DSPM-NLG: A Dual Supervised Pre-trained Model for Few-shot Natural Language Generation in Task-oriented Dialogue System

Yufan Wang[1,3*], **Bowei Zou**[2], **Rui Fan**[1,3], **Tingting He**[3†], **Ai Ti Aw**[2]
[1]National Engineering Research Center for E-Learning, Central China Normal University, China
[2]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[3]Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning,
National Language Resources Monitoring and Research Center for Network Media,
School of Computer, Central China Normal University, China
{yufan_wang,fanrui}@mails.ccnu.edu.cn
{zou_bowei,aaiti}@i2r.a-star.edu.sg
tthe@mail.ccnu.edu.cn

Abstract

In few-shot settings, fully conveying the semantic information of dialogue act is a crucial challenge for Natural Language Generation (NLG) in task-oriented dialogue systems. It is noteworthy that NLG and Spoken Language Understanding (SLU) form a natural dual problem pair. If the SLU module can successfully restore the generated response by the NLG module to the corresponding dialogue act, this would demonstrate that the response is effectively conveying the semantic information of the dialogue act. Based on this idea, a novel Dual Supervised Pre-trained Model for a few-shot Natural Language Generation (DSPM-NLG) is proposed to regularize the pre-training process. We adopt a joint model with a dual supervised framework to learn the dual correlation between NLG and SLU from a probabilistic perspective. In addition, a slot-masked strategy is designed to enable the model to focus more effectively on the key slot-value pairs. DSPM-NLG is continuously trained on publicly available and large-scale labeled data, allowing it to gain a thorough understanding of the duality between the two tasks and to enhance the pre-trained model's ability for semantic control and generalization. Experimental results illustrate that our proposed model demonstrates exceptional performance on the few-shot benchmark dataset, outperforming the previous state-of-the-art results.

1 Introduction

Task-oriented dialogue systems have been demonstrated to be effective in aiding users accomplish various tasks in multiple domains, such as airline ticket booking, restaurant and hotel reservations.

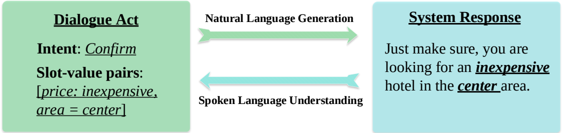

Dialogue Act
Intent: Confirm
Slot-value pairs:
[price: inexpensive, area = center]
Natural Language Generation
Spoken Language Understanding
System Response
Just make sure, you are looking for an inexpensive hotel in the center area.

Figure 1: NLG and SLU are two complementary components that form a natural duality. While NLG is the process of generating a response in natural language based on a structured semantic representation (in green), SLU is the act of transforming natural language into a structured semantic representation (in blue).

A complete task-oriented dialogue system typically consists of four components (Zhang et al., 2020): spoken language understanding (SLU), dialogue state tracking (DST), dialogue policy learning (DPL), and natural language generation (NLG). The NLG module aims to convert the dialogue act generated by DPL into a natural language, which can be abstracted as a semantically conditioned language generation task. As depcicted in Figure 1, the generated utterance should be sufficient to convey semantic information of the dialogue act, as well as being fluent, natural, and resembling human language to engage users' attention. As the primary module for user interaction, NLG plays a crucial impact in the performance of dialogue systems.

Recently, pre-trained models have revolutionized the field of natural language processing. The introduction of pre-trained models such as GPT2 (Radford et al., 2019) in the NLG task has resulted in a significant improvement in overall performance (Budzianowski and Vulić, 2019; Wu et al., 2019; Hosseini-Asl et al., 2020; Ham et al., 2020; Yang et al., 2020; Peng et al., 2021). Despite their superior performance on simple domains, they necessitate a great deal of high-quality labeled data and are challenging to generalize to the domain-specific.

Nevertheless, acquiring large amounts of domain-specific labeled data in practical scenarios

[*] Contribution during the internship at Institute for Infocomm Research.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 12389–12402
July 9-14, 2023 ©2023 Association for Computational Linguistics

is cost-prohibitive. It is essential that an NLG module is able to effectively generalize with limited domain-specific labeled data in few-shot settings. Recently, a paradigm of the few-shot learning utilizs the existing large-scale annotated data to train a pre-trained model such as GPT-2 (Radford et al., 2019) and subsequently is fine-tuned with only a few domain-specific labeled data to adapt to target domains. Thereby, the paradigm narrows the gap between pre-traineds model and downstream tasks. For instance, Peng et al. (2020) adopted the paradigm and achieved a state-of-the-art performance for few-shot NLG. However, in few-shot settings, one of the challenges of NLG is prone to omit important slot-value pairs and make it difficult to fully convey the semantic information of the dialogue act fully.

To go beyond this limitation, we explore further enhancing the semantically controlling ability of the pre-trained model. It is noteworthy that NLG and SLU are a natural dual problem pair, as illustrated in Figure 1. Ideally, the response generated by the NLG module can be restored to the corresponding dialogue acts by the SLU module. The two dual tasks are intrinsically connected due to the joint probabilistic correlation. Moreover, SLU can provide an additional supervision signal for NLG so that the NLG model better focuses on key slot-value pairs in the dialogue acts. Thus, we explicitly exploit the dual correlation between NLG and SLU to regularize the pre-training process and improve the semantically controlling ability of the pre-trained model.

In this paper, we propose a dual supervised pre-trained model for a few-shot Natural Language Generation (DSPM-NLG). DSPM-NLG consists of two primary, *the dual supervised pre-training* and *fine-tuning*. In the pre-training stage, the framework of dual supervised learning is introduced to learn the explicit joint probabilistic correlation between NLG and SLU from existing large-scale annotated data. Moreover, a slot-masked strategy is designed, which selects the key slot information detected by SLU, thereby constraining the NLG module to focus more on the slot-value pairs in the dialogue act. In the fine-tuning stage, the pre-trained model is fine-tuned with only a few domain-specific labels for adaptation. Experiments demonstrate that the semantically controlling and generalization abilities of DSPM-NLG are significantly improved. In general, the major contributions of this paper are described below:

- We propose a novel pre-trained framework for NLG based on dual supervised learning, which explicitly exploits the probabilistic correlation between NLG and SLU to regularize the pre-trained process.

- We design a slot-masked strategy that contributes to constraining the NLG module to focus more on the key slot-value pairs contained in the dialogue act.

- We carry out extensive ablation experiments to demonstrate the advantages of building the framework. The experimental results demonstrate that our model outperforms the existing state-of-the-art results on the few-shot benchmark dataset.

2 Related Work

Existing NLG models can be mainly summarized into two major categories. (1) Template-based NLG models (Langkilde and Knight, 1998; Stent et al., 2004) generate responses according to manually developed rules. These models generate responses that can convey the semantics information of certain predefined dialogue acts. Nevertheless, the handcraft templates are difficult to cover potentially unforeseen dialogue acts, and the generated response is not always natural. (2) Statistical-based NLG models (Wen et al., 2015; Dušek and Jurčíček, 2016; Tran and Nguyen, 2017; Su et al., 2018; Gao et al., 2019; Zhu et al., 2019; Wolf et al., 2019b; Su et al., 2020b,a) generate responses via training from massive annotated data. With the rise of attention mechanism, more approaches have been proposed, e.g., Hierarchical attention network (Su et al., 2018; Zhu et al., 2019; Chen et al., 2019). And then, some NLG works adapted a multi-task learning framework to improve the performance (Su et al., 2020b,a). In particular, some scholars exploit the relationship between SLU and NLG to improve the performance of two tasks (Su et al., 2019, 2020a; Zhu et al., 2020; Tseng et al., 2020; Chang et al., 2021). Subsequently, many works introduce pre-trained models (Budzianowski and Vulić, 2019; Edunov et al., 2019; Dai et al., 2019; Ham et al., 2020; Brown et al., 2020; Kale and Rastogi, 2020; Madotto et al., 2020) such as GPT2, and the overall performance of NLG is greatly improved.

Recently, to deal with the challenge of few-shot learning, data augmentation has been widely applied to NLG. Peng et al. (2020) proposed SC-GPT model. They pre-train GPT with large-scale NLG corpus collected from publicly available dialogue datasets and then fine-tuned the model on the target domain with few training instances. Xu et al. (2021) proposed a data augmentation approach that constructed dialogue acts and responses from the open-domain dialogues and applied the new data to SC-GPT.

Compared with previous work, we try to explore the duality between SLU and NLG in the pre-training stage. The difference between the proposed model and the previous methods is mainly reflected in the following two aspects: First, dual supervised learning is only applied in the pre-training. Thus, in few-shot settings, our model does not require any SLU annotated data and does not increase additional computation in fine-tuning and inference stages. It is worth mentioning that our model also avoids the error transfer between SLU and NLG in the inference stage. Second, in the pre-training stage, we collect a large amount of labeled data for SLU and NLG. The training of a large amount of labeled data enables the pre-trained model to have a strong semantically controlling ability rather than just learning the relationship between the two tasks in some specific domains to improve the performance of both tasks.

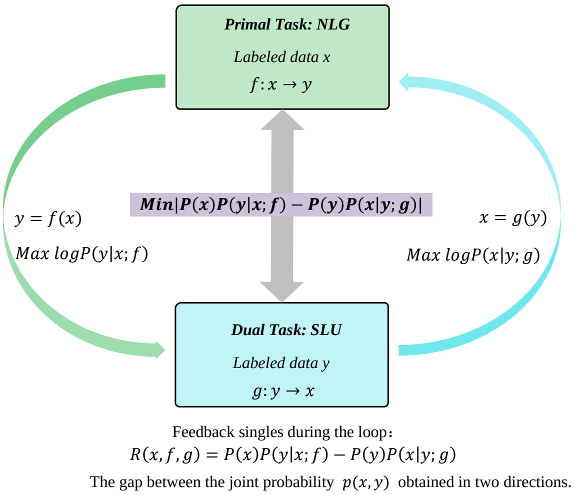
3 Background

Dual Supervised Learning Framework. The overall architecture of the dual supervised learning as shown in Figure 2. Assuming that we involve the dual tasks of NLG and SLU: the primal NLG task takes a sample from the semantics space X as input and maps it to the natural language space Y. The NLG task learns a mapping function $f\left(x;\theta_{x\to y}\right)$ parameterized by $\theta_{x\to y}$. In contrast, a dual task of SLU takes a sample from the natural language space Y as input and maps it to the semantics space X. The SLU task learns a mapping function $g\left(y;\theta_{y\to x}\right)$ parameterized by $\theta_{y\to x}$, where $x \in X$ and $y \in Y$. The joint probabilistic duality can be computed as followings:

$$P(x,y) = P(x)P\left(y \mid x\right) = P(y)P\left(x \mid y\right), \quad (1)$$

where $P(x)$, $P(y)$ denote the marginal distributions; $P(y|x)$, $P(x|y)$ are conditional probability.

For any $x \in X$, $y \in Y$, ideally, the conditional distributions of the primal and dual tasks should satisfy the following equality:



Feedback singles during the loop:
$R(x,f,g) = P(x)P(y|x;f) - P(y)P(x|y;g)$
The gap between the joint probability $p(x,y)$ obtained in two directions.

Figure 2: Illustration of the dual supervised learning.

$$P(x)P\left(y \mid x;\theta_{x\to y}\right) = P(y)P\left(x \mid y;\theta_{y\to x}\right), \quad (2)$$

where $\theta_{x\to y}$ and $\theta_{y\to x}$ are the learnable parameter of the model.

The core idea of dual supervised learning is to jointly model the two dual tasks by minimizing their loss functions and incorporating the probability duality constraint. A total of three loss functions are optimized. Obtain the maximum likelihood estimation of y_i from the labeled input x_i via the primal NLG task:

$$\min_{\theta_{xy}}(1/M)\sum_{i=1}^{M} l_{NLG}\left(f\left(x_i;\theta_{x\to y}\right),y_i\right). \quad (3)$$

Obtain the maximum likelihood estimation of x_i from the dual input y_i via the dual task:

$$\min_{\theta_{yx}}(1/M)\sum_{i=1}^{M} l_{SLU}\left(g\left(y_i;\theta_{y\to x}\right),x_i\right). \quad (4)$$

The probabilistic duality constraint is incorporated:

$$s.t\ P\left(x\right)P\left(y \mid x;\theta_{x\to y}\right) = P(y)P\left(x \mid y;\theta_{y\to x}\right), \quad (5)$$

where l_{NLG}, l_{SLU} are loss functions; M is the number of the samples and $s.t.$ denotes the constraint.

4 Methodology

4.1 Task Definition

The goal of NLG is to generate a natural language response containing the dialogue act's semantic information. A dialogue act ($\mathcal{DA}$) includes different types of system actions and slot-value pairs, the formal definition of $\mathcal{DA}$ is described as follows:

$$\mathcal{DA} = [A, (\text{slot}_1 = \text{value}_1), \ldots, (\text{slot}_k = \text{value}_k)], \quad (6)$$

where A indicates different types of system actions, such as *confirm*, *inform*, *request, etc.*; k is the number of slot-value pairs, which varies in different dialogue acts; slot-value pairs indicate critical structured semantic information of the dialogue act.

The formal definition of NLG is described as follows: given a $\mathcal{DA}$ consisting of a system action and k slot-pairs, a response $Y = [y_1, y_2, \ldots, y_n]$ can be generated by the NLG model, where n is the response length. For example, a $\mathcal{DA}$ is *[confirm, (price range = inexpensive)]* and the corresponding response is *"just to make sure, you are looking for an inexpensive hotel"*. The format of the SLU labels is described as follows: the utterance *"just to make sure, you are looking for an inexpensive hotel"* is labeled as *"O O O O O O O O O B-hotel-pricerange O"*, where "B-hotel-pricerange" and "O" are called slots. There is a one-to-one correspondence between a slot and a word.

4.2 Proposed Model

The section introduces the proposed DSPM-NLG model. The training procedure of DSPM-NLG mainly includes the dual supervised pre-training and fine-tuning stages. The overall architecture of DSPM-NLG is shown in Figure 3.

4.3 Dual Supervised Pre-training Stage

We inherit GPT-2 model (Radford et al., 2019) as our original pre-trained model in the proposed model. The GPT-2 model is a powerful language model which can be used for several downstream tasks. In order to enhance the generalization ability and semantically controlling ability of the pre-trained model, we continuously train the GPT-2 model on existing large-scale high-quality annotation pairs ($\mathcal{DA}$, response, slots)[1]. The pre-training dataset includes annotated training pairs from the MultiWOZ dataset (Eric et al., 2019) and schema-guided dialogue dataset (Rastogi et al., 2020). The total size of the dual supervised pre-training datasets is approximately 470k samples.

Encoder At the pre-training stage, the $\mathcal{DA}$ is pre-processed as a text sequence D. In the meanwhile, the response Y is pre-processed via appending response with a special start token [BOS] and an end token `[EOS]`. The input of our model is $X = \{D, Y\} = \{x_1, \cdots, x_m, x_{m+1}, \cdots, x_{m+n}\}$, where m is the length of the $\mathcal{DA}$ and n is the length of the response. The output of the last hidden layer is $H = \{h_0, \cdots, h_m, h_{m+1}, \cdots, h_{m+n}\}$, h_{m+1}, h_{m+n} denote the final hidden state of the special `[BOS]` and `[EOS]` token.

In the pre-training, the loss value is only compututed for Y corresponding to the hidden layer output $H_y = \{h_{m+1}, \cdots, h_i, \cdots, h_{n+m}\}$, where $h_i \in H_y$ denotes the final hidden state of the i^{th} token in H_y. For the NLG task, we utilize the final hidden state H_y to generate responses, and probability distribution $P(y' \mid x; \theta_{x\to y})$ of the generated tokens is calculated by:

$$\begin{aligned} P(y' \mid x; \theta_{x\to y}) &= softmax(h_i W_U + b_U), \\ f(x; \theta_{x\to y}) &= \arg\max_{y' \in \mathcal{Y}} \{P(y' \mid x; \theta_{x\to y})\}, \end{aligned} \quad (7)$$

where $f(x; \theta_{x\to y})$ is mapping function for NLG; $W_U \in R^{d\times|U|}$ and $b_U \in R^{|U|}$ are weight matrix and bias vector, respectively. d is the dimension of the hidden state vector. Besides, $|U|$ is the length of vocabulary, $\theta_{x\to y}$ is the learnable parameter of the model.

For the SLU task, we input the final hidden state H_y to another trainable linear layer, which is used to predict the slot of the corresponding input token. Then the probability distribution $P(x' \mid y; \theta_{y\to x})$ of slots is calculated by:

$$\begin{aligned} P(x' \mid y; \theta_{y\to x}) &= softmax(h_i W_S + b_S), \\ g(y; \theta_{y\to x}) &= \arg\max_{x' \in \mathcal{X}} \{P(x' \mid y; \theta_{y\to x})\}, \end{aligned} \quad (8)$$

where $g(y; \theta_{y\to x})$ is a mapping function for SLU; $W_S \in R^{d\times|S|}$ and $b_S \in R^{|S|}$ are weight matrix and bias vector, respectively. Besides, $|S|$ is the number of slot labels, and $\theta_{x\to y}$ is the learnable parameter of the model.

Loss Function In this section, we introduce the joint training procedure with dual supervised learning in detail. l_{NLG}, l_{SLU} are loss functions, and the loss values of NLG and SLU are computed as:

$$\begin{aligned} &\min_{\theta_{x\to y}} (E[l_{NLG}(f(x; \theta_{x\to y}), y)]), \\ &\min_{\theta_{y\to x}} (E[l_{SLU}(g(y; \theta_{y\to x}), x)]). \end{aligned} \quad (9)$$

The probabilistic duality constraint is incorporated:

$$s.t P(x) P(y \mid x; \theta_{x\to y}) = P(y) P(x \mid y; \theta_{y\to x}), \quad (10)$$

where $P(x)$ and $P(y)$ are the marginal distributions. Then, the method Lagrange multiplier is used to transfer the probability duality constraint into the objective function. The regularization term

[1] In this paper, we introduce the dual task SLU in the pre-training stage. The slots are denoted as ground-truth labels.

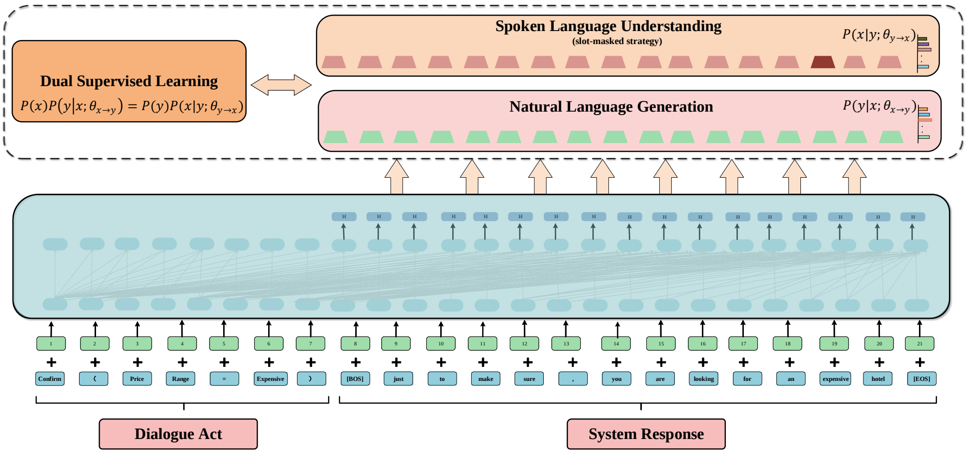


Figure 3: Illustration of the DSPM-NLG model.

is the constraint of the duality probabilistic. The new loss value of NLG is computed as:

$$\min_{\theta_{x\to y}} \left(E\left[l_{NLG}\left(f\left(x;\theta_{x\to y}\right), y\right)\right] + \lambda_{x\to y} l_{\text{duality}}\right), \tag{11}$$

where $\lambda_{x\to y}$ is a hyper-parameter. Besides, ℓ_{duality} denotes the regularization term. The regularization term is computed as:

$$\begin{aligned} \ell_{\text{duality}} = &\Big(\log \hat{P}(x) + \log P\left(y \mid x;\theta_{x\to y}\right) \\ &- \log \hat{P}(y) - \log P\left(x \mid y;\theta_{y\to x}\right)\Big)^2 . \end{aligned} \tag{12}$$

Note that the true marginal distribution of $P\left(x\right)$ and $P\left(y\right)$ are difficult to obtain. As an alternative, we relace them with empirical marginal distributions $\hat{P}\left(x\right)$ and $\hat{P}\left(y\right)$. $\hat{P}\left(x\right)$ is calculated by GPT-2 (language model). The empirical marginal distribution of $\hat{P}\left(y\right)$ is calculated by the statistics of the percentage of each slot in the collected labeled data. The meaning of the regularization term is to minimize the gap between $\hat{P}\left(x\right)P\left(y \mid x;\theta_{x\to y}\right)$ and $\hat{P}\left(y\right)\ P\left(x \mid y;\theta_{y\to x}\right)$. Thus, dual supervised learning enhances the process of supervised learning from the duality of the structure between NLG and SLU. The final NLG loss function is formulated as:

$$\begin{aligned} G_f = \nabla_{\theta_{x\to y}}(1/M)\sum_{j=1}^{M}&\left[l_{NLG}\left(f\left(x_j;\theta_{x\to y}\right), y_j\right)\right. \\ &\left.+\lambda_{x\to y}\ell_{\text{duality}}\left(x_j, y_j;\theta_{x\to y}, \theta_{y\to x}\right)\right], \end{aligned} \tag{13}$$

where M is the number of samples. The regularization term ℓ_{duality} is different from the SVM regularization term or the L1 regularization term. The regularization term of SVM or L1 is only dependent on the model. However, the regularization term ℓ_{duality} in dual supervised learning is both model and data-dependent. During the pre-training process, each training sample contributes to the regularization term. In addition, the probability distribution of SLU contributes to the regularization of the NLG model.

Slot-masked Strategy The slots use the beginning-inside-outside (BIO) data annotation standard (Athiwaratkun et al., 2020) in the SLU task. For example, the utterance *"just to make sure, you are looking for an inexpensive hotel"* is labeled as *"O O O O O O O O O B-hotel-pricerange O"*. We find that most slot labels in SLU are non-value slot "O". According to the statistics, the number of non-value slot labels ("O") is more than ten times that of the valued slots (e.g. "B-hotel-pricerange"). And the valued slot (not the "O" slot) contains critical semantic information and has great significance. Therefore, a slot-masked strategy is designed to select the vital slots detected by SLU. When calculating the loss value, the model only considers the valued slots, which makes it better focused on the key slots detected by SLU.

4.4 Fine-tuning Stage

We fine-tune DSPM-NLG on limited amounts of domain-specific labels for adaptation. The fine-tuning procedure follows standard supervised learning of NLG in few-shot sittings. The loss value of NLG is computed as follows:

$$\min_{\theta_{x\to y}} \left(E\left[l_{NLG}\left(f\left(x;\theta_{x\to y}\right), y\right)\right]\right). \tag{14}$$

It is worth mentioning that dual supervised learning is not applied in the fine-tuning stage, which avoids the error transfer between SLU and NLG.

5 Experimental Setup

Dataset Comparative experiments are conducted on the publicly available datasets for NLG, namely,

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	34.08	6.08	28.67	7.32	**38.35**	6.03	**31.25**	5.31	20.81	11.92	18.60	7.98	20.13	4.22
JM-NLG-sm	36.42	5.45	29.33	4.83	35.98	4.71	29.12	5.44	21.03	11.76	19.23	6.56	19.21	4.63
JM-NLG	37.53	4.76	29.30	4.49	37.04	4.62	30.15	4.93	21.31	11.04	19.38	6.51	20.02	3.92
DSPM-NLG-sm	**38.72**	3.76	29.76	4.31	36.46	**4.56**	30.23	4.87	21.82	11.21	19.74	6.44	20.32	3.26
DSPM-NLG	37.90	**3.34**	**30.33**	**3.93**	37.13	4.67	30.07	**4.45**	**22.31**	**10.32**	**20.36**	**6.32**	**20.83**	**3.13**

Table 1: Experimental results of our models and baseline models. The experimental results with the highest value is bolded. The "JM-NLG" adopts a multi-task learning method to jointly model NLG and SLU in the pre-training. The subscript "-sm" means without the slot-masked strategy.

FEWSHOTWOZ (Peng et al., 2020) and FEWSHOTSGD (Xu et al., 2021), respectively. The two datasets include seven domains and sixteen domains, respectively [2]. Compared with the other existing datasets, they have several favorable properties for few-shot learning: more domain, fewer training instances, and lower training overlap. For the FEWSHOTWOZ, each domain has 50 training instances, and the average number of test instances is 472.857. The overlap percentage is 8.82%. Since SLU has been introduced in the model, labels required for the SLU tasks are added to the standard NLG dataset in the pre-training stage. We obtain labeled data of the SLU according to the dialogue acts by the matching method.

Automatic Metrics In this paper, we continue previous evaluation metrics to evaluate the quality of the generated responses, including BLEU scores and slot error rate (ERR) (Wen et al., 2015). BLEU score is used to evaluate the fluency and naturalness of the generated response. And ERR is used to evaluate whether the generated response contains semantic information in the dialogue act. $ERR = (m\_slot + r\_slot)/k$, where k is the number of slots in a dialogue act, m_slot and r_slot denote the number of missing slots and redundant slots in the given realization, respectively.

Human Evaluation We conduct human evaluations of different models. We randomly select 100 responses generated by each model for human evaluation in the restaurant domain. Three workers are invited to independently rate the responses generated by each model according to the rules (Peng et al., 2020). The works are required to judge each response from 1(bad) to 3(good) in terms of informativeness and naturalness. Finally, we adopt the average score marked by the three volunteers as the final score of each response.

Model	information	Naturalness
SC-GPT	2.57	2.42
DSPM-NLG	2.64	2.49
Human	2.93	2.81

Table 2: Human evaluation on FEWSHOTWOZ.

Baseline Models To verify the effectiveness of the proposed model, several classic NLG models are compared.

SC-LSTM: Wen et al. (2015) design a semantic controlled LSTM cell with a reading gate to guide the response generation. The model is a canonical NLG model and achieves good performance on domain-specific.

GPT-2: The pre-trained GPT-2 (Radford et al., 2019) is directly fine-tuned on the domain-specific labeled data.

SC-GPT (strong baseline): Peng et al. (2020) regard the structured dialogue act as a sequence of tokens and feed the sequence to the generation model. We apply the obtained annotated data to SC-GPT as a strong baseline system.

6 Results and Analysis

We compare our model with previous state-of-the-art models. The overall results of NLG experiments on the FEWSHOTWOZ dataset are shown in Table 1. Although the strong baseline model has achieved solid results, our model outperforms previous state-of-the-art performance in most domains. For the FEWSHOTWOZ dataset, compared with the SC-GPT baseline, DSPM-NLG has a 3.82% absolute improvement in the BLEU score and a 2.76% absolute reduction in the ERR in the restaurant domain. As shown in Table 2, the DSPM-NLG model also achieves better performance in human evaluation indicators. The experimental results express the same trend with automatic evaluation indicators. The results of DSPM-NLG in BLEU on the

[2] See Appendix A for more details of two datasets.

Model	Restaurants	Hotels	Flights	Calendar	Banks	Weather	Buses	Services
GPT-2	08.98	08.84	12.18	05.27	06.09	10.52	07.77	09.79
DSPM-NLG	15.31	14.64	17.03	09.15	08.58	12.97	12.33	15.72

Model	Ridesharing	Media	Movies	Music	Rentalcars	Homes	Events	Travel
GPT-2	03.75	03.17	10.05	05.79	06.79	13.87	09.17	02.08
DSPM-NLG	09.13	07.16	09.86	09.36	09.14	14.54	13.23	11.07

Table 3: The results in BLEU on the FEWSHOTSGD dataset. And the DSPM-NLG model was pre-trained using only the MultiWOZ dataset.

FEWSHOTSGD are shown in Table 3. The results demonstrate that DSPM-NLG reaches stable performance and brings practical values to real-world applications. More importantly, we would like to explore the reason for the improved performance of DSPM-NLG [3]. Therefore, extensive ablation experiments are conducted to analyze the effectiveness of the proposed model.

6.1 Ablation Study

We provide integrated analysis results on the critical components of DSPM-NLG to gain detailed insights:

Effect of jointly modeling NLG and SLU. From the result, JM-NLG performs better than SC-GPT in some domains. In the pre-training stage, JM-NLG adopts a multi-task learning network that jointly trains two tasks. The loss function of JM-NLG not only learns the implicit correlations between tasks but also provides additional supervision signals, which constrains the joint model better to generate the slot-value pairs of the dialogue act. However, the model only takes advantage of the implicit association between the two tasks. Thus, the improvement of JM-NLG is slight.

Effect of the dual supervised pre-trained model. The experimental results show that, compared with the baseline models, DSPM-NLG-sm significantly improves both BLEU and ERR in most domains. The main reason is the dual supervised learning framework models the explicit joint probabilistic correlation between SLU and DST. In the pre-trained stage, the pre-trained model is continuously trained on large-scale dialogue-acts, responses, and slots annotated datasets, which helps the dual supervised learning framework learn the duality between SLU and NLG. And the objective function can be better optimized with large amounts of data. The result reveals the dual structure strengthens the supervised learning process.

Effect of the slot-masked strategy. To further verify the effectiveness of the designed slot-masked strategy, a statistical analysis is performed on the pre-training dataset in the SLU task. We find that the number of non-value slot labels (“O”) is more than ten times that of the valued slots. Although the loss function of SLU assigns a small loss value to the “O”-labeled slots, when the number of “O” slots is large, it may have a negative impact on the model. The slot-masked strategy can mask the “O”-labeled slots and select valued slot information. Therefore, the performance of JM-NLG and DSPM-NLG is further improved. In multi-task learning, the loss value of SLU has a significant impact on the model performance. Therefore, JM-NLG achieves a good performance. And we expect to get a considerable enhancement over DSPM-NLG. However, experimental results show that the performance improvement of DSPM-NLG is limited. To explain it, we think the dual regularization term is related to the loss value of SLU, and the value of the hyperparameter λ in the regularization term is generally small. Although the strategy is reasonable and feasible, the impact of the slot-masked strategy on DSPM-NLG is not significant.

6.2 In-depth Analysis

The generalizability and semantical controllability learned by the pre-trained model is critical to the performance of the model in the fine-tuning stage for few-shot learning. Next, experiments are conducted to analyze the generalization and semantically controlling abilities learned by DSPM-NLG.

Generalizeability (1) We analyze the performance of DSPM-NLG in *different training data sizes*. (2) We analyze the performance of different models on the *seen* dialogue acts and *unseen* dialogue acts in the restaurant domain.

To explore the performance of DSPM-NLG with different training data sizes, we conduct experiments with varying percentages of training data. 20%, 40%, 60%, 80%, and 100% of the training

[3] The parameter settings of the DSPM-NLG model are recorded in Appendix B.

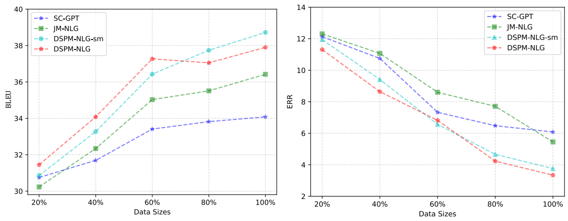


Figure 4: The experimental results of our models and baseline models under different training data sizes.

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	37.18	2.38	32.42	6.17
DSPM-NLG:	39.68	1.34	34.20	4.53

Table 4: The experimental results of our models and baseline models on the seen dialogue acts and unseen dialogue acts.

data are randomly selected from the restaurant domain. The experimental results are shown in Figure 4. Overall, the performance of these models improves in BLEU score and ERR as the size of training data increases. DMSP-NLG performs consistently better than SC-GPT and JM-NLG under different training data sizes. Our model achieves a significant improvement in 60% data size, which exceeds the performance of SC-GPT in 100% data size. In 100% data size, DSPM-NLG has a maximum slope compared to other models. It can be inferred that DSPM-NLG provides more large space for improvement when more numbers of domain labels are used for fine-tuning. The result reflects that our model has a stronger generalization ability than the baseline model.

In the restaurant domain, we split the test set into two subsets seen dialogue acts (DAs) and unseen dialogue acts. The dialogue acts that appear in the training set are called seen DAs ; otherwise, it is marked as unseen DAs. The performance of the unseen DAs can well reflect the generalization ability of the model. The performance of different models is compared on the seen DAs and unseen DAs, as shown in Table 4. On the two subsets, DSPM-NLG yields higher BLEU and lower ERR. It performs consistently better than SC-GPT and JM-NLG. What's more, the improvement of the model is more obvious in the unseen subset. Experiments demonstrate that DSPM-NLG has a strong generalization ability.

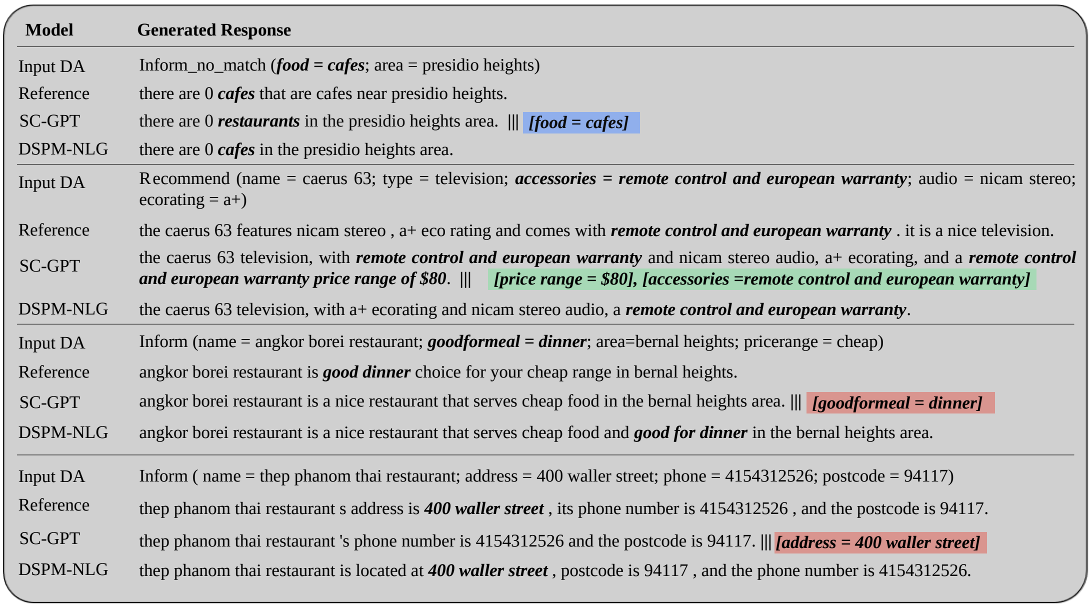
Controllability (1) We compare the generated responses of different models. (2) We analyze the performance of different models on the ERR.

Model	Wrong	Redundant	Omissive
SC-GPT	4.65	4.65	10.85
DSPM-NLG	3.10	2.32	3.10

Table 5: The statistics of generated responses for three types of errors in conveying dialogue acts.

As shown in Figure 5, we select a couple of cases from the FEWSHOTWOZ test set to specifically analyze the difference in generated response between our method and baseline models. We find that these NLG models have three types of errors in conveying dialogue acts: *Wrong* slot-value pairs, *Redundant* slot-value pairs, and *Omissive* slot-value pairs. In the first two cases, SC-GPT generates wrong slot-value pairs and redundant slot-value pairs, respectively. The appearance of the word "restaurant" in the dataset is relatively high. The SC-GPT baseline learns more about the data feature in the dataset than the semantic structure feature of dialogue acts. Consequently, in the baseline model, "`cafes`" is mislabeled as "`restaurants`", and "`accessories`","`pricerange`" are redundant. DSPM-NLG correctly conveys the semantic information of the dialogue act. This further indicates that DSPM-NLG is capable of constraining the NLG task with the semantic information detected by SLU so that our model can convey more accurate dialogue acts. In the fourth case, the baseline model misses a slot-value pair. For the slot "`goodformeal`","`address`" , our model accurately generates it. We think the main reason may be that the key slot information detected by SLU can supervise the generated response, whether it contains slot-value pairs of the dialogue act. And the slot-masked strategy can accurately select the key slot information detected by SLU to restrict the slots that need to be generated. The above results indicate the correctness of exploring the dual correlation between SLU and NLG.

To further quantitatively analyze three types of errors (*Wrong*, *Redundant*, *Omissive*) in conveying dialogue acts of the NLG model, we counted the percentage of three types of errors in the restaurant domain for SC-GPT and DSPM-NLG. The results are shown in Table 5. We found that SC-GPT is prone to omissive important slot-value pairs contained in dialogue acts. In particular, when the number of slot-value pairs in a dialogue act is *greater than **4***, *omissive* slot-value pairs of errors are more



Model	Generated Response
Input DA	Inform_no_match (***food = cafes***; area = presidio heights)
Reference	there are 0 ***cafes*** that are cafes near presidio heights.
SC-GPT	there are 0 ***restaurants*** in the presidio heights area. ||| ***[food = cafes]***
DSPM-NLG	there are 0 ***cafes*** in the presidio heights area.
Input DA	Recommend (name = caerus 63; type = television; ***accessories = remote control and european warranty***; audio = nicam stereo; ecorating = a+)
Reference	the caerus 63 features nicam stereo , a+ eco rating and comes with ***remote control and european warranty*** . it is a nice television.
SC-GPT	the caerus 63 television, with ***remote control and european warranty*** and nicam stereo audio, a+ ecorating, and a ***remote control and european warranty price range of $80***. ||| ***[price range = $80], [accessories =remote control and european warranty]***
DSPM-NLG	the caerus 63 television, with a+ ecorating and nicam stereo audio, a ***remote control and european warranty***.
Input DA	Inform (name = angkor borei restaurant; ***goodformeal = dinner***; area=bernal heights; pricerange = cheap)
Reference	angkor borei restaurant is ***good dinner*** choice for your cheap range in bernal heights.
SC-GPT	angkor borei restaurant is a nice restaurant that serves cheap food in the bernal heights area. ||| ***[goodformeal = dinner]***
DSPM-NLG	angkor borei restaurant is a nice restaurant that serves cheap food and ***good for dinner*** in the bernal heights area.
Input DA	Inform (name = thep phanom thai restaurant; address = 400 waller street; phone = 4154312526; postcode = 94117)
Reference	thep phanom thai restaurant s address is ***400 waller street*** , its phone number is 4154312526 , and the postcode is 94117.
SC-GPT	thep phanom thai restaurant 's phone number is 4154312526 and the postcode is 94117. ||| ***[address = 400 waller street]***
DSPM-NLG	thep phanom thai restaurant is located at ***400 waller street*** , postcode is 94117 , and the phone number is 4154312526.

Figure 5: Examples of generated responses from different models on FEWSHOTWOZ. Three different types of errors correspond to three colors (better viewed in color). The blue text means *Wrong*. The green text denotes *Redundant*. And the red text indicates *Omissive*.

Model	**$\lambda = 0$**		**$\lambda = 0.1$**		**$\lambda = 0.01$**		**$\lambda = 0.001$**	
	BLEU	ERR	BLEU	ERR	BLEU	ERR	BLEU	ERR
DSPM-NLG	34.08	6.08	**38.72**	**3.76**	35.73	4.63	34.6	5.75

Table 6: Valid BLEU and ERR with reference to λ.

serious. Compared with the baseline model, three types of errors of the DSPM-NLG model reduces "1.55%", "2.33%", and "7.75%", respectively. The experimental results reflect that our model effectively alleviates three types of errors in conveying dialogue acts. In particular, for the err of *omissive* slot-value pairs, the error rate of DSPM-NLG dropped significantly. The main reason may be that the joint probability between SLU and NLG constrains the model to accurately convey the semantic information of the dialogue act. In addition, the slot-masked strategy contributes to the reduction of *wrong* slot-value pairs. When these errors are reduced, ERR is reduced and the BLEU score is improved. The experimental results demonstrate that the DSPM-NLG model has a stronger semantic control ability than the baseline model.

Effects of λ. In the dual supervised learning framework, the Lagrange parameter λ setting greatly affects the model. Therefore, a sensitivity analysis of the λ is conducted. As shown in Table 6, we set λ and report the performance of different λ. From the result, $\lambda = 0.1$ is the optimal value for obtaining the best performance based on the dataset. When the value of $\lambda = 0$, the training of the model is the standard supervised learning process. We can see that, within a relatively large interval of λ, the performance of dual supervised learning is stronger than that of standard supervised learning.

7 Conclusion

In this paper, we proposed a novel dual supervised pre-trained model for NLG. We explore the duality between SLU and NLG from the perspective of joint probability in the pre-training stage. The slot-masked strategy is designed to constrain the DSPM-NLG model to focus on the slot-value pairs in dialogue acts. Thus, the proposed model endows the NLG module with strong semantically controlling and generalization abilities. Experiments on two benchmark datasets show significant improvement over previous state-of-the-art models in both automatic and human evaluations.

Acknowledgement

This research is substantially supported by the Key Research and Development Program of Hubei Province (2020BAB017), and the Institute for Scientific Research Center Program of National Language Commission (ZDI135-135), and the Institute for Infocomm Research of A*STAR (CR-2021-001). This research is also supported by the China Scholarship Council (202106770034).

Limitations

In the pre-training stage, the performance of DSPM-NLG depends on a large amount of annotated data. Despite the improved result, the annotated data is directly obtained from existing publicly available datasets, which has two main limitations: limited data volume and lack of data diversity. This renders limited scalability performance when dealing with complex tasks. When the data volume and diversity of the annotated data are rich enough, DSPM-NLG can fully learn the joint probability and mapping between dual tasks. Compared with the baseline model, the semantic controllability and generalization ability of DSPM-NLG will be improved more significantly.

References

Ben Athiwaratkun, Cicero Nogueira dos Santos, Jason Krone, and Bing Xiang. 2020. Augmented natural language for generative sequence labeling. *arXiv preprint arXiv:2009.13272*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Paweł Budzianowski and Ivan Vulić. 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*.

Ernie Chang, Vera Demberg, and Alex Marin. 2021. Jointly improving language understanding and generation with quality-weighted weak supervision of automatic labeling. *arXiv preprint arXiv:2102.03551*.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. *arXiv preprint arXiv:1905.12866*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Ondřej Dušek and Filip Jurčíček. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. *arXiv preprint arXiv:1606.05491*.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. *arXiv preprint arXiv:1903.09722*.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tür. 2019. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. *CoRR*, abs/1907.01669.

Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational ai. *Foundations and trends® in information retrieval*, 13(2-3):127–298.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using gpt-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.

Mihir Kale and Abhinav Rastogi. 2020. Template guided text generation for task-oriented dialogue. *arXiv preprint arXiv:2004.15006*.

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 *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 704–710, Montreal, Quebec, Canada. Association for Computational Linguistics.

Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul Crook, Bing Liu, Zhou Yu, Eunjoon Cho, and Zhiguang Wang. 2020. Continual learning in task-oriented dialogue systems. *arXiv preprint arXiv:2012.15504*.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Buildingtask bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. *arXiv preprint arXiv:2002.12328*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

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

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 79–86, Barcelona, Spain.

Shang-Yu Su, Yung-Sung Chuang, and Yun-Nung Chen. 2020a. Dual inference for improving language understanding and generation. *arXiv preprint arXiv:2010.04246*.

Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen. 2019. Dual supervised learning for natural language understanding and generation. *arXiv preprint arXiv:1905.06196*.

Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen. 2020b. Towards unsupervised language understanding and generation by joint dual learning. *arXiv preprint arXiv:2004.14710*.

Shang-Yu Su, Kai-Ling Lo, Yi-Ting Yeh, and Yun-Nung Chen. 2018. Natural language generation by hierarchical decoding with linguistic patterns. *arXiv preprint arXiv:1808.02747*.

Van-Khanh Tran and Le-Minh Nguyen. 2017. Neural-based natural language generation in dialogue using rnn encoder-decoder with semantic aggregation. *arXiv preprint arXiv:1706.06714*.

Bo-Hsiang Tseng, Jianpeng Cheng, Yimai Fang, and David Vandyke. 2020. A generative model for joint natural language understanding and generation. *arXiv preprint arXiv:2006.07499*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019a. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019b. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Qingyang Wu, Yichi Zhang, Yu Li, and Zhou Yu. 2019. Alternating recurrent dialog model with large-scale pre-trained language models. *arXiv preprint arXiv:1910.03756*.

Xinnuo Xu, Guoyin Wang, Young-Bum Kim, and Sungjin Lee. 2021. AUGNLG: few-shot natural language generation using self-trained data augmentation. *CoRR*, abs/2106.05589.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2020. Ubar: Towards fully end-to-end task-oriented dialog systems with gpt-2. *arXiv preprint arXiv:2012.03539*.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.

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.

Su Zhu, Ruisheng Cao, and Kai Yu. 2020. Dual learning for semi-supervised natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1936–1947.

A Data Statistics

Statistics	FEWSHOTWOZ	FEWSHOTSGD
# Domains	7	16
Avg. # Intents	8.14	6.44
Avg. # Slots	16.2	11.3
Avg. # Training Instances	50	35
Avg. # Test Instances	473	5618

Table 7: Data Statistics of two datasets.

B Experiment Setup

Using the Huggingface Transformers public library (Wolf et al., 2019a), we implement our model on PyTorch. The GPT-2-Medium model with 24 layers and 16 attention heads is chosen as the backbone, and byte pair encodings (Sennrich et al., 2015) is used for the tokenization. And the model uses Adam (Kingma and Ba, 2014) as the optimizer with an initial learning rate of 5e-5, a scheduler with a linear warm-up to update and adjust the learning rate. We set the maximum sequence length to 80 and the batch size to 8. The GPU used for the training is NVIDIA Quadro RTX 8000-64G. In the pre-training stage, we jointly (SLU and NLG) train GPT-2 until observing no obvious improvement in validation loss or up to 20 epochs. And we save the model parameters for the fine-tuning stage.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4

☑ B1. Did you cite the creators of artifacts you used?
Section 4

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4

C ☑ Did you run computational experiments?

Appendix A

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
n Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
n Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.