both in full data and few data settings. It indicates that BootTOD can achieve better generalization capability.

4. Qualitative Analysis

4.1. Ablation Study

Table 5 presents the ablation results of dialogue act prediction on DSTC2 and response selection on MWOZ. BootTOD without CLS Align performs the worst among all the variations. This indicates that CLS alignment loss is crucial for capturing the dialogue-level information, allowing the dialogue model to have better representation capabilities. Removing MLP Head also damages the performance. We find that removing MLP head makes

[2]TOD-BERT uses the response contrastive loss as the pre-training objective on full MWOZ training data so we don't report its results on few-shot setting.

the training unstable and adding a predictor serves as a decoder to learn future representation. Mask Align also contributes to performance, illustrating the importance of learning fine-grained token representations. Besides, the Stop gradient has a positive impact on dialogue act prediction but a negative impact on response selection. We believe it is due to the mismatch between the stop-gradient and the dual-encoder used in the response selection task.

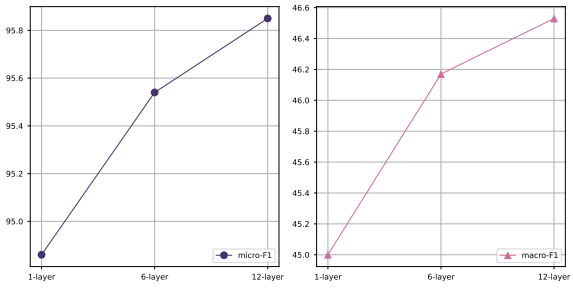
4.2. Hyper-parameter Analysis

Effect of Alignment Layers BootTOD uses the top-K Layer Representation for alignment loss L_{cls} and L_{mask}. Figure 3 shows the effect of varying the value of K. We find BootTOD gets improvements as the value of K increases. It indicates that different layers of the model can capture features of different granularities, thereby improving the performance of the downstream tasks.

Effect of Max Response Length The response consists of consecutive utterances, and we set the number of selectable utterances from 1 to max response length P. To explore the effect of varying the value of P, we set the P to 0, 3, All, and Fix respectively. $P = All$ denotes that we can randomly select any length of utterances from the whole utterances, while $P = Fix$ denotes that we must use the whole consecutive future utterances together. For example, if we have 5 future utterances $F = \{S_t, U_{t+1}, S_{t+1}, U_{t+2}, S_{t+2}\}$. $P = 3$ allows us to select any length no longer than 3, such as $\{S_t\}$ or $\{S_t, U_{t+1}, S_{t+1}\}$; $P = All$ allows us to select any length of future from the 5 utterances, that is $\{S_t\}$ or $\{S_t, U_{t+1}, S_{t+1}\}$ or F; $P = Fix$ can only select F. Figure 4 shows BootTOD generally gets improvements with increasing the P, indicating that more response targets are beneficial to learn more diverse dialogue representations. We also find that $P = Fix$ degrades performance compared to $P = All$. We argue that fixed response information will narrow down dialogue context representation space.

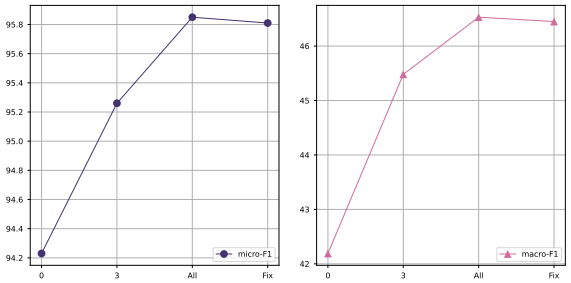
5. Non-Contrastive Methods Comparison

As supervised methods rely on labeled NLI datasets (Williams et al., 2018; Welleck et al., 2019) or dialogue act labels (He et al., 2022b), we didn't include them in a fairness comparison. Instead, we compared BootTOD with a recent non-contrastive method, FutureTOD (Zeng et al., 2023). FutureTOD proposes a non-contrastive framework that distills future knowledge into the representation of the previous dialogue. The results are displayed in Table 6, Table 7, Table 8, and Table 9 in the Appendix. Our method has demonstrated excellent performance on most metrics across all tasks compared to FutureTOD. This underscores the improvement of our BootTOD's performance in comparison to other non-contrastive methods.



(a) micro-F1 (b) macro-F1

Figure 3: Ablation study of Alignment Layers. We report the results of dialogue act prediction on DSTC2. The X-asix and Y-asix denotes the number of layers used for alignment and performance.



(a) micro-F1 (b) macro-F1

Figure 4: Ablation study of max future length P. We report the results of dialogue act prediction on DSTC2. The X-asix and Y-asix denotes the max future length P and performance.

6. Conclusion

In this paper, we propose a novel dialogue pre-training model, BootTOD, which learns task-oriented dialogue representations via a self-bootstrapping framework. Instead of contrastive counterparts, BootTOD aligns context and context+response representations and dismisses the requirements of contrastive pairs. Besides, BootTOD aligns the context representation with diverse targets to model the intrinsic one-to-many diversity of human conversations. We perform comprehensive experiments on various task-oriented dialogue tasks. BootTOD significantly outperforms TOD-BERT, DSE, and other strong baselines in all the scenarios.

Acknowledgements

We thank all anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Natural Science Foundation of China (NSFC No.62076031 and No.62076036). This work is partially supported by State Key Laboratory of Massive Personalized Customization System and Technology (No. H&C-MPC-2023-02-07(Q)).

7. Bibliographical References

Harold Abelson, Gerald Jay Sussman, and Julie Sussman. 1985. *Structure and Interpretation of Computer Programs*. MIT Press, Cambridge, Massachusetts.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

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*.

Siqi Bao, H. He, Fan Wang, and Hua Wu. 2019. Plato: Pre-trained dialogue generation model with discrete latent variable. In *Annual Meeting of the Association for Computational Linguistics*.

Robert Baumgartner, Georg Gottlob, and Sergio Flesca. 2001. Visual information extraction with Lixto. In *Proceedings of the 27th International Conference on Very Large Databases*, pages 119–128, Rome, Italy. Morgan Kaufmann.

Ronald J. Brachman and James G. Schmolze. 1985. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9(2):171–216.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

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*.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

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.

Xinlei Chen and Kaiming He. 2020. Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

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.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

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.

Georg Gottlob. 1992. Complexity results for nonmonotonic logics. *Journal of Logic and Computation*, 2(3):397–425.

Georg Gottlob, Nicola Leone, and Francesco Scarcello. 2002. Hypertree decompositions and tractable queries. *Journal of Computer and System Sciences*, 64(3):579–627.

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.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

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.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

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.

IJCAI Proceedings. IJCAI camera ready submission. https://proceedings.ijcai.org/info.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

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.

Hector J. Levesque. 1984a. Foundations of a functional approach to knowledge representation. *Artificial Intelligence*, 23(2):155–212.

Hector J. Levesque. 1984b. A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, pages 198–202, Austin, Texas. American Association for Artificial Intelligence.

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.

Shikib Mehri, Mihail Eric, and Dilek Z. Hakkani-Tür. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *ArXiv*, abs/2009.13570.

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*.

Bernhard Nebel. 2000. On the compilability and expressive power of propositional planning formalisms. *Journal of Artificial Intelligence Research*, 12:271–315.

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*.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

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*.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

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.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

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.

Sean Welleck, Jason Weston, Arthur D. Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *ACL*.

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*.

Weihao Zeng, Keqing He, Yejie Wang, Chen Zeng, Jingang Wang, Yunsen Xian, and Weiran Xu. 2023. FutureTOD: Teaching future knowledge to pre-trained language model for task-oriented dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6532–6546, Toronto, Canada. Association for Computational Linguistics.

Weihao Zeng, Keqing He, Zechen Wang, Dayuan Fu, Guanting Dong, Ruotong Geng, Pei Wang, Jingang Wang, Chaobo Sun, Wei Wu, and Weiran Xu. 2022. Semi-supervised knowledge-grounded pre-training for task-oriented dialog systems. In *Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD)*, pages 39–47, Abu Dhabi, Beijing (Hybrid). Association for Computational Linguistics.

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. Evaluation Details

We evaluated various pre-trained language models on four core task-oriented dialogue tasks, including intent recognition, dialogue state tracking, dialogue act prediction, and response selection. Here, we provide more details about these evaluation tasks and metrics.

Intent Recognition is a multi-class classification task that takes a dialogue utterance as input and predicts an intent label (Zeng et al., 2022). We use the [CLS] embeddings from the model as the dialogue representation. The model is trained with cross-entropy loss. We report classification accuracy and recall.

Dialogue State Tracking is a multi-class classification task, which involves identifying the slot values for each (domain, slot) pair at each dialogue turn, based on a pre-defined ontology. The model takes dialogue history as input and 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 the Joint acc and Slot acc. The Joint acc considers true if and only if the predicted values exactly match its ground truth values at each dialogue turn. The slot acc individually compares each (domain, slot, value) triplet to its ground truth label.

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

Response Selection is a ranking task that aims to retrieve the most relative system response from a candidate pool based on dialogue history. We also use MWOZ and DSTC2 as our evaluation datasets. We use a dual-encoder strategy, which calculates similarity scores between dialogue history and candidate responses. We train this model with random system responses from the corpus as negative samples. We report k-to-100 accuracy. This metric represents the ratio of the ground-truth response being ranked in the top-k positions when compared to 99 randomly sampled responses, as determined by the scores computed by the dual-encoder.

B. Non-Contrastive Methods Comparison

We present the performance of non-contrastive methods in intent recognition, dialogue state tracking, dialogue act prediction, and response selection in the Table 6, Table 7, Table 8 and Table 9 respectively.

	Acc(all)	Acc(in)	Acc(out)	Recall(out)
FutureTOD	87.2%	96.0%	90.0%	47.6%
BootTOD	**88.2%**	**96.1%**	**91.1%**	**52.7%**

Table 6: The performance of non-contrastive methods on the OOS dataset for Intent recognition. Acc(all), Acc(in), Acc(out) denotes the overall accuracy, in-domain intent accuracy, and out-of-domain intent accuracy.

	Joint Acc	Slot Acc
FutureTOD	50.4%	97.1%
BootTOD	**50.7**%	**97.2**%

Table 7: The performance of non-contrastive methods on the MWOZ 2.1 for Dialogue State Tracking. Joint Acc and Slot Acc denote the joint goal accuracy and slot accuracy.

	MWOZ		DSTC2	
	micro-F1	macro-F1	micro-F1	macro-F1
FutureTOD	**92.0**%	81.9%	94.6%	44.6%
BootTOD	91.8%	**82.3**%	**95.9**%	**46.5**%

Table 8: The performance of non-contrastive methods on the MWOZ and DSTC2 for Dialogue Act Prediction.

	MWOZ		DSTC2	
	1-to-100	3-to-100	1-to-100	3-to-100
FutureTOD	68.5%	87.9%	58.4%	72.6%
BootTOD	**68.8**%	87.6%	**59.1**%	72.3%

Table 9: The performance of non-contrastive methods on the MWOZ and DSTC2 for Response Selection. 1-to-100 and 3-to-100 denote the ratio of the ground-truth response being ranked at the top-1 or top-3 given 100 candidates.