GALAXY: A Generative Pre-trained Model for Task-Oriented Dialog with Semi-supervised Learning and Explicit Policy Injection

Wanwei He[1,2*†], Yinpei Dai[3*], Yinhe Zheng[3], Yuchuan Wu[3], Zheng Cao[3], Dermot Liu[3]
Peng Jiang[3], Min Yang[1‡], Fei Huang[3], Luo Si[3], Jian Sun[3], Yongbin Li[3‡]
[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
[2]University of Chinese Academy of Sciences, China
[3]Alibaba Group, China
{ww.he, min.yang}@siat.ac.cn, {yinpei.dyp, shuide.lyb, f.huang, luo.si, jian.sun}@alibaba-inc.com

Abstract

Pre-trained models have proved to be powerful in enhancing task-oriented dialog systems. However, current pre-training methods mainly focus on enhancing dialog understanding and generation tasks while neglecting the exploitation of dialog policy. In this paper, we propose GALAXY, a novel pre-trained dialog model that explicitly learns dialog policy from limited labeled dialogs and large-scale unlabeled dialog corpora via semi-supervised learning. Specifically, we introduce a dialog act prediction task for policy optimization during pre-training and employ a *consistency regularization* term to refine the learned representation with the help of unlabeled dialogs. We also implement a gating mechanism to weigh suitable unlabeled dialog samples. Empirical results show that GALAXY substantially improves the performance of task-oriented dialog systems, and achieves new state-of-the-art results on benchmark datasets: In-Car, MultiWOZ2.0 and MultiWOZ2.1, improving their end-to-end combined scores by 2.5, 5.3 and 5.5 points, respectively. We also show that GALAXY has a stronger few-shot ability than existing models under various low-resource settings. For reproducibility, we release the code and data at https://github.com/siat-nlp/GALAXY.

Introduction

Task-oriented dialog (TOD) systems aim to help users accomplish certain tasks through conversations. Fundamental abilities of a TOD system include: (1) *Dialog understanding*: extracting structured semantics from user utterances; (2) *Policy planning*: determining a Dialog Act (DA) that leads to successful task completion; and (3) *Dialog generation*: producing appropriate responses (Figure 1). With the recent progress of Pre-trained Language Models (PLMs), remarkable performances improvements are achieved by casting TODs as generative language modeling tasks (Peng et al. 2020a; Lin et al. 2020), which benefit from the rich linguistic knowledge embedded in PLMs.

[*]Equal Contribution.
[†]Work done while the author was interning at Alibaba.
[‡]Corresponding authors.

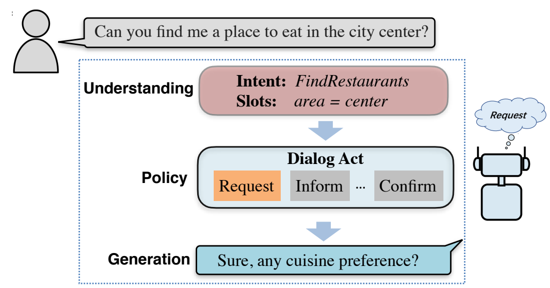

Figure 1: Given the input user utterance, a task-oriented dialog system needs to perform understanding, policy planning, and generation successively to complete the reply.

However, as reported in previous studies (Zhang et al. 2020b; Kulhánek et al. 2021), there are intrinsic differences between the distribution of human conversations and plain texts. Directly fine-tuning plain-text-trained PLMs on downstream dialog tasks hinders the model from effectively capturing conversational linguistic knowledge and thus leads to sub-optimal performances (Mehri et al. 2019; Zeng and Nie 2021; Wu and Xiong 2020). Current attempts to tackle this issue try to build Pre-trained Conversation Models (PCMs) by directly optimizing vanilla language model objectives on dialog corpora (Mehri, Eric, and Hakkani-Tur 2020; Zhang et al. 2020b; Henderson et al. 2019), which shows improved results on both dialog understanding (Wu et al. 2020) and generation (Peng et al. 2020b).

Despite these reported advances, few approaches are proposed to further enrich the pre-training process of PCMs with the knowledge of dialog policy. Specifically, existing methods either ignore explicit policy modeling or use latent variables without considering external dialog policy information (Bao et al. 2020), which hinders the possibility of learning controllable policy during pre-training. The optimization of dialog policy is usually formulated as a DA prediction task, which is crucial in TOD systems (Su et al. 2017; Liu et al. 2018). Therefore, we hypothesize that explicitly incorporating the DA annotations into the pre-training process can also facilitate learning better representations for policy optimization to improve the overall end-to-end performance.

A naive way to utilize these labels is to design a multi-task learning process (Sun et al. 2020) that directly combines vanilla unsupervised pre-training losses such as MLM (Devlin et al. 2018) with a supervised DA classification loss. However, this approach has several drawbacks when generalizing to large-scale pre-training paradigms: (1) The DA annotation schema is inconsistent among existing corpora, making it challenging to collect large-scale DA annotations; (2) A vast majority of available dialogs do not have DA labels. A naive joint training process without careful regularization would lead to highly over-fitting on those labeled samples, resulting in low performance; (3) All supervision signals from unlabeled data are self-supervised without any explicit inference over the DA space, so the linguistic knowledge PCMs can extract is only the general type, and the knowledge of dialog policy can not be effectively explored.

In this study, we propose a novel generative pre-trained model called GALAXY, aiming to inject the knowledge of dialog policy explicitly into pre-training at low cost while maintaining its strong ability on dialog understanding and generation. To begin with, we build a unified DA taxonomy for TOD and examine eight existing datasets to develop a new labeled dataset named *UniDA* with a total of 975K utterances. We also collect and process a large-scale unlabeled dialog corpus called *UnDial* with 35M utterances, whose scenarios ranging from online forums to customer services. Then, we propose a semi-supervised pre-training paradigm that applies *consistency regularization* (Verma et al. 2019) on all data. It minimizes the bi-directional KL-divergence between model predictions made on dropout-perturbed samples, which facilitates better representation learning from unlabeled dialog corpora. Since a large proportion of UnDial is from the Internet and not well-suited to our DA taxonomy, we add a learnable control gate on the KL loss of unlabeled data, so that only good samples are allowed for the consistent regularization, other samples are restricted back to normal self-supervised objectives. Experiments show that GALAXY substantially improves TOD systems and achieves new state-of-the-art results on In-Car, MultiWOZ2.0, and MultiWOZ2.1, pushing the end-to-end combined score to 107.45, 110.35, and 110.76, respectively. We also observe that GALAXY has a strong few-shot ability under various low-resource settings.

In summary, our main contributions are three-fold:

- To the best of our knowledge, this is the first study to use semi-supervised pre-training to model explicit dialog policy for PCMs.
- Experiments show our model has learned the knowledge of dialog policy, and achieves new state-of-the-art performance on several TOD benchmarks;
- We collect a new labeled dataset *UniDA* as well as a large-scale unlabeled dialog corpus *UnDial*, hoping that can help bring forward the research in this area.

Related Work

Pre-trained Language Models (PLMs) are trained on large-scale textual corpora with Transformer (Devlin et al. 2018; Radford et al. 2019), which significantly improve dialog systems performance. Budzianowski and Vulić (2019) is the first work to validate the possibility of fine-tuning the information of all sub-tasks in a single paragraph of text on GPT-2. SimpleTOD (Hosseini-Asl et al. 2020) and SOLOIST (Peng et al. 2020a) further generalize this idea to an end-to-end setting where the semantic labels are generated instead of using ground truth values and also consider database results in the training process. Yang, Li, and Quan (2020) leverage the entire dialog session as the input sequence and demonstrate superior performance using self-generated responses during evaluation.

Pre-trained Conversation Models (PCMs) are variants of PLMs particularly adapted for conversational modeling. The main adaptation methods can be roughly divided into three types. The first is training PLMs on dialog corpora instead of plain texts with vanilla language model objectives. Recent work, such as DialoGPT (Zhang et al. 2020b), Meena (Adiwardana et al. 2020) and Blender (Roller et al. 2020) are trained on billions of open-domain dialogs, demonstrating powerful dialog generation performances. TOD-BERT (Wu et al. 2020) shows a great few-shot ability in various understanding tasks via pre-training BERT on extensive task-oriented dialog data. The second line is to design new dialog-oriented pre-training objectives (Bao et al. 2020; He et al. 2020, 2021; Xu and Zhao 2021; Su et al. 2021; Dai et al. 2021). Bao et al. (2020) use discrete latent variables to tackle the one-to-many mapping problem in open-domain dialog generation. Xu and Zhao (2021) propose to simulate the conversation features only using plain texts. The third is to integrate dialog annotations into the pre-training stage. Yu et al. (2020) use labels of dialog understanding as supervision to pre-train BERT. Peng et al. (2020b) use labeled conditional generation data to enhance dialog generation performance. Different from them, we are the first to utilize labels of dialog policy to improve PCMs.

Semi-supervised Learning (SSL) learns from both unlabeled and labeled data. Approaches differ on what information to acquire from the structure of the unlabeled samples. Many initial results were based on generative models, such as variational autoencoders (Kingma and Welling 2019) and generative adversarial networks (Goodfellow et al. 2014). Pseudo-Labeling (Lee et al. 2013) is another widely used method, where unlabeled data is used as further training data after predicted by a model trained on labeled data. One line of recent research shows promising results by jointly training labeled data with supervised learning and unlabeled data with self-supervised learning (Sun et al. 2020). This lies in the paradigm of multi-task learning, where lower layers are often shared across all tasks while the top layers are task-specific. Consistency regularization (Verma et al. 2019) is also a prominent method in SSL, which improves classification performance by minimizing the discrepancy between predictions made on perturbed unlabeled data points. Recently, SimCSE (Gao, Yao, and Chen 2021) leverages dropout as the perturbed method and uses a contrastive objective as the regularization loss to learn sentence representations. Inspired by SimCSE, we adopt the same dropout method for perturbation, and use the bidirectional KL-divergence as in Liang et al. (2021) as our regularization loss, hoping to learn better representations that encodes the knowledge of dialog policy for downstream tasks.

Pre-training Dialog Datasets

In this section, we describe the new dialog datasets used for pre-training, including a labeled dialog dataset (*UniDA*) and a large-scale unlabeled dialog corpus (*UnDial*).

Labeled Dataset: *UniDA*

Dialog policy[1] is tasked to predict dialog acts (DAs) given dialog context. Although DAs are general tags to describe speakers' communicative behaviors (Bunt 2009), current DA annotations in task-oriented dialog are still limited and lack of unified taxonomy because each dataset is small and scattered. Recently, Paul, Goel, and Hakkani-Tür (2019) propose a universal task-oriented DA schema, but their dataset is still insufficient for pre-training purposes and the schema lacks some important features such as *not_sure* and *dont_understand*. To this end, we follow ISO (Bunt et al. 2010) and propose a more comprehensive unified DA taxonomy for task-oriented dialog, which consists of 20 frequently-used DAs. Base on that, we align the annotations of eight existing benchmarks: MultiWOZ (Budzianowski et al. 2018), Frames (Asri et al. 2017), MSRe2e (Li et al. 2018), SGD (Rastogi et al. 2020), DSTC2 (Henderson, Thomson, and Williams 2014), SimJoint (Shah et al. 2018), STAR (Mosig, Mehri, and Kober 2020) and DailyDialog (Li et al. 2017). We add DailyDialog, an open-domain dialog dataset, to accommodate our dialog policy for more general types. Finally, a new dataset *UniDA* is obtained. Table 1 shows more detailed statistics.

Name	# Dialogs	# Utterance	# Unified DA
MultiWOZ	10,433	142,968	11
Frames	1,369	19,986	14
MSRe2e	10,087	74,686	12
SGD	22,825	463,284	9
DSTC2	3,235	44,532	7
SimJoint	3,008	24,112	6
STAR	6,652	107,846	11
DailyDialog	13,117	98,366	9
UniDA	70,726	975,780	20
Unified DAs	*request, select, reqalts, affirm, not_sure, inform, impl-confirm, expl-confirm, notify_success, notify_failure, hi, bye, negate, repeat, welcome, thank_you, direct, dont_understand, propose, offer*		

Table 1: Statistics of the labeled dataset UniDA.

# Datasets	14
# Dialog Sessions	14M
# Utterances	35M
Avg. Utterances per Dialog	2.5
Avg. Tokens per Utterance	14.6

Table 2: Statistics of the unlabeled dataset UnDial.

Unlabeled Dataset: *UnDial*

Large clean dialogs are difficult to acquire. We build the unlabeled dialog corpora from various available sources, ranging from online forum chatting logs to customer service conversations. We select 14 existing dialog corpora and perform careful processing on all data. Then we acquire a large-scale unlabeled dialog dataset *UnDial*, which consists of 35M utterances. Table 2 shows the statistics of our final pre-training unlabeled data.

Method

In this section, we first introduce the model architecture. Then we describe each objective used in our pre-training and the proposed semi-supervised pre-training paradigm.

Model Architecture

We choose UniLM (Dong et al. 2019) as our backbone model. It contains a bi-directional encoder for understanding and a uni-directional decoder for generation, which is naturally suitable for task-oriented dialog modeling. The encoder and the decoder are weight-shared. We adopt a similar scheme of input representation in Bao et al. (2020), where the input embeddings consist of four elements: tokens, roles, turns, and positions. Role embeddings are like segmentation embeddings in BERT and are used to differentiate which role the current token belongs to, either user or system. Turn embeddings are assigned to each token according to its turn number. Position embeddings are assigned to each token according to its relative position within its belonging sentence.

[1]In some datasets, the dialog act is defined as a combination of an act and its semantic contents. To unify different datasets, we neglect the contents and only use dialog acts as the annotations. We also focus on the text-in-text-out TOD systems in this paper, and leave the spoken DA in the future research.

Pre-training Objectives

Four objectives are employed in our dialog pre-training process: response selection, response generation, DA prediction and consistency regularization. Figure 2 illustrates the procedure of pre-training.

Response Selection. Many work (Wu et al. 2020; Bao et al. 2020; Henderson et al. 2019) show that the response selection task can capture the coherency between dialog contexts and responses and thus benefit dialog understanding. We follow their implementation and model this task as a binary classification problem. Specifically, for a context response pair (c, r) from the corpus, the positive example (with label $l = 1$) is obtained by concatenating c with its corresponding response r, and the negative example (with label $l = 0$) is constructed by concatenating c with a response r^- that is randomly selected from the corpus. A binary cross-entropy loss is defined as:

$$\mathcal{L}_{\text{RS}} = -\log p\left(l = 1|c, r\right) - \log p\left(l = 0|c, r^-\right) \quad (1)$$

in which the classification probability $p\left(l|c, r\right)$ is calculated by feeding the concatenated sequence of c and r into the bi-directional encoder and adding a binary classification head

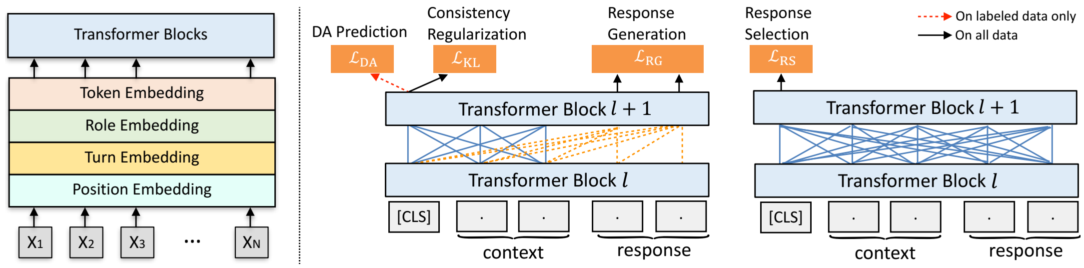

Figure 2: Architecture of our pre-trained dialog model. The left part illustrate the input representations, which contain embeddings of tokens, roles, turns, and positions. The right part shows the pre-trained objectives. Blue lines denote the bi-directional attention. Dashed yellow lines denote the uni-directional attention.

on the extracted representation h_{cls} of token [CLS] from the last transformer layer:

$$p\left(l=1|c,r\right)=\text{sigmoid}\left(\phi_a(h_{cls})\right)\in\mathbb{R}^1 \tag{2}$$

where ϕ_a is a fully-connected neural network with the output layer of size 1. sigmoid is the sigmoid function acts on each dimension of the input vector.

Response Generation. The response generation task aims to predict the dialog response r auto-regressively based on the dialog context c. We adopt the standard negative log-likelihood loss for the generation task:

$$\mathcal{L}_{\text{RG}}=-\sum_{t=1}^{T}\log p\left(r_t|c,r_{<t}\right) \tag{3}$$

where r_t is the t-th word in r, $r_{<t}=\{r_1,...,r_{t-1}\}$ represents the words of previous steps.

DA Prediction. For a context response pair (c,r) sampled from UniDA, the DA prediction task aims to predict the DA label a of the response r based merely on the context c. Note that, since there are some responses in UniDA are associated with multiple DAs, we model the DA prediction task as a multi-label classification problem. We denote $a=(a_1,a_2,...,a_N)$, where N is the total number of dialog acts. A multi-dimensional Bernoulli distribution is used for dialog acts: $p(a|c)=\prod_i^N p(a_i|c)$. Taking the dialog context c as input, we add a multi-dimensional binary classifiers on h_{cls} to predict each act a_i. The binary classification loss is:

$$\begin{aligned}\mathcal{L}_{\text{DA}}=-\sum_{i=1}^{N}&\{y_i\log p(a_i|c)\\&+(1-y_i)\log\left(1-p(a_i|c)\right)\}\end{aligned} \tag{4}$$

$$p\left(a|c\right)=\text{sigmoid}\left(\phi_b\left(h_{cls}\right)\right)\in\mathbb{R}^N \tag{5}$$

where ϕ_b is a fully-connected neural network with the output layer of size N. $y_i\in\{0,1\}$ is the true label of a_i.

Consistency Regularization. For UnDial, the DA annotations are unavailable. In that case, we need to infer the DA labels based on the given dialog context c. Instead of using $p(a|c)$ in Eq. (5), we use a categorical distribution $q(a|c)$ for dialog acts:

$$q\left(a|c\right)=\text{softmax}\left(\phi_b(h_{cls})\right)\in\mathbb{R}^N \tag{6}$$

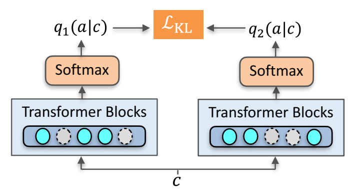

Figure 3: The procedure of computing $\mathcal{L}_{\text{KL}}$.

where softmax is the softmax function, ϕ_b is the same feed-forward neural network in Eq. (5). So $\sum_{i=1}^{N}q(a_i|c)=1$. Then we employ a dropout-based consistency regularization to learn better representations (Gao, Yao, and Chen 2021). Concretely, given the same dialog context c, we feed c to go through the forward pass of the model twice. Due to the randomness of the dropout mechanism in transformers, we can get two different sets of hidden features, and therefore, two different categorical distributions of dialog policy, denoted as $q_1(a|c)$ and $q_2(a|c)$. Then the Kullback-Leibler (KL) divergence between these two output distributions is calculated as $\mathcal{D}_{KL}(q_1\|q_2)$. We minimize the bidirectional KL divergence as in (Liang et al. 2021) between the two distributions to regularize the model predictions, which is defined as:

$$\mathcal{L}_{KL}=\frac{1}{2}\left(\mathcal{D}_{KL}\left(q_1\|q_2\right)+\mathcal{D}_{\text{KL}}\left(q_2\|q_1\right)\right) \tag{7}$$

Figure 3 illustrate the procedure of computing $\mathcal{D}_{KL}$.

Semi-supervised Pre-training Paradigm

We aim to leverage semi-supervised pre-training to learn better pre-trained representations from both the labeled and unlabeled data. For the labeled dataset UniDA, we use all objectives to optimize. The total loss $\mathcal{L}_{\text{label}}$ is computed as:

$$\mathcal{L}_{\text{label}}=\mathcal{L}_{\text{RS}}+\mathcal{L}_{\text{RG}}+\mathcal{L}_{\text{DA}}+\mathcal{L}_{\text{KL}} \tag{8}$$

For the unlabeled data UnDial, since some dialogs collected from the open-domain Internet are too noisy to be compatible with our DA taxonomy, we propose to use a gating mechanism to select a high-quality subset of UnDial for prediction. In practice, we compute a soft gating score $g\in[0,1]$ based on the entropy of $q(a|c)$ to control whether a data point is adopted for consistency regularization in the current iteration.

$$g = \min\left\{\max\left\{0, \frac{E_{max} - (E + \log E)}{E_{max}}\right\}, 1\right\} \quad (9)$$

where $E_{max} = \log N$ is the Maximum Entropy of N-dimensional probability distribution. E is the current entropy of $q(a|c)$, i.e., $E = \sum_i^N q(a_i|c) \log q(a_i|c)$. In practice, we use the perturbed distribution $q_1(a|c)$ as the approximation of $q(a|c)$ to calculate the gate score.

Hence, we have the loss $\mathcal{L}_{\text{unlabel}}$ for the unlabeled data to adjust it adaptively by the gate g as following:

$$\mathcal{L}_{\text{unlabel}} = \mathcal{L}_{\text{RS}} + \mathcal{L}_{\text{RG}} + g\mathcal{L}_{\text{KL}} \quad (10)$$

The final loss $\mathcal{L}_{\text{pre}}$ is computed as:

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{unlabel}} + \mathcal{L}_{\text{label}} \quad (11)$$

In the pre-training process, we mix and shuffle UniDA and UnDial, and randomly sample batches from the mixed corpus.

Fine-tuning and Inference

In the fine-tuning stage, we concentrate on task-oriented dialog tasks. For tasks that contained necessary semantic labels (e.g., belief states and dialog acts), we re-organize the response r to contain those labels, and generate them together. Suppose the sequence of the labels is d. Thus the new response $r^* = (d, r)$ is the concatenation of d and r and is generated in the downstream tasks. For tasks that do not have semantic labels, we generate the initial response r. We also maintain the DA prediction task to alleviate the model discrepancy between pre-training and fine-tuning (Zeng and Nie 2021). Therefore, The fine-tuning loss is as follows:

$$\mathcal{L}_{\text{fine}} = \mathcal{L}_{\text{RS}} + \mathcal{L}_{\text{RG}} + \alpha\mathcal{L}_{\text{DA}} \quad (12)$$

where $\alpha = 1$ for tasks that provide DA annotations and $\alpha = 0$ for tasks that contain no DA annotations.

Experimental Settings

Evaluation Datasets

We evaluate the end-to-end dialog system performance of GALAXY on two well-studied task-oriented dialog benchmarks: Stanford In-Car Assistant (In-Car) (Eric and Manning 2017), MultiWOZ (Budzianowski et al. 2018). In-Car consists of dialogs between a user and an in-car assistant system covering three tasks: calendar scheduling, weather information retrieval, and point-of-interest navigation. Following the data processing in (Zhang et al. 2020a), we divide the dataset into training/validation/testing sets with 2425/302/304 dialogs respectively. MultiWOZ is a large-scale human-human dataset spanning seven domains, which is one of the most challenging datasets in task-oriented dialog due to its complex ontology and diverse language styles. We evaluate our model on MultiWOZ2.0 (the original version) and MultiWOZ2.1 (a revised version) since both are popular benchmarks with various competing models. Following the data processing in Yang, Li, and Quan (2020), we obtain 8438/1000/1000 dialogs for training/validation/testing respectively. We also adopt delexicalized responses for task-oriented generation, which allows the model to learn value-independent parameters (Zhang, Ou, and Yu 2020).

Evaluation Metrics

We use `BLEU` (Papineni et al. 2002) to measure the response generation quality. Metrics relate to task completion are used for separate datasets to facilitate comparison with prior works. For MultiWOZ, we report `Inform`, `Success`, as a combined score (`Comb`) is also computed via (`Inform` + `Success`)×0.5+`BLEU` as an overall quality measure as in Mehri, Srinivasan, and Eskenazi (2019). For In-Car, we use `Match` and `SuccF1` following Lei et al. (2018), and calculate a similar combined score (`Comb`) via (`Match` + `SuccF1`)×0.5+`BLEU`.

Experimental Results

In our experiments, we focus on the setting of end-to-end dialog modeling (E2E), in which no ground-truth immediate labels are provided to the model. GALAXY is initialized with UniLM and then performs semi-supervised pre-training with UniDA and UnDial. Notably, we removed the validation and testing set of MultiWOZ from UniDA during pre-training for fairness. We compare GALAXY with all published work on respective datasets. We also compare different pre-trained conversation models (PCMs) and different semi-supervised pre-training methods to verify the efficacy of GALAXY. In addition, we conduct an extensive discussion and analysis to reveal the internal performance of GALAXY.

Benchmark Performance

As shown in Table 3 and Table 4, GALAXY achieves new state-of-the-art combined scores on all datasets, improving In-Car by 2.5 points (from 104.95 to 107.45), MultiWOZ2.0 by 5.3 points (from 105.05 to 110.35), and MultiWOZ2.1 by 5.5 points (from 105.25 to 110.76). Note that in both tables, GALAXY is the only model that can obtain best `Success` while maintaining `BLEU` at a very high level, which means that GALAXY can take better dialog policy than other models to facilitate task completion, and therefore generate better responses. Our model can also achieve competitive results in `Inform` on par with other best baselines. We also report the results of GALAXY (w/o pre-train) without the pre-training procedure on more dialog corpora. From both tables, GALAXY also achieves comparable results with previous best models, indicating that our model architecture is competitive for dialog modeling.

Comparison with Other PCMs

We verify that GALAXY has a much better ability to fulfill task-oriented dialog tasks than other PCMs due to modeling dialog policy during pre-training. To alleviate the discrepancy brought from model structure, we use UniLM (Dong et al. 2019) and PLATO (Bao et al. 2020) as our baselines. We also train both models on our pre-training dialog datasets (UniDA and UnDial) with their original objectives and perform the same fine-tuning process on MultiWOZ2.0. We denote the new models as TOD-UniLM and TOD-PLATO, respectively. As shown in Table 5, the results of both models are worse than GALAXY due to the lack of using important information of dialog policy.

Model	MultiWOZ2.0				MultiWOZ2.1			
	Inform	Success	BLEU	Comb	Inform	Success	BLEU	Comb
SimpleTOD (Hosseini-Asl et al. 2020)	84.40	70.10	15.01	92.26	85.00	70.50	15.23	92.98
DoTS (Jeon and Lee 2021)	86.59	74.14	15.06	95.43	86.65	74.18	15.90	96.32
SOLOIST (Peng et al. 2020a)	85.50	72.90	16.54	95.74	–	–	–	–
MinTL (Lin et al. 2020)	84.88	74.91	17.89	97.79	–	–	–	–
PPTOD (Su et al. 2021)	89.20	79.40	18.62	102.92	87.09	79.08	19.17	102.26
UBAR (Yang, Li, and Quan 2020)	**95.40**	80.70	17.00	105.05	**95.70**	81.80	16.50	105.25
GALAXY(w/o pre-train)	93.10	81.00	18.44	105.49	93.50	81.70	18.32	105.92
GALAXY	94.40	**85.30**	**20.50**	**110.35**	95.30	**86.20**	**20.01**	**110.76**

Table 3: E2E performances on MultiWOZ2.0/2.1. All results are from original papers. 'w/o pre-train' means using original weights of UniLM for initialization.

Model	Match	SuccF1	BLEU
SEDST (Jin et al. 2018)	84.50	82.90	19.30
TSCP (Lei et al. 2018)	84.50	81.10	21.90
LABES (Zhang et al. 2020a)	**85.80**	77.00	22.80
FSDM (Shu et al. 2019)	84.80	82.10	21.50
GALAXY (w/o pre-train)	81.90	83.30	22.00
GALAXY	85.30	**83.60**	**23.00**

Table 4: E2E performances on In-Car. All results are from original papers. 'w/o pre-train' means using original weights of UniLM for initialization.

Model	Inform	Success	BLEU	Comb
UniLM	92.40	81.40	18.45	105.35
PLATO	91.20	77.20	16.68	100.88
TOD-UniLM	93.50	81.30	19.13	106.53
TOD-PLATO	92.10	79.40	17.23	102.98
GALAXY	**94.40**	**85.30**	**20.50**	**110.35**

Table 5: E2E performances of different pre-trained conversation models on MultiWOZ2.0.

Comparison with Other Semi-supervised Pre-training Methods

As shown in Table 6, we also compare GALAXY with other semi-supervised pre-training methods on MultiWOZ2.0. Specifically, we employ three baselines: Pseudo-Labeling, Variation Autoencoder (VAE), and multi-task learning. For multi-task learning, we discard the $\mathcal{L}_{\mathrm{KL}}$ loss for GALAXY, which represents that model does not perform any inference over DA labels on UnDial. We denote this method as GALAXY$_{multi}$. The results in Table 6 show that VAE has the worst performance because it is difficult to pre-train stochastic latent variables well. Multi-task learning is the most substantial baseline among the three methods, which indicates the importance of integrating DA annotations in the pre-training process. However, without inference on unlabeled dialog samples, GALAXY$_{multi}$ can not explore the stored knowledge of dialog policy thoroughly.

Low Resource Evaluation

Many recent works (Peng et al. 2020b; Wu et al. 2020) have demonstrated that pre-trained models have a solid few-shot ability in the understanding and conditional generation tasks. We also evaluate GALAXY in the simulated low resource setting on MultiWOZ2.0, showing that it is more sample-efficiency than existing models. Specifically, we use 5%, 10% and 20% of the training set data to train our models and baselines. To be fair, we discard the (1-X%) training data of MultiWOZ from UniDA in the pre-training process under each X% setting, eliminating the influence of using any external data. Experimental results in Table 7 show that GALAXY significantly outperforms other models under all low-resource settings.

Model	Inform	Success	BLEU	Comb
Pseudo-Labeling	90.10	80.30	16.79	101.99
VAE	89.00	76.40	16.48	99.18
GALAXY$_{multi}$	93.90	82.30	19.17	107.27
GALAXY	**94.40**	**85.30**	**20.50**	**110.35**

Table 6: E2E performance of different semi-supervised pre-training methods on MultiWOZ2.0.

Analysis and Discussion

In this section, we try to answer three questions: (1) How does our semi-supervised method work during the pre-training process? (2) How much improvements does $\mathcal{L}_{\mathrm{DA}}$, $\mathcal{L}_{\mathrm{KL}}$ and the gating mechanism contribute? (3) How can our model improve task completion in real cases?

Learning Curve. In order to figure out how consistency regularization loss can influence the pre-training, we monitor the predicted DA accuracy and $\mathcal{L}_{\mathrm{KL}}$. Specifically, we conduct a simulated experiment where 10% UniDA and 100% UnDial are used for training, and the rest of UniDA is held out as a testing set. Then we observe the testing DA F1 score and the $\mathcal{L}_{\mathrm{KL}}$ loss on the rest of UniDA data. Note that our goal is to mimic the actual case that whether the model can learn well given limited labeled data and large unlabeled data. As we can see from Figure 4, $\mathcal{L}_{\mathrm{KL}}$ decreases to zero at the beginning, indicating that the model falls into the *collapsing* mode (Chen and He 2021), which means all outputs collapse to a constant. However, since we have the $\mathcal{L}_{\mathrm{DA}}$ loss on labeled data, the collapsing problem can be tackled in the following iterations. On the other hand, the regularization loss $\mathcal{L}_{\mathrm{KL}}$ performs on the labeled data can also avoid over-fitting to some extent, which is shown in Figure 4 that the testing DA F1 score keeps increasing during the pre-training without degradation.

Model	5% data			10% data			20% data		
	`Inform`	`Success`	`BLEU`	`Inform`	`Success`	`BLEU`	`Inform`	`Success`	`BLEU`
DAMD	56.60	24.50	10.60	62.00	39.40	14.50	77.90	70.30	12.10
SOLOIST	69.30	52.30	11.80	69.90	51.90	14.60	74.00	60.10	15.24
MinTL	75.48	60.96	13.98	78.08	66.87	15.46	82.48	68.57	13.00
PPTOD	79.86	63.48	14.89	84.42	68.36	15.57	84.94	71.70	17.01
UBAR	73.04*	60.28*	16.03*	79.20*	68.70*	16.09*	82.50*	66.60*	**17.72***
GALAXY	**80.59**	**67.43**	**17.39**	**87.00**	**75.00**	**17.65**	**89.55**	**75.85**	17.54

Table 7: E2E results of low-resource experiments. 5% (400 dialogs), 10% (800 dialogs), 20% (1600 dialogs) of training data is used to train each model. $*$ denotes our re-implementation results.

Model	`Inform`	`Success`	`BLEU`
GALAXY	**94.40**	**85.30**	**20.50**
$-g$	94.20	83.50	19.26
$-\mathcal{L}_{\mathrm{DA}}$	89.10	79.90	18.77
$-\mathcal{L}_{\mathrm{KL}}$	93.90	82.30	19.17
$-\mathcal{L}_{\mathrm{DA}} - \mathcal{L}_{\mathrm{KL}}$	93.30	81.20	19.54

Table 8: E2E results of ablation study on MultiWOZ2.0.

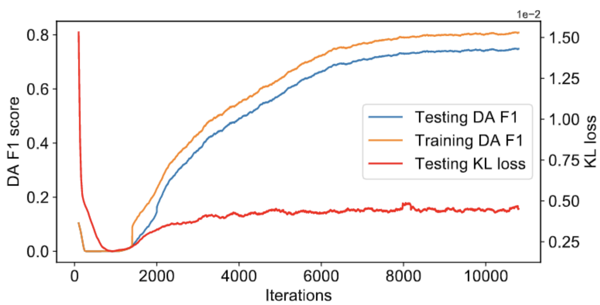

Figure 4: Learning curves of train/test DA F1 scores and the $\mathcal{L}_{\mathrm{KL}}$ loss.

Ablation Results. Table 8 shows the ablation results of GALAXY on MultiWOZ2.0. Without $\mathcal{L}_{\mathrm{DA}}$, GALAXY performs worst because of the collapsing problem. GALAXY without $\mathcal{L}_{\mathrm{KL}}$ equals to multi-task learning, but the results are not as good as our semi-supervised learning due to the inadequate utilization of unlabeled data. If we discard both losses, which backs to the use of common pre-training objectives $\mathcal{L}_{\mathrm{RS}}$ and $\mathcal{L}_{\mathrm{RG}}$, we can acquire 106.79 in `Comb`, suggesting that our pre-training dialog datasets are high-quality and can facilitate task-oriented dialog training. We also examine the function of the gating mechanism. Note that adding the gate g is essential for improving model performance, indicating that it can filter inappropriate data for our semi-supervised pre-training. Figure 5 shows the predicted gating scores of four utterances from UnDial and the DAs annotated manually for the corresponding responses.

Case Study. Figure 6 illustrates a case where GALAXY chooses correct dialog acts for the first two turns so that the whole conversation can steer towards successful task completion. On the contrary, UBAR takes a wrong DA *notify-failure* at the beginning turn and a redundant DA *request* at the second turn, which leads to a failure for the interaction.

Context	i need either the email address , or just zip code. **(Gate: 1.0)**
Response	zip code : 24627. **(DA: inform)**
Context	i need to return an item , can you help me? **(Gate: 0.91)**
Response	sure , may i have your name please? **(DA: request)**
Context	i pour a little liquor out for habeas. **(Gate: 0.41)**
Response	i pour it into corpus. **(DA: N.A.)**
Context	one word : justice. **(Gate: 0.19)**
Response	let me guess , you drive a 1980 ford pinto. **(DA: N.A.)**

Figure 5: Examples of predicted gating scores give the context. Responses are also annotated with DAs for analysis. 'N.A.' means we cannot find a suitable DA for the response.

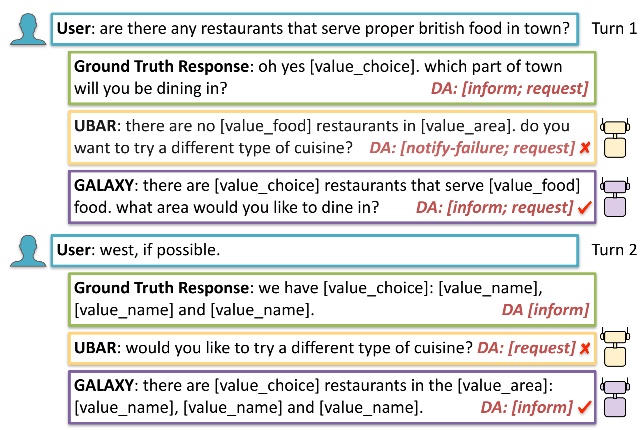

Figure 6: Case Study: Delexicalized responses generated by GALAXY and UBAR on MultiWOZ2.0 test data.

Conclusion

In this paper, we propose GALAXY, a pre-trained conversation model that learns dialog policy explicitly in the pre-training process via semi-supervised learning. We introduce a dialog act prediction task for policy optimization and use a consistency regularization loss to learn better representations on unlabeled dialog corpora. A gating mechanism is also used to weigh suitable unlabeled samples. Experiments show that our model creates new SOTA results on several task-oriented dialog benchmarks and outperforms existing models by a large margin in various low-resource settings. We hope that GALAXY, and the newly collected labeled dataset *UniDA* and large-scale unlabeled corpus *UnDial*, can inspire researchers to explore the new paradigm to build pre-trained conversation models for task-oriented dialog. In the future, we will extend this paradigm into more dialog tasks.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (No. 61906185), Youth Innovation Promotion Association of CAS China (No. 2020357), Shenzhen Science and Technology Innovation Program (Grant No. KQTD20190929172835662), Shenzhen Basic Research Foundation (No. JCYJ20200109113441941). This work was supported by Alibaba Group through Alibaba Research Intern Program. We also thank Dr. Yichi Zhang for the cute cartoon in Figure 1.

References

Adiwardana, D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Asri, L. E.; Schulz, H.; Sharma, S.; Zumer, J.; Harris, J.; Fine, E.; Mehrotra, R.; and Suleman, K. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. *arXiv preprint arXiv:1704.00057*.

Bao, S.; He, H.; Wang, F.; Wu, H.; and Wang, H. 2020. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Budzianowski, P.; and Vulić, I. 2019. Hello, it's GPT-2–how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 2018. MultiWOZ–A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. *arXiv preprint arXiv:1810.00278*.

Bunt, H. 2009. The DIT++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, 13–24.

Bunt, H.; Alexandersson, J.; Carletta, J.; Choe, J.-W.; Fang, A. C.; Hasida, K.; Lee, K.; Petukhova, V.; Popescu-Belis, A.; Romary, L.; et al. 2010. Towards an ISO standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *CVPR*, 15750–15758.

Dai, Y.; Li, H.; Li, Y.; Sun, J.; Huang, F.; Si, L.; and Zhu, X. 2021. Preview, Attend and Review: Schema-Aware Curriculum Learning for Multi-Domain Dialogue State Tracking. In *ACL-IJCNLP*, 879–885.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H.-W. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Eric, M.; and Manning, C. D. 2017. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414*.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *arXiv preprint arXiv:2104.08821*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

He, W.; Sun, Y.; Yang, M.; Ji, F.; Li, C.; and Xu, R. 2021. Multi-goal multi-agent learning for task-oriented dialogue with bidirectional teacher–student learning. *Knowledge-Based Systems*, 213: 106667.

He, W.; Yang, M.; Yan, R.; Li, C.; Shen, Y.; and Xu, R. 2020. Amalgamating knowledge from two teachers for task-oriented dialogue system with adversarial training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3498–3507.

Henderson, M.; Casanueva, I.; Mrkšić, N.; Su, P.-H.; Wen, T.-H.; and Vulić, I. 2019. Convert: Efficient and accurate conversational representations from transformers. *arXiv preprint arXiv:1911.03688*.

Henderson, M.; Thomson, B.; and Williams, J. D. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, 263–272.

Hosseini-Asl, E.; McCann, B.; Wu, C.-S.; Yavuz, S.; and Socher, R. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

Jeon, H.; and Lee, G. G. 2021. Domain State Tracking for a Simplified Dialogue System. *arXiv preprint arXiv:2103.06648*.

Jin, X.; Lei, W.; Ren, Z.; Chen, H.; Liang, S.; Zhao, Y.; and Yin, D. 2018. Explicit state tracking with semi-supervision for neural dialogue generation. In *CIKM*, 1403–1412.

Kingma, D. P.; and Welling, M. 2019. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*.

Kulhánek, J.; Hudeček, V.; Nekvinda, T.; and Dušek, O. 2021. Augpt: Dialogue with pre-trained language models and data augmentation. *arXiv preprint arXiv:2102.05126*.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896.

Lei, W.; Jin, X.; Kan, M.-Y.; Ren, Z.; He, X.; and Yin, D. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1437–1447.

Li, X.; Wang, Y.; Sun, S.; Panda, S.; Liu, J.; and Gao, J. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv preprint arXiv:1807.11125*.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Liang, X.; Wu, L.; Li, J.; Wang, Y.; Meng, Q.; Qin, T.; Chen, W.; Zhang, M.; and Liu, T.-Y. 2021. R-Drop: Regularized Dropout for Neural Networks. *arXiv preprint arXiv:2106.14448*.

Lin, Z.; Madotto, A.; Winata, G. I.; and Fung, P. 2020. Mintl: Minimalist transfer learning for task-oriented dialogue systems. *arXiv preprint arXiv:2009.12005*.

Liu, B.; Tur, G.; Hakkani-Tur, D.; Shah, P.; and Heck, L. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. *arXiv preprint arXiv:1804.06512*.

Mehri, S.; Eric, M.; and Hakkani-Tur, D. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *arXiv preprint arXiv:2009.13570*.

Mehri, S.; Razumovskaia, E.; Zhao, T.; and Eskenazi, M. 2019. Pretraining methods for dialog context representation learning. *arXiv preprint arXiv:1906.00414*.

Mehri, S.; Srinivasan, T.; and Eskenazi, M. 2019. Structured fusion networks for dialog. *arXiv:1907.10016*.

Mosig, J. E.; Mehri, S.; and Kober, T. 2020. Star: A schema-guided dialog dataset for transfer learning. *arXiv preprint arXiv:2010.11853*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 311–318.

Paul, S.; Goel, R.; and Hakkani-Tür, D. 2019. Towards universal dialogue act tagging for task-oriented dialogues. *arXiv preprint arXiv:1907.03020*.

Peng, B.; Li, C.; Li, J.; Shayandeh, S.; Liden, L.; and Gao, J. 2020a. SOLOIST: Building Task Bots at Scale with Transfer Learning and Machine Teaching. *arXiv preprint arXiv:2005.05298*.

Peng, B.; Zhu, C.; Li, C.; Li, X.; Li, J.; Zeng, M.; and Gao, J. 2020b. Few-shot natural language generation for task-oriented dialog. *arXiv preprint arXiv:2002.12328*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1: 9.

Rastogi, A.; Zang, X.; Sunkara, S.; Gupta, R.; and Khaitan, P. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8689–8696.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Shuster, K.; Smith, E. M.; et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Shah, P.; Hakkani-Tur, D.; Liu, B.; and Tur, G. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *NAACL (Industry Papers)*, 41–51.

Shu, L.; Molino, P.; Namazifar, M.; Xu, H.; Liu, B.; Zheng, H.; and Tur, G. 2019. Flexibly-structured model for task-oriented dialogues. *arXiv preprint arXiv:1908.02402*.

Su, P.-H.; Budzianowski, P.; Ultes, S.; Gasic, M.; and Young, S. 2017. Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. *arXiv preprint arXiv:1707.00130*.

Su, Y.; Shu, L.; Mansimov, E.; Gupta, A.; Cai, D.; Lai, Y.; and Zhang, Y. 2021. Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System. *CoRR*, abs/2109.14739.

Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Tian, H.; Wu, H.; and Wang, H. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8968–8975.

Verma, V.; Kawaguchi, K.; Lamb, A.; Kannala, J.; Bengio, Y.; and Lopez-Paz, D. 2019. Interpolation consistency training for semi-supervised learning. *arXiv preprint arXiv:1903.03825*.

Wu, C.-S.; Hoi, S.; Socher, R.; and Xiong, C. 2020. TOD-BERT: pre-trained natural language understanding for task-oriented dialogue. *EMNLP 2020*.

Wu, C.-S.; and Xiong, C. 2020. Probing task-oriented dialogue representation from language models. *arXiv preprint arXiv:2010.13912*.

Xu, Y.; and Zhao, H. 2021. Dialogue-oriented Pre-training. *arXiv preprint arXiv:2106.00420*.

Yang, Y.; Li, Y.; and Quan, X. 2020. UBAR: Towards Fully End-to-End Task-Oriented Dialog Systems with GPT-2. *arXiv preprint arXiv:2012.03539*.

Yu, T.; Zhang, R.; Polozov, A.; Meek, C.; and Awadallah, A. H. 2020. SCoRe: Pre-Training for Context Representation in Conversational Semantic Parsing. In *International Conference on Learning Representations*.

Zeng, Y.; and Nie, J.-Y. 2021. An Investigation of Suitability of Pre-Trained Language Models for Dialogue Generation–Avoiding Discrepancies. *ACL 2021*.

Zhang, Y.; Ou, Z.; Wang, H.; and Feng, J. 2020a. A Probabilistic End-To-End Task-Oriented Dialog Model with Latent Belief States towards Semi-Supervised Learning. *arXiv preprint arXiv:2009.08115*.

Zhang, Y.; Ou, Z.; and Yu, Z. 2020. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *AAAI*, volume 34, 9604–9611.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation. *ACL 2020*.