



北京邮电大学

Beijing University of Posts and Telecommunications

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

Weihao Zeng, Keqing He, Yejie Wang,
Dayuan Fu, Weiran Xu

¹ Beijing University of Posts and Telecommunications

² Meituan, Beijing, China

ZengWH@bupt.edu.cn



Content



- 1 Motivations
 - 2 Contributions
 - 3 Methodology
 - 4 Experimental Setup
 - 5 Qualitative Analysis
 - 6 Challenges
 - 7 Conclusion
- 

- Motivations

1. Contrastive methods suffer from selecting noisy positive and negative pairs
2. Ignoring the one-to-many property in conversation

- Contributions

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

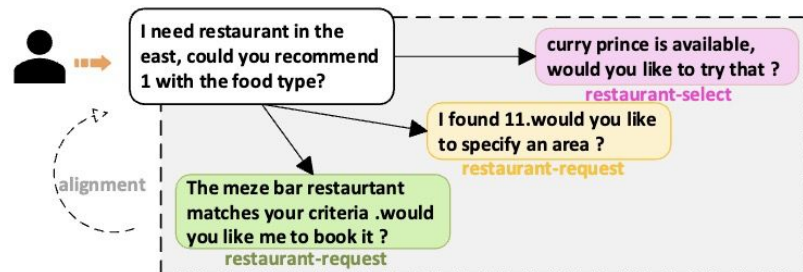


Figure 1: The same context may have multiple appropriate responses in a task-oriented dialogue.

Alignment objectives:

1. Dialogue Representation Alignment Loss
2. Token Representation Alignment Loss
3. Mask Language Model Loss

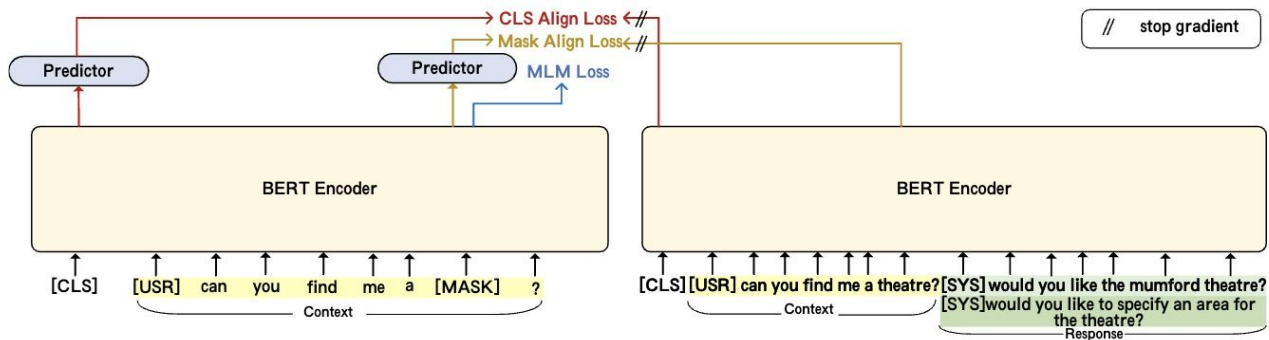


Figure 2: Overall architecture of our proposed BootTOD.

- Main Results

Intent recognition

	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 $p < 0.01$.

Dialogue state tracking

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 $p < 0.01$.

- Main Results

Dialogue Act Prediction

	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%
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%
	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 $p < 0.01$.

Response Selection

	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%
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%
	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 $p < 0.01$.

• Ablation Study

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.

Effect of Max Response Length

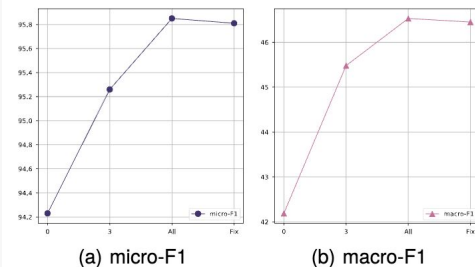


Figure 4: Ablation study of max future length P . We report the results of dialogue act prediction on DSTC2. The X-axis and Y-axis denotes the max future length P and performance.

Hyper-parameter Analysis

Effect of Alignment Layers

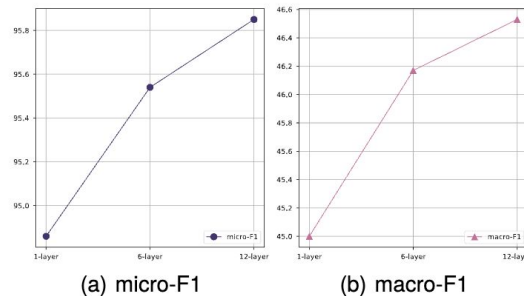


Figure 3: Ablation study of Alignment Layers. We report the results of dialogue act prediction on DSTC2. The X-axis and Y-axis denotes the number of layers used for alignment and performance.

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 con-text+response representations and dismisses the requirements of contrastive pairs. Besides, Boot-TOD 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.

Thank You

THANKS!

