Few-shot Natural Language Generation for Task-Oriented Dialog

Baolin Peng, Chenguang Zhu, Chunyuan Li
Xiujun Li, Jinchao Li, Michael Zeng, Jianfeng Gao
Microsoft Research, Redmond
{bapeng,chezhu,chunyl,xiul,jincli,nzeng,jfgao}@microsoft.com

Abstract

As a crucial component in task-oriented dialog systems, the Natural Language Generation (NLG) module converts a dialog act represented in a semantic form into a response in natural language. The success of traditional template-based or statistical models typically relies on heavily annotated data, which is infeasible for new domains. Therefore, it is pivotal for an NLG system to generalize well with limited labelled data in real applications. To this end, we present FEWSHOTWOZ, the first NLG benchmark to simulate the few-shot learning setting in task-oriented dialog systems. Further, we develop the SC-GPT[1] model. It is pre-trained on a large set of annotated NLG corpus to acquire the controllable generation ability, and fine-tuned with only a few domain-specific labels to adapt to new domains. Experiments on FEWSHOTWOZ and the large Multi-Domain-WOZ datasets show that the proposed SC-GPT significantly outperforms existing methods, measured by various automatic metrics and human evaluations.

1 Introduction

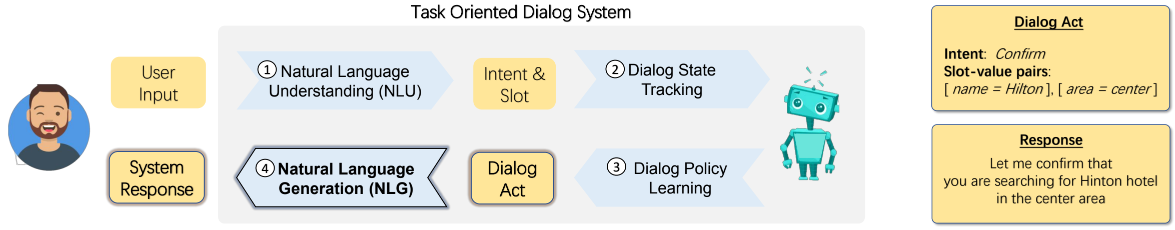
Task-oriented dialog systems are becoming increasingly popular, as they can assist users in various daily activities such as ticket booking and restaurant reservations. In a typical task-oriented dialog system, the *Natural Language Generation* (NLG) module plays a crucial role: it converts a system action (*e.g.,* often specified in a semantic form selected by a dialog policy) into a final response in natural language. Hence, the response should be *adequate* to represent semantic dialog actions, and *fluent* to engage users' attention. As the ultimate interface to interacts with users, NLG plays a significant impact on the users' experience.

Existing methods for NLG can be broadly summarized into two major categories. (*i*) *Template-based methods* require domain experts to handcraft templates for each domain, and the system fills in slot-values afterward (Cheyer and Guzzoni, 2014; Langkilde and Knight, 1998). Thus, the produced responses are often adequate to contain the required semantic information, but not always fluent and nature, hurting users' experiences. (*ii*) *Statistical language models* such as neural networks (Gao et al., 2019) learn to generate fluent responses via training from labelled corpus. One canonical model is *semantically conditioned LSTM* (SC-LSTM) (Wen et al., 2015b), which encodes dialog acts with one-hot representations and uses it as an extra feature to inform the sentence generation process. Despite its good performance on simple domains, it requires large amounts of domain-specific annotated data which is not available for many domains in real-world applications. Even worse, this renders severe scalability issues when the number of possible combinations of dialog acts grows exponentially with the number of slots in more complex domains.

We revisit the current research benchmarks for NLG, and notice that each dialog domain is extensively labelled to favor model training. However, this is in contrast to the real-world application scenarios, where only very limited amounts of labelled data are available for new domains. To simulate such a few-shot learning setting, we have developed a new benchmark dataset, called FEWSHOTWOZ, based on the MultiWOZ (Budzianowski et al., 2018) and Cambridge NLG datasets (Wen et al., 2016a). FEWSHOTWOZ consists of dialog utterances from 7 domains. For each domain, we provide less than 50 labeled utterances for fine-tuning. We believe that FEWSHOTWOZ can better inspire research to address the challenge of learning data-hungry statistical models with very limited amounts of labelled data in real-world scenarios.

To deal with the challenge of few-shot learning,

[1] **S**emantically-**C**onditioned **G**enerative **P**re-**T**raining


Task Oriented Dialog System
User Input
System Response
① Natural Language Understanding (NLU)
④ Natural Language Generation (NLG)
Intent & Slot
Dialog Act
② Dialog State Tracking
③ Dialog Policy Learning
Dialog Act
Intent: Confirm
Slot-value pairs:
[name = Hilton], [area = center]
Response
Let me confirm that you are searching for Hinton hotel in the center area

(a) The overall framework of a task-oriented dialog system

(b) Dialog act & Response

Figure 1: Illustration of the NLG module in the overall task-oriented dialog system. (a) The NLG module is highlighted with glowing black bounding boxes. (b) One example of dialog act (including intent and slot-value pairs) and its corresponding natural language response.

we develop the SC-GPT model. SC-GPT is a multi-layer Transformer neural language model, trained in three steps: (*i*) Pre-trained on plain text, similar to GPT-2 (Radford et al.); (*ii*) Continuously pre-trained on large amounts of dialog-act labeled utterances corpora to acquire the ability of controllable generation; (*iii*) Fine-tuned for a target domain using very limited amounts of domain labels. Unlike GPT-2, SC-GPT generates semantically controlled responses that are conditioned on the given semantic form, similar to SC-LSTM but requiring much less domain labels to generalize to new domains.

In summary, our key contributions are three-fold:

- A new benchmark FewShotWOZ is introduced to simulate the few-shot adaptation setting where only a handful of training data from each domain is available.
- We propose a new model SC-GPT. To our best knowledge, this work is the first study of exploiting state-of-the-art pre-trained language models for NLG in task-oriented dialog systems.
- On the MultiWOZ dataset, SC-GPT creates a new SOTA, outperforming previous models by 4 points in BLEU. On FewShotWOZ, SC-GPT outperforms several strong baselines such as SC-LSTM and HDSA (Chen et al., 2019), showing that SC-GPT adapts to new domain much more effectively, requiring much smaller amounts of in-domain labels. We release our code[2] and dataset[3] for reproducible research.

2 Background

A typical task-oriented spoken dialog system uses a pipeline architecture, as shown in Figure 1 (a), where each dialog turn is processed using a four-step procedure. (*i*) Transcriptions of users input are first passed to the natural language understanding (NLU) module, where the users intention and other key information are extracted. (*ii*) This information is then formatted as the input to dialog state tracking (DST), which maintains the current state of the dialog. (*iii*) Outputs of DST are passed to the dialog policy module, which produces a dialog act based on the facts or entities retrieved from external resources (such as a database or a knowledge base). (*iv*) The dialog act emitted by the dialog policy module serves as the input to the NLG, through which a system response in natural language is generated. In this paper, we focus on the NLG component of task-oriented dialog systems, *i.e.,* how to produce natural language responses conditioned on dialog acts.

Specifically, *dialog act* $\mathcal{A}$ is defined as the combination of intent $\mathbf{I}$ and slot-value pairs $\{(s_i, v_i)\}_{i=1}^{P}$:

$$\mathcal{A} = [\underbrace{\mathbf{I}}_{\text{Intent}}, \underbrace{(s_1, v_1), \cdots, (s_P, v_P)}_{\text{Slot-value pairs}}] \qquad (1)$$

where P is the number of pairs[4], which varies in different dialog acts.

- *Intents* are usually used to distinguish different types of system actions. Typical examples include *inform*, *request*, *confirm*, *select etc.*

[2] https://github.com/pengbaolin/SC-GPT

[3] Project website: https://aka.ms/scgpt

[4] In some literature, dialog act denotes only the type of system actions, slot-value pairs are defined as meaning representations. Throughout this paper, we follow the usage in Budzianowski et al. (2018) and use dialog acts to indicate system action and associated slot-value pairs.

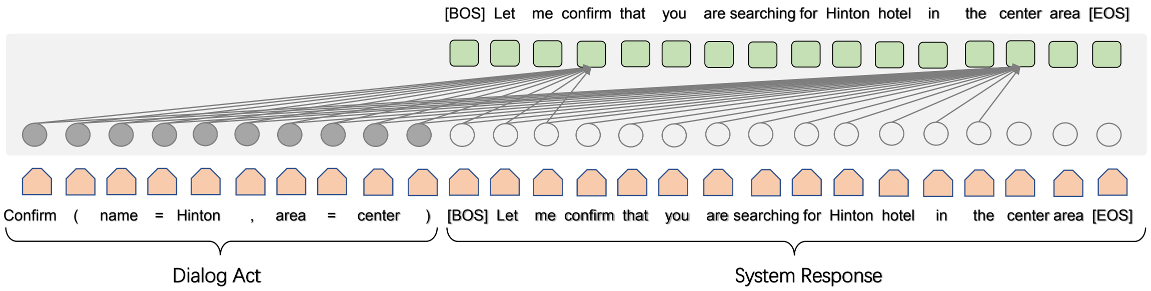

[BOS] Let me confirm that you are searching for Hinton hotel in the center area [EOS]
Confirm (name = Hinton , area = center) [BOS] Let me confirm that you are searching for Hinton hotel in the center area [EOS]
Dialog Act
System Response

Figure 2: Illustration of SC-GPT. In this example, SC-GPT generates a new word token (*e.g.*, "`confirm`" or "`center`") by attending the entire dialog act and word tokens on the left within the response.

- *Slot-value pairs* indicate the category and content of the information to express in the utterance, respectively.

The goal of NLG is to translate $\mathcal{A}$ into a natural language response $\boldsymbol{x} = [x_1, \cdots, x_T]$, where T is the sequence length. In Figure 1 (b), we show an example of the dialog act: `confirm` *(name=Hilton, area=center)*, and the corresponding natural language response is "*Let me confirm that you are searching for Hilton in the center area*".

3 Semantically Conditioned GPT

We tackle this generation problem using conditional neural language models. Given training data of N samples $\mathcal{D} = \{(\mathcal{A}_n, \boldsymbol{x}_n)\}_{n=1}^N$, our goal is to build a statistical model parameterized by $\boldsymbol{\theta}$ to characterize $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\mathcal{A})$. To leverage the sequential structure of response, one may further decompose the joint probability of $\boldsymbol{x}$ using the chain rule, casting an auto-regressive generation process as follows:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\mathcal{A}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(x_t|x_{<t}, \mathcal{A}) \tag{2}$$

where $x_{<t}$ indicates all tokens before t.

Learning $\boldsymbol{\theta}$ is performed via maximizing the log-likelihood (MLE) of the conditional probabilities in (2) over the entire training dataset:

$$\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D}) = \sum_{n=1}^{|\mathcal{D}|} \sum_{t=1}^{T_n} \log p_{\boldsymbol{\theta}}(x_{t,n}|x_{<t,n}, \mathcal{A}_n) \tag{3}$$

In this paper, we employ the Transformers (Vaswani et al., 2017) to parameterize the conditionals in (2). To enable strong generalization and controllable ability for the learned model, we propose the following three-stage procedure as the training recipe.

Massive Plain Language Pre-training. Large models trained on massive training corpus usually generalize better to new domains. Inspired by this, we inherit the GPT-2 architecture (Radford et al.) as the backbone language model. GPT-2 is an auto-regressive language model that leverages 12-24 layers of masked, multi-head self-attention Transformers. GPT-2 is pre-trained on extremely massive text data OpenWebText (Radford et al.). It has demonstrated superior performance on characterizing human language data distribution and knowledge transfer. Given text prompts, GPT-2 can often generate realistic sentences.

Dialog-Act Controlled Pre-training. To enable the guidance of dialog act in response generation, we propose to continuously pre-train the GPT-2 model on large amounts of annotated (dialog act, response) pairs. The pre-training dataset[5] includes annotated training pairs from Schema-Guided Dialog corpus, MultiWOZ corpus, Frame corpus, and Facebook Multilingual Dialog Corpus. The total size of the pre-training corpus is around 400k examples.

We firstly pre-process dialog act $\mathcal{A}$ into a sequence of control codes using the following format:

$$\mathcal{A}' = [\ \mathbf{I}\ \ (\ \ s_1\ \ =\ \ v_1\ \ ,\ \ \cdots\ \ s_P\ \ =\ \ v_P\ \ )\ ] \tag{4}$$

Meanwhile, the output sequence $\boldsymbol{x}'$ is pre-processed via appending $\boldsymbol{x}$ with a special start token `[BOS]` and an end token `[EOS]`. Finally, the sequentialized dialog act $\mathcal{A}'$ is concatenated with its augmented response $\boldsymbol{x}'$, and then fed into GPT-2. During training, the prediction loss is only computed for $\boldsymbol{x}'$, and $\mathcal{A}'$ provides the attended conditions. Since the dialog act represents the semantics of the generated sentences, we follow the naming

[5] The domains appearing in fine-tuning are excluded.

Statistics	E2E NLG	BAGEL	RNNLG	FEWSHOTWOZ
# Domains	1	1	4	7
Avg. # Intents	1	8	11.25	8.14
Avg. # Slots	8	10	21	16.15
Avg. # Delexicalised DAs in Training	109	23.9	794.5	50
Avg. # Delexicalised DAs in Testing	7	14.3	566.5	472.857
Overlap Percentage	100%	99.6%	94.00%	8.82%
Avg. # Training Instances	42056	363	4625.5	50
Avg. # Testing Instances	630	41	1792.5	472.86

Table 1: Comparison of existing NLG datasets, including E2E NLG (Novikova et al., 2017), BAGEL(Mairesse et al., 2010), Cambridge NLG(Wen et al., 2016a) and the proposed FEWSHOTWOZ.

convention of SC-LSTM, and term our model as *Semantically Conditioned Generative Pre-training* (SC-GPT). The overall architecture of SC-GPT is illustrated in Figure 2.

Fine-tuning. For a new domain, a dialog act usually contains novel intents or slot-value pairs, and annotated training samples are often limited. We fine-tune SC-GPT on limited amounts of domain-specific labels for adaptation. The fine-tuning follows the same procedure of dialog-act controlled pre-training, as described above, but uses only a few dozens of domain labels.

It is worth noticing that the above recipe has several favorable properties:

- *Flexibility.* SC-GPT operates on a sequence of tokens without delexicalization, which means that SC-GPT does not assume a fixed one-hot or tree-structured dialog act representation vectors. Hence, it has great flexibility in extending to novel dialog acts.
- *Controllability.* In contrast to GPT-2 that generates natural sentences without high-level semantic guidance, SC-GPT can generate sentences with adequate intent and slot-value information and maintain its fluency.
- *Generalizability.* SC-GPT is able to generalize significantly better than SC-LSTM, due to the pre-training on massive plain text corpora and annotated dialog datasets.

4 Dataset: FEWSHOTWOZ

Revisiting NLG Benchmarks. The three commonly used NLG datasets in developing and evaluating task-oriented dialog systems are E2E NLG (Novikova et al., 2017) BAGEL (Mairesse et al., 2010) and RNNLG (Wen et al., 2016a), as summarized in Table 1. We observe two issues from their shared statistics: (*i*) All the datasets contain a large number of labelled training samples for each domain, ranging from hundreds to tens of thousands. However, the cost of labeling is high in practice, *e.g.,* labeling 50 utterances is 5 hours per domain. Creating such an extensively annotated dataset for each new domain is prohibitively expensive. (*ii*) The percentage of distinct delexicalised dialog acts between training and testing data is quite small. For example, the delexicalised dialog acts in testing is 100% covered by the training set for the E2E NLG dataset. It renders difficulties in evaluating the model's generalization ability for new domains.

FEWSHOTWOZ. To build a setting for more pragmatic NLG scenarios, we introduce a new dataset FEWSHOTWOZ to better reflect real application complexity, and encourage the community to develop algorithms that are capable of generalizing with only a few domain-specific labels for each (new) domain. The dataset statistics are shown in the last column of Table 1. We see that FEWSHOTWOZ is different from the other datasets in three aspects: (*i*) *More domains*. FEWSHOTWOZ contains seven domains in total, which is larger than any existing NLG datasets. (*ii*) *Less training instances*. Importantly, FEWSHOTWOZ has a much smaller number of training instances per domain, aiming to evaluate the few-shot learning ability. (*iii*) *Lower training/testing overlap*. FEWSHOTWOZ has only 8.82% overlap, significantly smaller than the other datasets, which amount to more than 90% overlap. The average number of intents per instance in `Attraction`/ `Taxi`/ `Train` domain is 2, 1.33, and 2.05, respectively. In contrast, there is only one intent for each example in the other datasets. The NLG task defined on FEWSHOTWOZ requires the models to learn to generalize over new compositions of intents. The

Statistics	Restaurant	Hotel	Laptop	TV	Attraction	Taxi	Train
# Intent	9	10	13	13	5	2	5
# Slot	21	19	22	22	10	7	13
# DAs in training	50	50	50	50	50	40	50
# DAs in testing	129	78	1379	680	340	47	657
Overlap Percentage	35.56	60.26	2.61	5.74	13.82	72.34	6.55
Avg. #DAs per Instance	1	1	1	1	2	1.33	2.05
# Training Instances	50	50	50	50	50	40	50
# Testing Instances	129	78	1379	680	340	47	657

Table 2: FEWSHOTWOZ statistics over 7 different domains.

details of FEWSHOTWOZ is shown in Table 2.

Collection Protocols. We construct FEWSHOTWOZ via re-organizing data samples from RNNLG and MultiWOZ datasets (Budzianowski et al., 2018). For each domain in RNNLG, we first group utterances according to their delexicalised dialog acts, and keep only one utterance as the target sentence. To ensure diversity, we consider three domains from MultiWOZ: `Attraction`, `Taxi`, and `Train`. Since MultiWOZ is a cross-domain dataset, the dialog act of an utterance may exist in multiple domains. We choose to keep utterances whose dialog act appears only in one domain. Similar delexicalising processing is applied to ensure that each dialog act has only one target utterance. Finally, to simulate the few-shot learning in practice, we randomly sample 50 training examples for each domain, except the `Taxi` domain, which has 40 examples.

5 Related Work

Pre-trained Models. Pre-trained language models (PLMs) have substantially advanced the state-of-the-art across a variety of natural language processing (NLP) tasks (Peters et al., 2018; Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019; Keskar et al., 2019; Raffel et al., 2019). PLMs are often trained to predict words based on their context on massive text data, and the learned models can be fine-tuned to adapt to various downstream tasks. The closest line of research to ours are GPT-2 (Radford et al.), CTRL (Keskar et al., 2019) and Grover (Zellers et al., 2019). GPT-2 first investigated missive Transformer-based auto-regressive language models with large-scale text data for pre-training. After fine-tuning, GPT-2 achieves drastic improvements on several generation tasks. One drawback of GPT-2 is the lack of high-level semantic controlling ability in language generation. To alleviate this issue, CTRL (Keskar et al., 2019) was introduced to train the model based on pre-defined codes such as text style, content description, and task-specific behavior, meanwhile Grover (Zellers et al., 2019) was proposed to generate news articles conditioned on authors, dates *etc.* Although conceptually similar to our SC-GPT, CTRL and Grover cannot be readily applied to NLG in task-oriented dialog systems, as the conditioning codes are quite different. Another controllable generation work for GPT-2 is PPLM (Dathathri et al., 2019), which provides a decoding scheme to guide the generation process using key-words or classifiers, without re-training the model. In this paper, we focus on pre-training an NLG model conditioned on finer-grained semantic dialog acts, which are more desirable for dialog systems.

Dialog. Various dialog systems have been developed (Gao et al., 2019), including task-oriented dialog systems such as Rasa[6], Microsoft Bot Framework[7], and Conversational Learner[8], and chit-chat systems such as XiaoIce (Zhou et al.), DialoGPT (Zhang et al., 2019), Meena (Adiwardana et al., 2020). In this paper, we focus on task-oriented systems, particularly the NLG module. With the blooming of deep learning, neural sequential models have shown powerful capability and flexibility in NLG. Extensive efforts have been made, including new architecture choices such as RNNs (Wen et al., 2015a), attention RNNs (Dušek and Jurčíček, 2016), SC-LSTM (Wen et al., 2015b) and its variants (Tran et al., 2017; Tran and Nguyen, 2017), as well as learning objectives (Zhu et al., 2019). However, they all require large amounts of annotated data to reach satisfactory performance. A more realistic scenario is to require much less

[6] https://rasa.com/

[7] https://dev.botframework.com/

[8] https://www.microsoft.com/en-us/research/project/conversation-learner/

Model	Restaurant		Laptop		Hotel		TV		Attraction		Train		Taxi	
	BLEU ↑	ERR ↓	BLEU ↑	ERR ↓	BLEU ↑	ERR ↓	BLEU ↑	ERR ↓	BLEU ↑	ERR ↓	BLEU ↑	ERR ↓	BLEU ↑	ERR ↓
SC-LSTM	15.90	48.02	21.98	80.48	31.30	31.54	22.39	64.62	7.76	367.12	6.08	189.88	11.61	61.45
GPT-2	29.48	13.47	27.43	11.26	35.75	11.54	28.47	9.44	16.11	21.10	13.72	19.26	16.27	9.52
SC-GPT	**38.08**	**3.89**	**32.73**	**3.39**	**38.25**	**2.75**	**32.95**	**3.38**	**20.69**	**12.72**	**17.21**	**7.74**	**19.70**	**3.57**

Table 3: Performance of different methods on FEWSHOTWOZ

Model	Informativeness	Naturalness
SC-LSTM	2.29	2.13
GPT-2	2.54^{*}	2.38^{*}
SC-GPT	$2.64^{*\dagger}$	$2.47^{*\dagger}$
Human	2.92	2.72

* $p < 0.005$, comparison with SC-LSTM
† $p < 0.05$, comparison with GPT

Table 4: Human evaluation on FEWSHOTWOZ. Statistical significance is computed with a two-tailed t-test.

labeling and improve the sample efficiency of models, This is especially important when deploying the models to new domains, where dialog acts need to be labelled from scratch. Our paper aims to formally set up such a research scenario by proposing a new dataset FEWSHOTWOZ, and a new model SC-GPT.

6 Experiments

In this section, we evaluate the proposed SC-GPT on the FEWSHOTWOZ and MultiWOZ datasets to answer two research questions: (i) Is SC-GPT an effective model for strong generalization and controllability in dialog response generation? (ii) Does FEWSHOTWOZ meet the goal of effectively evaluating the generalization of NLG models in the few-shot learning setting?

6.1 Experimental Setup

Implementation details. The model was built upon Huggingface Pytorch Transformer (Wolf et al., 2019). We use GPT2-Medium with 345M parameters[9] as the initial checkpoint, and byte pair encodings (Sennrich et al., 2015) for the tokenization. Linear rate scheduler with start rate as 5e-5 was used for both pre-training and fine-tuning. Adam (Kingma and Ba, 2014) with weight decay was used to optimize the parameters. For pre-training, the model was trained with a mini-batch of 8 on an 8 Nvidia V100 machine until observing no significant progress on validation loss or up to 20 epochs, whichever is earlier. For fine-tuning on FEWSHOTWOZ, models were trained on each domain separately with five epochs.

Automatic metrics. Following Wen et al. (2015b), BLEU scores and the slot error rate (ERR) are reported. BLEU score evaluates how natural the generated utterance is compared with human readers. ERR measures the exact matching of the slot tokens in the candidate utterances. ERR $= (p+q)/M$, where M is the total number of slots in the dialog act, and p, q is the number of missing and redundant slots in the given realisation. For each dialog act, we generate five utterances and select the top one with the lowest ERR as the final output.

Human evaluation. We conducted the human evaluation using Amazon Mechanical Turk to assess subjective quality. We recruit master level workers (who have good prior approval rates) to perform a human comparison between generated responses from two systems (which are randomly sampled from comparison systems). The workers are required to judge each utterance from 1 (bad) to 3 (good) in terms of informativeness and naturalness. *Informativeness* indicates the extent to which generated utterance contains all the information specified in the dialog act. *Naturalness* denotes whether the utterance is as natural as a human does. To reduce judgement bias, we distribute each question to three different workers. Finally, we collected in total of 5800 judges.

Baselines. We compare with three baseline methods. (i) **SC-LSTM** (Wen et al., 2015b) is a canonical model and a strong baseline that uses an additional dialog act vector and a reading gate to guide the utterance generation. (ii) **GPT-2** (Radford et al.) is used to directly fine-tune on the domain-specific labels, without pre-training on the large-scale corpus of (dialog act, response) pairs. (iii) **HDSA** (Chen et al., 2019) is a state-of-the-art

[9] We also experimented using GPT2 with 117M parameters but observed significant poor performance.

Model	Entity F1	BLEU
SC-LSTM (Wen et al., 2015b)	80.42	21.6
HDSA (Chen et al., 2019)	87.30	26.48
GPT-2	87.70	30.71
SC-GPT	**88.37**	**30.76**

Table 5: Performance on MultiWOZ

Model	Data size						
	0.1%	0.5%	1%	5%	10%	20%	50%
SC-LSTM	9.05	15.15	15.38	18.26	18.97	19.99	21.07
HDSA	9.40	15.32	18.27	22.19	22.89	24.16	25.01
GPT-2	11.96	18.88	20.29	24.18	25.39	26.25	27.40
SC-GPT	**12.70**	**19.65**	**20.67**	**24.45**	**25.67**	**26.37**	**27.89**

Table 6: BLEU score of different models on MultiWOZ using training data of different sizes.

model on MultiWOZ. It leverages dialog act structures to enable transfer in the multi-domain setting, showing superior performance than SC-LSTM.

6.2 FewShotWOZ

Table 3 reports the automatic evaluation performance of different methods on FewShotWOZ. SC-LSTM fails to learn the generation effectively in this few-shot learning setting. The generated utterances are poor in quality and suffer from inaccurate slot rendering. In addition, GPT-2 performs consistently better than SC-LSTM in all the domains. It reveals the feasibility of using a pre-trained language model for NLG, though only limited annotations are available for fine-tuning. Importantly, SC-GPT performs significantly better than GPT and SC-LSTM in terms of both BLEU and ERR. In all the domains, SC-GPT reduces the ERR to a significantly lower level, revealing its strong controllability power. This verifies the importance of pre-training on large annotated dialog data, as SC-GPT learns how to generate utterances specified by the dialog acts accurately.

Table 4 shows the human assessment on FewShotWOZ. The results exhibit the same trend with automatic evaluation. SC-GPT outperforms GPT-2 and SC-LSTM significantly in both metrics, *i.e.,* SC-GPT can better control the generation to convey information in the dialog act while maintaining good fluency. Note that the gap between SC-GPT and human annotation is still large, indicating that the proposed FewShotWOZ exhibits an under-explored research area, and provides a large space to encourage future research for improvement.

Model	**Informativeness**	**Naturalness**
SC-LSTM	2.14	2.33
HDSA	2.34	2.42
SC-GPT	2.71^*	2.69^*
Human	2.77	2.61

$^*\ p < 0.005$

Table 7: Human evaluation on MultiWOZ. Statistical significance was computed with a two-tailed t-test between SC-GPT and HDSA.

6.3 MultiWOZ

The results on MultiWOZ are shown in Table 5. Following Chen et al. (2019), Entity F1 (Wen et al., 2016b) is used to evaluate the entity coverage accuracy (including all slot values, days, numbers, and reference, *etc.*). Again, SC-GPT achieves the best performance on BLEU score. Note that GPT-2 performs similarly with SC-GPT on the full MultiWOZ dataset, this is because MultiWOZ contains 57k utterances, which is large enough for GPT-2 to achieve good performance. The results also confirm that with enough annotated data, conditional language model formulation performs significantly better than HDSA, a strong competitor that leverages graph/tree-structure information to encode dialog acts.

To study how SC-GPT performs with different training data sizes. We further conduct experiments with varying percentages of training data on MultiWOZ, ranging from 0.1% (50 examples) to 50%. As shown in Table 6, the observations are consistent with FewShotWOZ. SC-GPT performs consistently better than GPT-2, HDSA, and SC-LSTM for a wide range of dataset sizes, and the improvement is more substantial when the fewer numbers of in-domain labels are used for fine-tuning.

Table 7 shows the human assessment results on MultiWOZ. The results are consistent with the automatic evaluation. It is interesting to see that (*i*) the gap between the new state-of-the-art method (*i.e.,* SC-GPT) and human performance on FewShotWOZ (as shown in Table 4) is much larger than that on MultiWOZ; (*ii*) the human rating on the naturalness of SC-GPT is even higher than humans on MultiWOZ, while there is a visible gap on FewShotWOZ. These results demonstrate that FewShotWOZ presents a challenging few-shot learning setting, SG-GPT serves as a simple and strong baseline in this setting, and the combined

#	Model	Generated Responses from Different Models
1	***Input DA***	Laptop{inform(name=satellite proteus 84; type=laptop; memory=8 gb; drive=1 tb; weight=2.3 kg)}
	Reference	the satellite proteus 84 is a laptop with a 1 tb drive , 8 gb memory and weighs in at 2.3 kg
	SC-LSTM	the satellite proteus 84 is a laptop with 8 gb , with a 1 tb drive , and is for business computing , and is for business computing % *[businesscomputing=true]* *[weight=2.3kg]*
	GPT-2	the satellite proteus 84 laptop has a 1 tb drive , weighs 2.3 kg and is used for business computing % *[businesscomputing=true]* *[memory=8 gb]*
	SC-GPT	the satellite proteus 84 is a laptop with 8 gb memory , 1 tb drive , and a weight of 2.3 kg
2	***Input DA***	Restaurant{inform_only_match(name=marnee thai and thanh long; pricerange=moderate; near=outer sunset; goodformeal=dinner)}
	Reference	marnee thai and thanh long are the only 2 moderately priced restaurants near outer sunset that are good for dinner
	SC-LSTM	marnee thai and thanh long is is moderate restaurant restaurant and good good good dinner % *[near=outer sunset]* *[inform_only_match]*
	GPT-2	there is a moderately priced restaurant called marnee thai and thanh long that is near the outer sunset area % *[goodformeal=dinner]* *[inform_only_match]*
	SC-GPT	marnee thai and thanh long is the only restaurant that serves moderately priced food near outer sunset and good for dinner
3	***Input DA***	Train{inform(time=50 minutes)}, hotel{request(stars=?; area=?), inform(choice=5)}
	Reference	travel time is 50 minutes . i found 5 hotels you may like . do you have a star rating request or an area you prefer
	SC-LSTM	there are 5 trains that are 50 minutes . do you have a preferred departure location ? % *[train{inform(choice=5),request(departure=?)}]* *[hotel{request(stars=?; area=?), inform(choice=5)}]*
	HDSA	there are 5 trains that meet your criteria . do you have a preference on the area or star rating ? % *[train{inform(choice=5,* *time=50 minutes)}]* *[hotel{inform(choice=5)}]*
	SC-GPT	there are 5 hotels that meet your criteria . the trip will last 50 minutes . do you have an area preference or star rating you would like ?

Table 8: Examples of generated utterances from different models, along with its corresponding dialog acts (DAs) and references. The first two examples are sampled from FewShotWOZ and the last one is from MultiWOZ. Each generated utterance is followed by a brief description explaining the errors (starting with "%"). (Better viewed in color. wrong , redundant , missing information)

Model	Seen		Unseen	
	BLEU ↑	ERR ↓	BLEU ↑	ERR ↓
SC-LSTM	23.05	40.82	12.83	51.98
GPT-2	30.43	3.26	27.92	17.36
SC-GPT	**40.28**	**1.09**	**36.69**	**4.96**

Table 9: Performance of different methods on seen DAs and unseen DAs in restaurant domain.

provides a platform for researchers to develop NLG models that are able to generalize to new domains and generate semantically conditioned and fluent responses.

6.4 Analysis

We perform detailed analysis to investigate SG-GPT's *flexibility*, *controllability* and *generalizability*. The test set is split into two subsets - *seen* and *unseen*. If a dialog act of an example appears in the training set, the example is marked as *seen*; otherwise, it is marked as *unseen*. Table 9 compares different models on the seen and unseen subsets in the `restaurant` domain. SC-GPT yields higher BLEU and lower ERR, and the improvement is more significant on the unseen set. For example, SC-GPT reduces ERR to 4.96, an order of magnitude lower than SC-LSTM and only 1/3 of GPT-2. This demonstrates that SC-GPT generalizes well to novel dialog acts, and is able to precisely ground in them to compose fluent responses. This is further confirmed by the quantitative comparison in Table 8, where we compare the generated utterance examples of different models. While the baseline methods prone to over-generate or miss important slots, SC-GPT can successfully generate fluent natural language utterances that share precise semantic conditions with the ground-truth references.

#	Model	Generated Responses from Different Models
	Original DA	ack_makereservation (price = moderate ; restaurantcuisine = chinese ; location = center ; peoplecount = 2)
	Reference	ok making a reservation for restaurant moderate chinese in center for two people?
1	***Input DA***	ack_makereservation (price = moderate ; restaurantcuisine = japanese ; location = bellevue square ; peoplecount = 2)
	GPT-2	ok making reservation for moderate japanese restaurant in bellevue square for 2 people?
	SC-GPT	ok making a reservation for a moderate japanese cuisine restaurant in bellevue square for 2 people?
2	***Input DA***	ack_makereservation (restaurantname = sakura ; price = moderate ; restaurantcuisine = japanese ; location = bellevue sqare ; peoplecount = 2)
	GPT-2	ok making a reservation for 2 people sakura moderate japanese restaurant located bellevue sqare for 10 minutes?
	SC-GPT	ok making a reservation for sakura, moderate japanese cuisine in bellevue square for 2 people?
3	***Input DA***	ack_makereservation (price = expensive ; restaurantcuisine = japanese ; location = center ; peoplecount = 2)
	GPT-2	ok making reservation for japanese cuisine expensive?
	SC-GPT	ok making reservation for expensive japanese cuisine?
#	**Model**	**Generated Responses from Different Models**
	Original DA	ack_makereservation (childrenfriendly = true)
	Reference	make reservation for a children friendly restaurant ?
1	***Input DA***	ack_makereservation (restaurantname = sakura ; childrenfriendly = true ; price = moderate)
	GPT-2	make reservation at sakura child friendly restaurant and make table moderate price?
	SC-GPT	make reservation for restaurant sakura moderate price and children friendly restaurant?

Table 10: Examples of generated utterances with novel dialog acts. SC-GPT produces better utterances than GPT-2 for with edited dialog acts. Since both models produce similar responses to references for the original dialog act, the results are not shown here. (Better viewed in color. insert a slot, substitute a slot value, delete a slot).

We further simulate the process when deploying SC-GPT for a new domain, using the examples provided in the RASA dialog toolkit [10]. We first fine-tune SC-GPT using a few training examples (only 16 instances in this new domain), and then generate utterances based on novel dialog acts that are unseen in training data, shown in Table 10. In practice, it is desirable for an NLG system to deal with an extending domain whose dialog acts change dynamically. We simulate the setting by editing the original input dialog acts, such as inserting or deleting a slot, or substituting a slot value.

Since SC-LSTM is infeasible in the setting of an extending domain, we compare SC-GPT with GPT-2. Results show that SC-GPT produces better utterances than GPT-2. SC-GPT can generate reasonably good natural language responses with different combinations of editing operations, showing its high flexibility to generalize to new dialog acts with very limited training data, and produce controllable responses.

7 Conclusion and Future Work

In this paper, we have made two major contributions towards developing a more pragmatic NLG module for task-oriented dialog systems: (i) A new benchmark FewShotWOZ is introduced to simulate the few-shot learning scenarios with scarce labelled data in real-world applications. (ii) A new model SC-GPT is proposed to endow the NLG module with strong semantically controlling and generalization ability. Empirical results on both FewShotWOZ and MultiWOZ show that SC-GPT achieves the best overall performance in both automatic and human evaluations.

There are two interesting directions for future work. The first is to design mechanisms to generate more interpersonal responses which are proven to help improve user experiences (Li et al., 2016; Zhou et al.). The other is to generalize the generative pre-training idea to all four modules in

[10] https://github.com/RasaHQ/rasa/tree/master /examples/restaurantbot

the dialog system pipeline for end-to-end training. Since these four modules process information in order, one may organize their input/output as segments, and pre-train a segment-level autoregressive model.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3696–3709, Florence, Italy. Association for Computational Linguistics.

Adam Cheyer and Didier Guzzoni. 2014. Method and apparatus for building an intelligent automated assistant. US Patent 8,677,377.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: a simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

Ondřej Dušek and Filip Jurčíček. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51, Berlin, Germany. Association for Computational Linguistics.

Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 704–710.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

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 preprint arXiv:1907.11692*.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The e2e dataset: New challenges for end-to-end generation. *arXiv preprint arXiv:1706.09254*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Van-Khanh Tran and Le-Minh Nguyen. 2017. Natural language generation for spoken dialogue system using RNN encoder-decoder networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 442–451, Vancouver, Canada. Association for Computational Linguistics.

Van-Khanh Tran, Le-Minh Nguyen, and Satoshi Tojo. 2017. Neural-based natural language generation in dialogue using RNN encoder-decoder with semantic aggregation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 231–240, Saarbrücken, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Tsung-Hsien Wen, Milica Gašić, Dongho Kim, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015a. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 275–284, Prague, Czech Republic. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. 2016a. Multi-domain neural network language generation for spoken dialogue systems. *arXiv preprint arXiv:1603.01232*.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015b. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016b. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *NeurIPS*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, (Just Accepted):1–62.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2019. Multi-task learning for natural language generation in task-oriented dialogue. 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 1261–1266, Hong Kong, China. Association for Computational Linguistics.