Code: https://github.com/jasonwu0731/ToD-BERT

TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue

Chien-Sheng (Jason) Wu, Steven Hoi,
Richard Socher, Caiming Xiong

Salesforce AI Research
EMNLP 2020

Dialogue Systems:
Chit-Chat (CC) v.s. Task-Oriented Dialogue (TOD)

Chit-Chat Dialogue Systems

- No Specific goal
- Focus on generating natural responses
- The more turns the better
- Using variants of generation models, ex: Seq2Seq, VAE, etc.

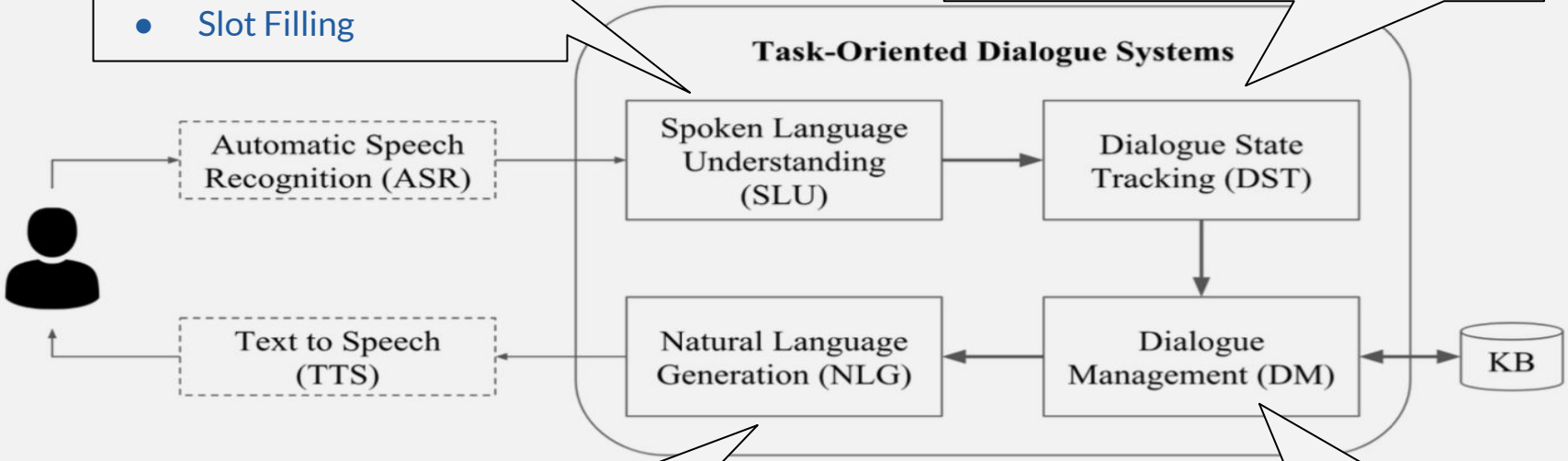
Task-Oriented Dialogue Systems

- Help users achieve their goal
- Focus on understanding users, tracking states, and generating next actions.
- The less turns the better
- Combination of rules and statistical components.

Modularized TOD Systems

- Domain Identification
- Intent Detection
- Slot Filling

- Represent system's belief of user's goal as slot-value pairs



Figu … arized task-ori

- Mapping actions and states into natural language.

- Query knowledge base
- Dialogue policy (generate next action)

Pre-training for Task-oriented Dialogues

Motivation & Challenge

Motivation

- BERT and its variances do **NOT** show much improvement in conversational applications as observed in other NLP tasks.
- One possible reason could be the **intrinsically difference** of linguistic patterns between human conversations and writing text, resulting in a large gap of data distributions.

Challenge

- Dataset?
  - Unlike open-domain corpora, which are easy-to-get from social media (Twitter or Reddit), task-oriented corpora are **small** in size, **expensive** to collect, and **scattered** in different sources.
- Objective functions?

Research Questions

- Can we leverage existing TOD datasets to improve natural language understanding?
- Will pre-training on TOD datasets learn better representations for downstream tasks?
- How to do the pre-training? Can it really transfer knowledge effectively?

Aggregated Pre-training Datasets

We collect **nine** different task-oriented datasets which are English, human-human, and multi-turn. In total, there are 100,707 dialogues, which contain 1,388,152 utterances over 60 domains. Dataset statistics is shown in Table 1.

Name	# Dialogue	# Utterance	Avg. Turn	# Domain
MetaLWOZ (Lee et al., 2019)	37,884	432,036	11.4	47
Schema (Rastogi et al., 2019)	22,825	463,284	20.3	17
Taskmaster (Byrne et al., 2019)	13,215	303,066	22.9	6
MWOZ (Budzianowski et al., 2018)	10,420	71,410	6.9	7
MSR-E2E (Li et al., 2018)	10,087	74,686	7.4	3
SMD (Eric and Manning, 2017)	3,031	15,928	5.3	3
Frames (Asri et al., 2017)	1,369	19,986	14.6	3
WOZ (Mrkšić et al., 2016)	1,200	5,012	4.2	1
CamRest676 (Wen et al., 2016)	676	2,744	4.1	1

Table 1: Data statistics for task-oriented dialogue datasets.

P.S. MWOZ test set is excluded from the pre-training step since we use it as one of the evaluated datasets for downstream tasks.

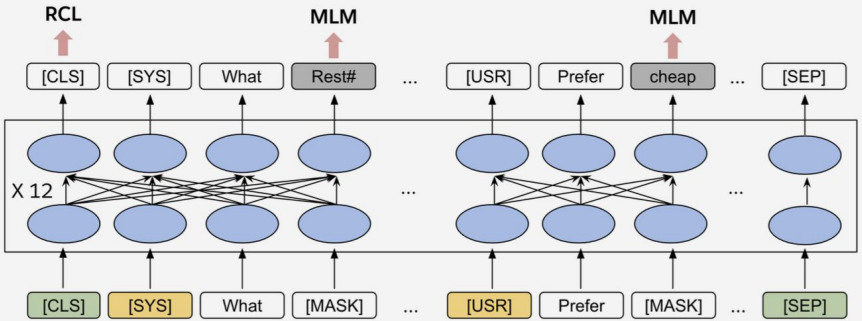
Pre-training Objectives

MLM & RCL

- Masked-Language Modeling (MLM)
  - Randomly select a few tokens in the input and replace with the special token [MASK].
  - M is the total number of masked tokens and P(x_m) is the predicted probability of the token x_m.
- Response Contrastive Loss (RCL)
  - We randomly split a dialogue into two pieces (X and Y), and use the same encoder to encode each of them.
  - We do in-batch negative training with cross-entropy loss. The number of negative samples increases as the batch size increases.

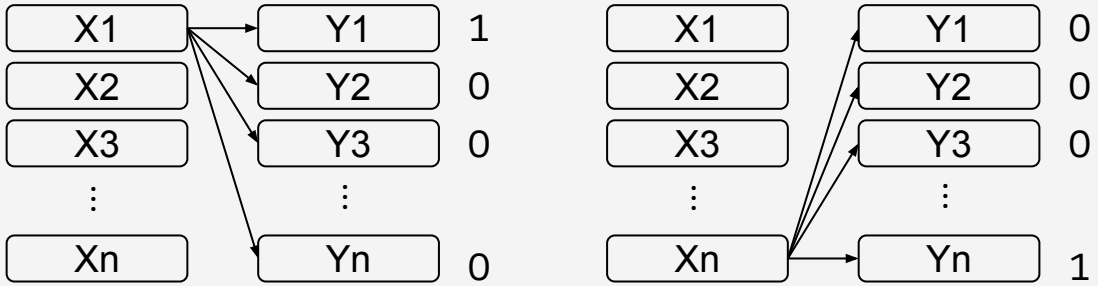
$$L_{mlm} = -\sum_{m=1}^{M} \log P(x_m)$$

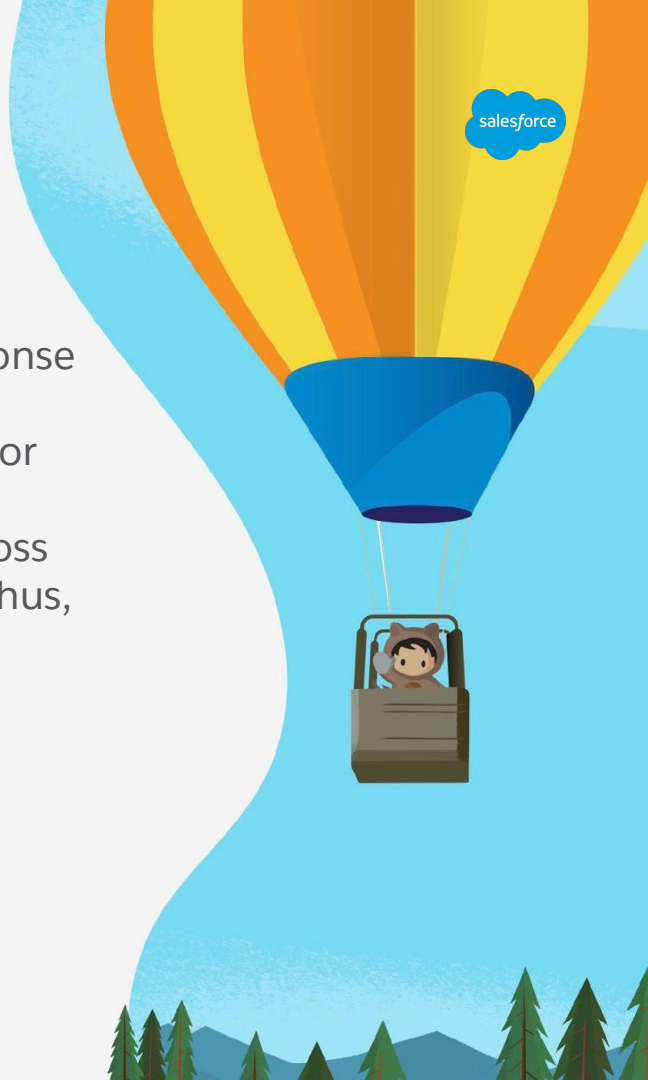
$$L_{rcl} = -\sum_{i=1}^{b} \log M_{i,i},$$
$$M = \text{Softmax}(CR^T) \in \mathbb{R}^{b \times b}.$$



Response Contrastive Loss

- It does not require any additional human annotation.
- Intuitively, we encourage the model to capture underlying dialogue sequential order, structure information, and response similarity.
- In-batch negative training is data/computational efficient for contrastive learning.
- We use cross-entropy loss instead of binary cross-entropy loss because negative responses may not be “true negatives”. Thus, we treat the task as a ranking task instead of a multi-label classification task.





Pre-training Details

- We select BERT because
  - it is the most widely used model in NLP research recently.
  - We use BERT-base model with 12 layers with hidden size = 768.

- To capture speaker information and the underlying interaction behavior in dialogue
  - We add two special tokens, [USR] and [SYS], to the byte-pair embeddings.
  - We prefix the special token to each user utterance and system response
  - Concatenate all the utterances in the same dialogue into one flat sequence.

- Our unified datasets and training objectives can be easily applied to pre-train any existing language models.

- Model Variances:
  - TOD-BERT-mlm: model only trained on the MLM objective
  - TOD-BERT-jnt: model trained on both MLM and RCL objectives

Before Fine-tuning: Linear Probe

- We probe the output representation using **one single-layer perceptron** on top of a "**fixed**" pre-trained language model and only fine-tune that layer for a downstream task with the same number of hyper-parameters.

	Domain (acc)	Intent (acc)	Dialogue Act (F1-micro)
GPT2	63.5%	74.7%	85.7%
DialoGPT	63.0%	65.7%	84.2%
BERT	60.5%	71.1%	85.3%
TOD-BERT-mlm	63.9%	70.7%	83.5%
TOD-BERT-jnt	**68.7%**	**77.8%**	**86.2%**

Table 3: Probing results of different pre-trained language models using a single-layer perceptron.

"Probing methods are proposed to determine what information is carried intrinsically by the learned embeddings"

Tenney et al., 2019

Fine-tuning & Evaluation

- We care the most in this paper whether TOD-BERT, a pre-trained language model using aggregated TOD corpora, can show any advantage over BERT.

- We avoid adding too many additional components on top of its architecture when fine-tuning on each downstream task.

- We select four crucial task-oriented downstream tasks to evaluate: **intent recognition, dialogue state tracking, dialogue act prediction**, and **response selection**.

- We conduct experiments on
  - the full dataset using all the training samples
  - Partial datasets using a few training samples (reported results over 3 runs)

Intent Detection

- Intent task is a **multi-class classification** problem
- The OOS intent dataset has 15,100/3,100/5,500 samples for the train, validation, and test sets, respectively. It covers **151 intent classes** over **ten domains**, including 150 in-scope intent and one out- of-scope intent. The out-of-scope intent means that a user utterance that does not fall into any of the predefined intents. Each of the intents has 100 training samples.

$$P_{int} = \text{Softmax}(W_1(F(U))) \in \mathbb{R}^I$$

"An evaluation dataset for intent classification and out-of-scope prediction", Larson et al., EMNLP 2019.

	Model	Acc (all)	Acc (in)	Acc (out)	Recall (out)
1-Shot	BERT	29.3% ± 3.4%	35.7% ± 4.1%	81.3% ± 0.4%	0.4% ± 0.3%
	TOD-BERT-mlm	38.9% ± 6.3%	47.4% ± 7.6%	81.6% ± 0.2%	**0.5%** ± 0.2%
	TOD-BERT-jnt	**42.5%** ± 0.1%	**52.0%** ± 0.1%	**81.7%** ± 0.1%	0.1% ± 0.1%
10-Shot	BERT	75.5% ± 1.1%	88.6% ± 1.1%	84.7% ± 0.3%	16.5% ± 1.7%
	TOD-BERT-mlm	76.6% ± 0.8%	90.5% ± 1.2%	84.3% ± 0.2%	14.0% ± 1.3%
	TOD-BERT-jnt	**77.3%** ± 0.5%	**91.0%** ± 0.5%	**84.5%** ± 0.4%	**15.3%** ± 2.1%
Full (100-Shot)	FastText*	-	89.0%	-	9.7%
	SVM*	-	91.0%	-	14.5%
	CNN*	-	91.2%	-	18.9%
	GPT2	83.0%	94.1%	87.7%	32.0%
	DialoGPT	83.9%	95.5%	87.6%	32.1%
	BERT	84.9%	95.8%	88.1%	35.6%
	TOD-BERT-mlm	85.9%	96.1%	89.5%	**46.3%**
	TOD-BERT-jnt	**86.6%**	**96.2%**	**89.9%**	43.6%

Table 2: Intent recognition results on the OOS dataset, one of the largest intent corpus. Models with * are reported from Larson et al. (2019).

Dialogue State Tracking (DST)

- We treat DST as a **multiple multi-class classification** problem using a predefined ontology. The model is trained with cross-entropy loss summed over all the pairs.
- We compare BERT to TOD-BERT-jnt on the MWOZ 2.1 dataset and find the latter has 2.4% joint goal accuracy improvement.
- The improvement is more obvious in limited data setting, where 1% of data contains around 84 dialogues.

$$S_i^j = Sim(G_j(F(X)), F(v_i^j)) \in \mathbb{R}^1$$

	Model	Joint Acc	Slot Acc
1% Data	BERT	6.4% ± 1.4%	84.4% ± 1.0%
	TOD-BERT-mlm	**9.9%** ± 0.6%	**86.6%** ± 0.5%
	TOD-BERT-jnt	8.0% ± 1.0%	85.3% ± 0.4%
5% Data	BERT	19.6% ± 0.1%	92.0% ± 0.5%
	TOD-BERT-mlm	28.1% ± 1.6%	**93.9%** ± 0.1%
	TOD-BERT-jnt	**28.6%** ± 1.4%	93.8% ± 0.3%
10% Data	BERT	32.9% ± 0.6%	94.7% ± 0.1%
	TOD-BERT-mlm	**39.5%** ± 0.7%	**95.6%** ± 0.1%
	TOD-BERT-jnt	37.0% ± 0.1%	95.2% ± 0.1%
25% Data	BERT	40.8% ± 1.0%	95.8% ± 0.1%
	TOD-BERT-mlm	44.0% ± 0.4%	**96.4%** ± 0.1%
	TOD-BERT-jnt	**44.3%** ± 0.3%	96.3% ± 0.2%
Full Data	DSTReader*	36.4%	-
	HyST*	38.1%	-
	ZSDST*	43.4%	-
	TRADE*	45.6%	-
	GPT2	46.2%	96.6%
	DialoGPT	45.2%	96.5%
	BERT	45.6%	96.6%
	TOD-BERT-mlm	47.7%	96.8%
	TOD-BERT-jnt	**48.0%**	**96.9%**

Table 5: Dialogue state tracking results on MWOZ 2.1. We report joint goal accuracy and slot accuracy for the full data setting and the simulated few-shot settings.

Dialogue Act (DA) Prediction

- DA is a **multi-label classification** problem because a system response may contain multiple dialogue acts.
- The model is trained with binary cross-entropy loss and the i-th dialogue act is considered as a triggered dialogue act if A_i > 0.5.

$$A = Sigmoid(W_2(F(X))) \in \mathbb{R}^N$$

		MWOZ (13)		DSTC2 (9)		GSIM (6)	
		micro-F1	macro-F1	micro-F1	macro-F1	micro-F1	macro-F1
1% Data	BERT	84.0% ± 0.6%	66.7% ± 1.7%	77.1% ± 2.1%	25.8% ± 0.8%	67.3% ± 1.4%	26.9% ± 1.0%
	TOD-BERT-mlm	**87.5%** ± 0.6%	**73.3%** ± 1.5%	79.6% ± 1.0%	26.4% ± 0.5%	**82.7%** ± 0.7%	**35.7%** ± 0.3%
	TOD-BERT-jnt	86.9% ± 0.2%	72.4% ± 0.8%	**82.9%** ± 0.4%	**28.0%** ± 0.1%	78.4% ± 3.2%	32.9% ± 2.1%
10% Data	BERT	89.7% ± 0.2%	78.4% ± 0.3%	88.2% ± 0.7%	34.8% ± 1.3%	98.4% ± 0.3%	45.1% ± 0.2%
	TOD-BERT-mlm	90.1% ± 0.2%	78.9% ± 0.1%	**91.8%** ± 1.7%	**39.4%** ± 1.7%	99.2% ± 0.1%	45.6% ± 0.1%
	TOD-BERT-jnt	**90.2%** ± 0.2%	**79.6%** ± 0.7%	90.6% ± 3.2%	38.8% ± 2.2%	**99.3%** ± 0.1%	**45.7%** ± 0.0%
Full Data	MLP	61.6%	45.5%	77.6%	18.1%	89.5%	26.1%
	RNN	90.4%	77.3%	90.8%	29.4%	98.4%	45.2%
	GPT2	90.8%	79.8%	92.5%	39.4%	99.1%	45.6%
	DialoGPT	91.2%	79.7%	**93.8%**	**42.1%**	99.2%	45.6%
	BERT	91.4%	79.7%	92.3%	40.1%	98.7%	45.2%
	TOD-BERT-mlm	**91.7%**	79.9%	90.9%	39.9%	99.4%	45.8%
	TOD-BERT-jnt	**91.7%**	**80.6%**	**93.8%**	41.3%	**99.5%**	**45.8%**

Table 4: Dialogue act prediction results on three different datasets. The numbers reported are the micro and macro F1 scores, and each dataset has different numbers of dialogue acts.

Response Selection (RS)

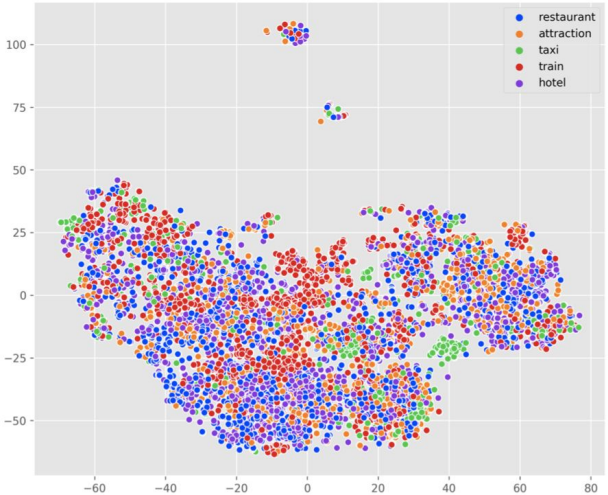
- RS is a **ranking problem** between source X and target Y. We use **dual-encoder** as ranker.
- Source X can be truncated, and we limit the context lengths to the most recent 256 tokens in our experiments.

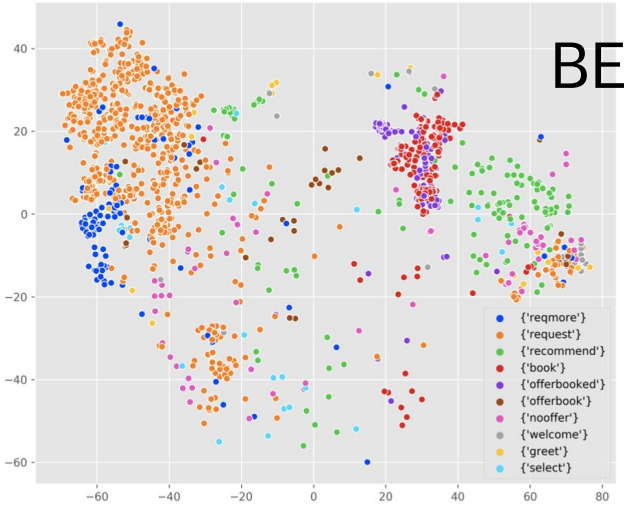
$$r_i = Sim(F(X), F(Y_i)) \in \mathbb{R}^1$$

		MWOZ		DSTC2		GSIM	
		1-to-100	3-to-100	1-to-100	3-to-100	1-to-100	3-to-100
1% Data	BERT	7.8% ± 2.0%	20.5% ± 4.4%	3.7% ± 0.6%	9.6% ± 1.3%	4.0% ± 0.4%	10.3% ± 1.1%
	TOD-BERT-mlm	13.0% ± 1.1%	34.6% ± 0.4%	12.5% ± 6.7%	24.9% ± 10.7%	7.2% ± 4.0%	15.4% ± 8.0%
	TOD-BERT-jnt	-	-	**37.5%** ± 0.6%	**55.9%** ± 0.4%	**12.5%** ± 0.9%	**26.8%** ± 0.8%
10% Data	BERT	20.9% ± 2.6%	45.4% ± 3.8%	8.9% ± 2.3%	21.4% ± 3.1%	9.8% ± 0.1%	24.4% ± 1.2%
	TOD-BERT-mlm	22.3% ± 3.2%	48.7% ± 4.0%	19.0% ± 16.3%	33.8% ± 20.4%	11.2% ± 2.5%	26.0% ± 2.7%
	TOD-BERT-jnt	-	-	**49.7%** ± 0.3%	**66.6%** ± 0.1%	**23.0%** ± 1.0%	**42.6%** ± 1.0%
Full Data	GPT2	47.5%	75.4%	53.7%	69.2%	39.1%	60.5%
	DialoGPT	35.7%	64.1%	39.8%	57.1%	16.5%	39.5%
	BERT	47.5%	75.5%	46.6%	62.1%	13.4%	32.9%
	TOD-BERT-mlm	48.1%	74.3%	50.0%	65.1%	36.5%	60.1%
	TOD-BERT-jnt	**65.8%**	**87.0%**	**56.8%**	**70.6%**	**41.0%**	**65.4%**

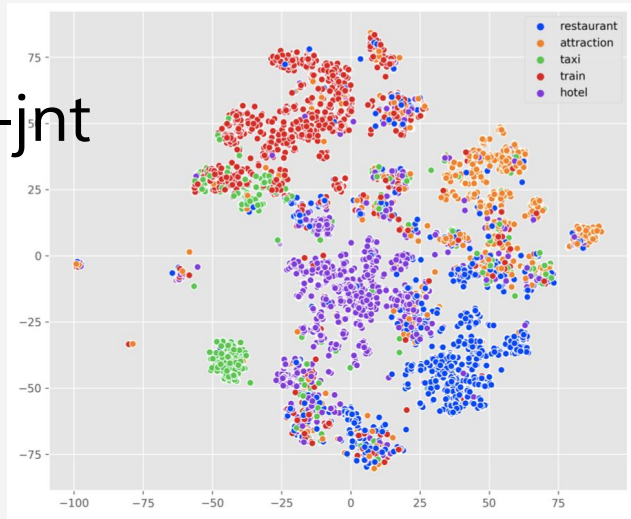
Table 6: Response selection evaluation results on three corpora for 1%, 10% and full data setting. We report 1-to-100 and 3-to-100 accuracy, which is similar to recall1 and recall@3 given 100 candidates.

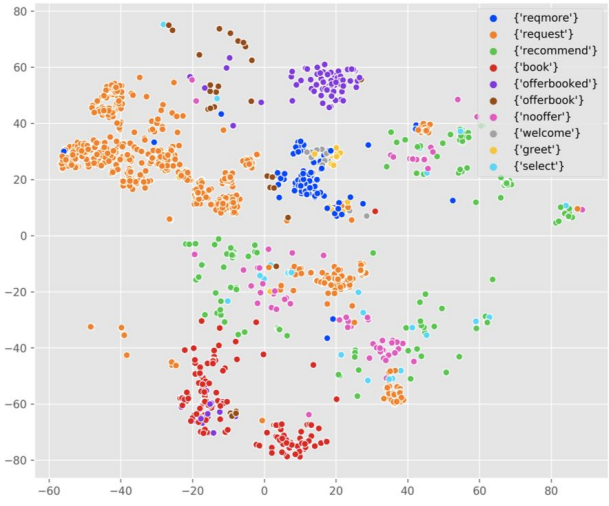
BERT





TOD-BERT-jnt





Code: https://github.com/jasonwu0731/ToD-BERT

- TODBERT/TOD-BERT-MLM-V1: TOD-BERT pre-trained only using the MLM objective
- TODBERT/TOD-BERT-JNT-V1: TOD-BERT pre-trained using both the MLM and RCL objectives
- TODBERT/TOD-DistilBERT-JNT-V1: TOD-DistilBERT pre-trained using both the MLM and RCL objectives

```
import torch
from transformers import *
tokenizer = AutoTokenizer.from_pretrained("TODBERT/TOD-BERT-JNT-V1")
tod_bert = AutoModel.from_pretrained("TODBERT/TOD-BERT-JNT-V1")
```

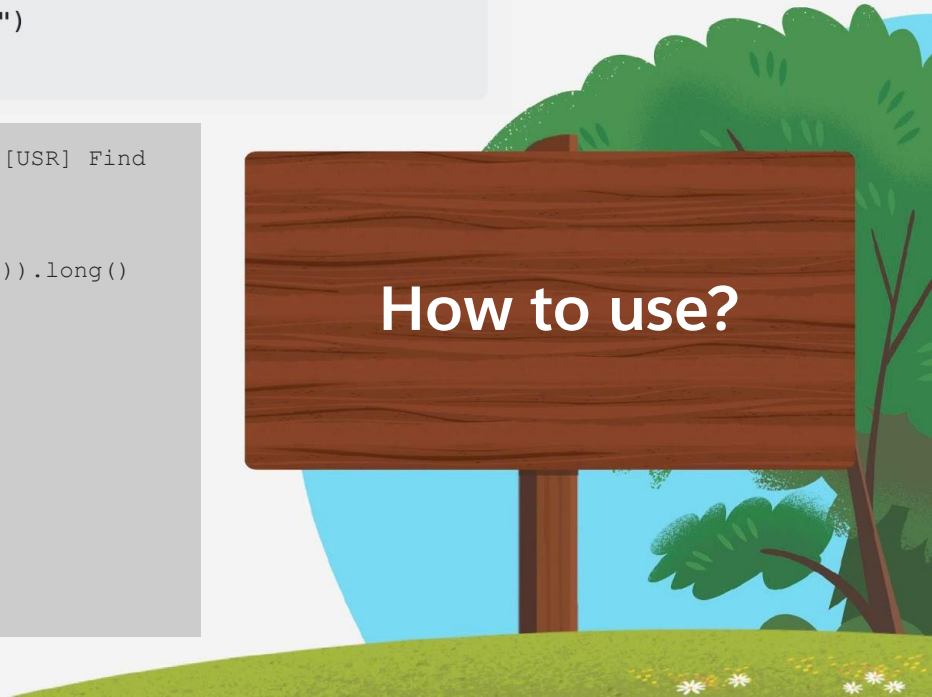
```
input_text = "[CLS] [SYS] Hello, what can I help with you today? [USR] Find
me a cheap restaurant nearby the north town."

input_tokens = tokenizer.tokenize(input_text)
story = torch.Tensor(tokenizer.convert_tokens_to_ids(input_tokens)).long()

if len(story.size()) == 1:
    story = story.unsqueeze(0) # batch size dimension

if torch.cuda.is_available():
    tod_bert = tod_bert.cuda()
    story = story.cuda()

with torch.no_grad():
    input_context = {"input_ids": story,
                     "attention_mask": (story > 0).long()}
    hiddens = tod_bert(**input_context)[0]
```



Conclusion

1. TOD-BERT improves TOD downstream tasks, especially for few-shot scenarios.
2. Unsupervised pre-training objectives on TOD corpora, especially the response contrastive objective, is beneficial in dialogue pre-training.
3. TOD-BERT can be easily plugged in to any state-of-the-art models, and the training strategy can be used to pre-train any pre-trained LMs.

If you have any questions, please feel free to contact Jason by email.
(wu.jason@salesforce.com)