UniPCM: Universal Pre-trained Conversation Model with Task-aware Automatic Prompt

Yucheng Cai [1], **Wentao Ma** [2], **Yuchuan Wu** [2], **Shuzheng Si** [3], **Yuan Shao** [2],
Zhijian Ou [1] , **Yongbin Li** [2,†]

[1]Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University,
[2]Alibaba Group, [3]Peking University
cyc22@mails.tsinghua.edu.cn, ozj@tsinghua.edu.cn, {mawentao.mwt}@alibaba-inc.com

Abstract

Recent researches have shown that multi-task instruction tuning after pre-training greatly improves the model's robustness and transfer ability, which is crucial for building a high-quality dialog system. However, most previous works on multi-task instruction tuning rely heavily on human-defined input format or prompt, which is not optimal in quality and quantity.In this work, we propose to use **T**ask-aware **A**utomatic **P**rompt generation (TAP) to automatically generate high-quality prompts. Using the high-quality prompts generated, we scale the corpus of the pre-trained conversation model to 122 datasets from 15 dialog-related tasks, resulting in **Uni**versal **P**re-trained **C**onversation **M**odel (**UniPCM**), a powerful foundation model for various conversational tasks and different dialog systems. Extensive experiments have shown that UniPCM is robust to input prompts and capable of various dialog-related tasks. Moreover, UniPCM has strong transfer ability and excels at low resource scenarios, achieving SOTA results on 9 different datasets ranging from task-oriented dialog to open-domain conversation. Furthermore, we are amazed to find that TAP can generate prompts on par with those collected with crowdsourcing.

Keywords: Automatic prompt generation, Multi-task training, Pre-trained conversation model, Dialog system

1. Introduction

Dialogue system modeling, a classic research topic in the field of human-machine interaction, serves as an important application area (Si et al., 2022a; Lin et al., 2020b; Song et al., 2023; Dang et al., 2024). The advancements in dialogue systems for those applications have been significantly boosted by the use of pre-trained language models (PLMs), including GPT-2 (Radford et al., 2019) , BERT (Kenton and Toutanova, 2019), and T5 (Raffel et al., 2020) , combined with task-specific fine-tuning on annotated data (Hosseini-Asl et al., 2020; Yang et al., 2021; Heck et al., 2020; Lee, 2021; Liu et al., 2022; Zhao et al., 2023; Liu et al., 2023). However, most of the models trained under the 'pretrain-finetune' paradigm are limited to specific tasks or datasets, and the dialog systems built upon those models can only respond to certain input formats, which lacks robustness and transfer ability.

To relieve such problems, multi-task training, which has achieved great success in language model pre-training, has been introduced to pre-trained conversation models (PCM), which performs multi-task instruction tuning on a pre-trained language model basis. Recent progress in multi-task pre-training (Ouyang et al., 2022; Sanh et al., 2022; Mishra et al., 2022; Wang et al., 2022) has shown that the robustness and transfer ability of language models are greatly improved by pre-training with multiple tasks.

However, the previous works on multi-task training rely heavily on human-defined input format or prompt. We find those artificially constructed prompts still have two major weaknesses that can be harmful when used to train a language model:

(1) **Human labor required and limited in quantity.** Previous works like Supernatural-instruction (Wang et al., 2022) only have one task instruction for each dataset. Using only one prompt for a task to train the model may cause overfitting to the prompt and hinder the model to catch the essence of the tasks and transfer to other prompts. (2) **Hard to understand and limited in quality.** Human labelers easily incorporate their own understandings into the prompts or simply obtain the prompt by paraphrasing the dataset description, which makes the prompts quite long and unnatural, containing specific knowledge of the datasets.

To address the problems mentioned above, we propose a task-aware auotmatic prompt generation method TAP. The TAP method can generate numerous high-quality prompts automatically, which relieves human labor and improves the quality and quantity of the prompts as well. Those high-quality prompts can greatly help train a universal pre-trained conversational model. With the high-qualit prompts, the model can learn better on the larger pre-training datasets and understand each task better. Using the 303 high-quality prompts automatically generated for the 15 tasks, we scale our pre-training corpus to 122 datasets and 26,625,486 instances, which, to our knowledge, is currently the

---

† Corresponding author

largest annotated conversational dataset, covering almost all dialog-related tasks. Moreover, we propose a multi-prompt training mechanism to make use of the generated prompts to better fuse different tasks. The resulting model UniPCM is a powerful foundation model for various conversational tasks and different dialog systems.

Through comprehensive experiments in multiple conversational tasks, we find that UniPCM has strong ability on various dialog tasks, which outperforms the T5 baseline by 7.14% in the few-shot setting on DialoGLUE (Mehri et al., 2020) and achieves SOTA results on 9 different datasets ranging from task-oriented dialog to open-domain conversation. Moreover, the model is robust to input format and can respond to different input prompts. Furthermore, to comprehensively evaluate the quality of the generated prompts, we conduct human evaluation and automatic evaluation. Our generated prompts achieve higher average scores than human-written prompts by 9.50% on three proposed metrics in human evaluation and improve by 2.40% when used to finetune downstream tasks.

In summary, our main contributions are:

- We propose a task-aware automatic prompt generation method TAP to better fuse the datasets from different tasks in the multi-task training, which can generate numerous high-quality prompts based on extracted task information. The proposed method greatly reduces human effort in prompt engineering and improves the quality of generated prompts.

- Leveraging the high-quality prompts generated, we pre-train a unified pre-trained conversation model UniPCM by scaling the pre-training datasets to 122 dialog-related datasets from 15 dialog-related tasks, resulting in a powerful conversation model UniPCM. The pre-trained model and the datasets collected will be released to public.

- We conduct extensive experiments on 10 dialog-related benchmarks including 6 types of task. Results on few-shot and full-data experiments show the superiority of our proposed method and model. The code is released at https://github.com/cycrab/unipcm.

2. Related Work

2.1. Multi-task Language Model Training

Recent researches have shown that multi-task language model training can greatly improve the model's transfer ability, resulting improved performance in few-shot and zero-shot settings (Raffel et al., 2020; Wei et al., 2021; Sanh et al., 2022; Si et al., 2023b, 2024). Although negative transfer may occur when the number of tasks is limited, the model will still benefit from training if scaling the number of task (Aribandi et al., 2021).

To implement multi-task training, some signals are given to the model to distinguish one task from another. Initially, researchers do multi-task pre-training using a unified text to text format directly (Raffel et al., 2020; Lu et al., 2022). Simply adding the name of the task will help the model better understand the relation between task and reduce negative transfer problem (Zhang et al., 2018b). Recent works use crowdsourcing prompts and instructions to perform multi-task pre-training, which achieves great success (Sanh et al., 2022; Wang et al., 2022; Ouyang et al., 2022). The resulting models show strong transfer ability, and can even chat with humans fluently in open domain (Ouyang et al., 2022).

Our work improves over the previous works in that we use automatically generated prompts instead of the crowdsourcing ones to enable multi-tasking, which reduces human labor as well as improves the quality of the prompts. Furthermore, we propose and formulate multi-prompt training mechanism, which relieves several problems in multi-task pre-training, including task imbalance, uneven data quality and difference between the importance of tasks. Moreover, we prove that multi-prompt training can improve model's performance on unseen prompts.

2.2. Automatic Prompt Generation

It has been shown that prompt engineering can be of great benefit to reduce the gap between language model pre-training and finetuning on downstream tasks (Gao et al., 2021; Zhong et al., 2021). To reduce human labor in prompt engineering, various approaches have been proposed to generate prompts automatically. AutoPrompt (Shin et al., 2020) use gradient-based prompt search to automatically generate prompts. However, the prompt generated are not coherent, and may confuse models in multi-task scenarios. Researchers proposed in (Gao et al., 2021; Zhou et al., 2022) to use T5 (Raffel et al., 2020) or large language model to fill in the blank bewteent the input and output for automatically generating coherent prompts. However, the prompts generated do not necessarily contain task information and may be highly related to certain input case or dataset. Different from previous works, our work aims to generating prompts for multi-task pre-training. Therefore, our method TAP models task in automatic prompt generation to help the model understand the relation between the tasks and the prompts, as well as improve the quality of generated prompts.

2.3. Pre-training for Dialog Systems

The PLMs have shown strong capabilities in many NLP tasks (Si et al., 2022a, 2023b; Zhao et al., 2023; Hu et al., 2023; Li et al., 2024; Si et al., 2022b, 2023a). It has been shown that pre-training can
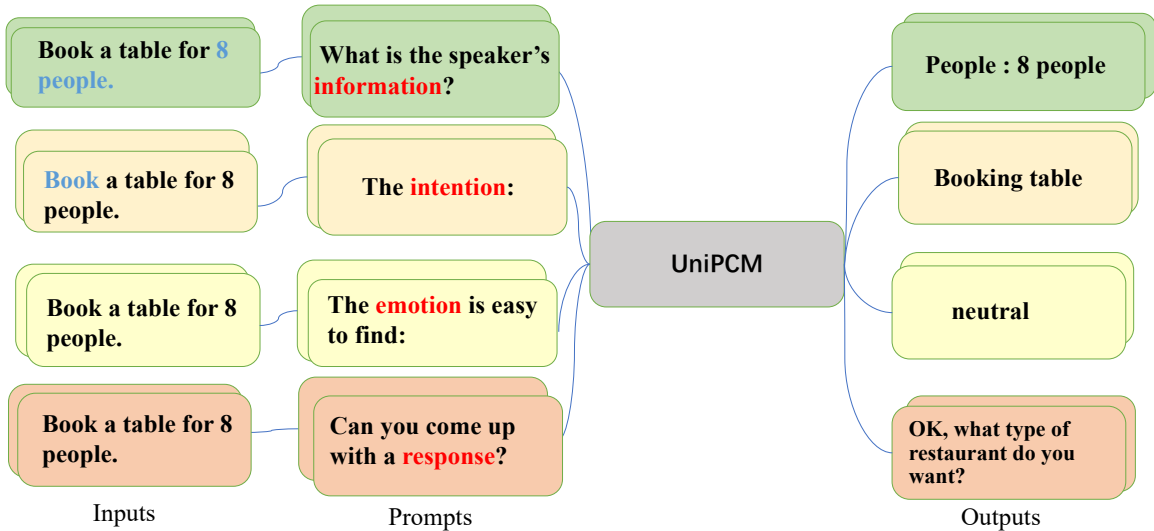
Figure 1: An illustration of our proposed model UniPCM, which unifies all tasks into an 'input-prompt-output' format. Prompts are crucial as they help the model understand the task it should perform.

greatly improve performance for dialog systems, improving coherency of the generated response and transfer ability (Roller et al., 2021; Zhang et al., 2020b; Su et al., 2022). Models trained on large-scale online open-domain dialogues, for example, BlenderBot (Roller et al., 2021), DialoGPT (Zhang et al., 2020b) and Meena (Adiwardana et al., 2020), can perform well on the chit-chat task, while models pre-trained on certain tasks can improve performance on corresponding tasks. For example, in task-oriented dialog, works like TOD-BERT (Wu et al., 2020), CONV-BERT (Mehri et al., 2020), PP-TOD (Su et al., 2022), GALAXY (He et al., 2021) improve the performance on relevant datasets.

However, to interact with human fluently in open-domain (Ouyang et al., 2022), the dialog system should not only be capable of various tasks, but also be robust to different input prompts. Recent progress in building powerful open-domain dialog systems mainly used crowdsourcing to annotate large-scale, multi-task datasets to improve the systems' performance (Shuster et al., 2022; Ouyang et al., 2022). Different from their approaches, we propose to leverage the existing large scale datasets that are dialog-related to perform multi-task pre-training. Chen et al. (2022b) also trains their dialog foundation model over large scale dialog-related datasets. However, they do not aim to building a dialog system, therefore they do not improve their model's robustness to input prompts, neither do they evaluate their model's transfer ability in few-shot or zero-shot scenarios. In contrast, we use generated prompts to perform multi-task prompt pre-training to improve the model's transfer ability and robustness to different input prompts.

3. Method

To train our UniPCM, we first unify all the dialog-related tasks into an 'input-prompt-output' format,

which is shown in Figure 1. Then we propose task-aware automatic prompt generation TAP to generate high-quality prompts. Finally, based on the prompts and corpus, we train our UniPCM using the proposed multi-prompt training mechanism.

3.1. Task-aware Automatic Prompt Generation

Our task-aware automatic prompt generation method TAP, as illustrated in Figure 2, mainly contains the following 3 parts:

3.1.1. Finding Signals for Task Information

Before generating prompts, we extract task-related signals to help us to find the information about the task t. In this work, we mainly focus on 3 kinds of signals that can be used as hint of the task for the model to generate prompts upon. We discuss their availability, limitation and effectiveness to generate prompts.

Instructions: Task descriptions, or instructions are usually available for datasets. Moreover, huge amounts of instructions are annotated or collected by researchers or crowdsourcing workers (Wang et al., 2022; Ouyang et al., 2022). Instructions are usually long and difficult for language model to understand directly, therefore it's hard to directly use them as input to generate prompts in an unsupervised way. However, instructions contain almost all important information for the task and it's not hard to extract key information from it.

In our work, we use tf-idf methods to filter out irrelevant words. [1] Then we get all the 1-gram, 2-grams and 3-grams of the remaining words to get scored by a Bert model (Devlin et al., 2019) using their similarity with the task name. The words that

---

[1] We inplemented tf-idf using the gensim package: https://radimrehurek.com/gensim/
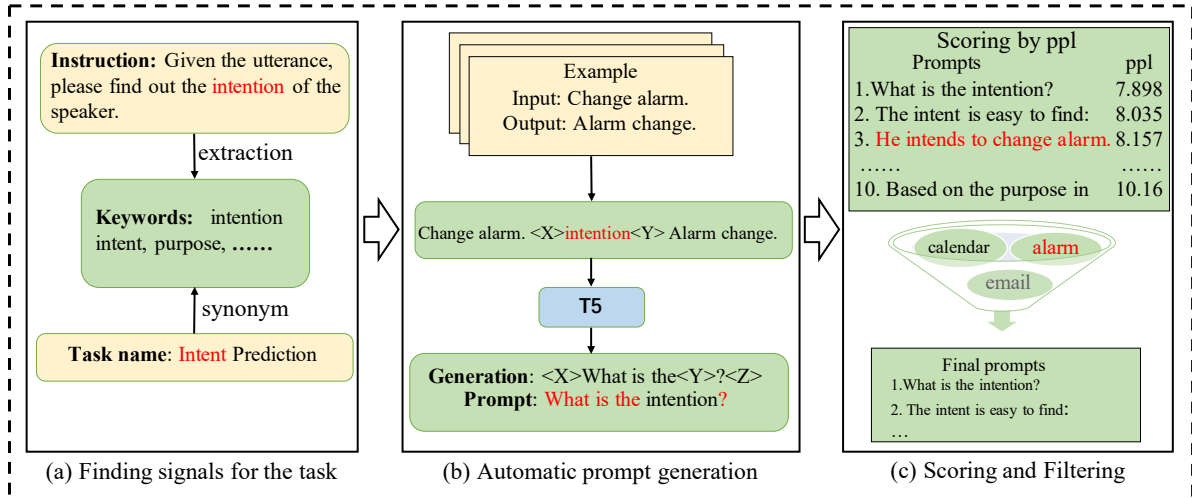
17044

Figure 2: An illustration of our proposed method TAP. (a) We collect existing signals to generate keywords, either by extracting from instructions or searching synonyms of the task name. (b) We automatically generate prompts based on input examples and task-related keywords using a T5 model, harvesting prompts by concatenating generated words after the sentinel token. (c)We select the prompts by average perplexity on task-related examples and filter out those prompts that may contain information about the labels.

have similarity score above a threshold are deemed as keyword and used to generate prompts.

Task Name: Task name is always available for task and very concise, ideal for automatic prompt generation. However, one task can only have one task name, making it difficult to generate diverse prompts. Therefore, we propose to use the thesaurus tool to paraphrase the task name to form diverse key words. Also, the task name is used to select the keywords extracted from the instruction, which has already been discussed.

Keywords: Keywords are ideal input for automatic prompt generation, as keywords are both concise and representitive of the task information. However, keywords are not readily available and should be inferred from other task-related information like instructions or the task name. If the quality or the number of the keywords generated by the instructions or the task name do not meet the researchers need, researchers can quickly summarize the task and write some high-quality keywords themselves.

3.1.2. Automatic Prompt Generation Using Keywords

Getting the task signals (in the form of keywords in this work), we can generate prompts automatically using a pre-trained language model T5 (Raffel et al., 2020). T5 is pre-trained to fill missing spans for a sentence. For example, given the input "Thank you <X> me to your party <Y> week", T5 is trained to generate "<X> for inviting <Y> last <Z>", meaning that "for inviting" replaces the placeholder <X> and "last" replaces the placeholder <Y>. This is well suited for prompt generation, as we want to generate a prompt with the keywords that is coherent

given input and output.

Given an instance of input-output pair (X_t, Y_t) in task t, along with one of the keywords k_t^i, we define a transform $\mathcal{T}(X_t, Y_t, k_t^i)$[2]:

$$X_t, Y_t, k_t^i \rightarrow X_t \langle X \rangle k_t^i \langle Y \rangle Y_t \qquad (1)$$

where $\langle X \rangle, \langle Y \rangle$ are sentinel tokens for T5 generation. We generate the prompts according to the T5 generation probability $P_{T5}(\mathcal{T}(X_t, Y_t, k_t^i))$ and harvest the prompts generated after the sentinel tokens. The final prompts are reorganized as $P_t = x \oplus k_t^i \oplus y$, where $\oplus$ denotes the concatenation of token sequence and x, y are the corresponding content generated after sentinel tokens $\langle X \rangle, \langle Y \rangle$ by the T5 model.

For one single input-output instance and one keyword, we get the top 5 prompts according to generation probability. Using multiple instances and multiple keywords, we generate numerous prompts for selection. To avoid generating prompts that are specific to one single input instance and overlook the task information, we only retain the prompts that appear multiple times, which is empirically set as 2 in our work, in different instances.

3.1.3. Scoring and Filtering of Generated Prompts

Given input X_t and prompt P_t in task t, the probability of generating the correct answer Y_t: $p(Y_t|X_t, P_t)$ should be optimized, therefore we evaluate the

---

[2]In fact, multiple reasonable transforms can be defined. In our experiments, we use the transform $\tilde{\mathcal{T}}(X_t, Y_t, k_t^i) = X_t, Y_t, k_t^i \rightarrow \langle X \rangle X_t \langle Y \rangle k_t^i \langle Z \rangle Y_t$, along with the transform T introduced in the paper to better model tasks that need a prefix like question answering.

quality of the generated prompts according to the probability, as all of Y_t, X_t, P_t are available and the probability can be directly calculated. We choose those prompts P_t that has higher average log probability $\sum \log(p(Y_t|X_t, P_t))$ among the datasets in task.

After that we filter out prompts that may contain certain biased information about the output Y_t using a prohibited words list extracted from the outputs. The prohibited words mainly fall into the classification type class, as the output Y_t is selected from certain labels. For example, in emotion classification task, the word "positive" or "negative" is often generated by the model, as the output Y_t contains those two words with high frequency. However, using such prompts for output generation, the result will be biased, reducing the generated performance. Therefore, we filter out those case-sensitive prompts to encourage those prompts that accurately reflect the task.

3.2. Multi-task Prompt Training

Using the generated prompts, as well as the collected corpus, we perform multi-task prompt training. With the training instances $(X_t^i, Y_t^i)(i = 1, 2, \cdots, N_t)$ from K different tasks and the generated prompts $P_t^j (j = 1, 2, \cdots, M_t)$, the training objective function can be written as:

$$\mathcal{J}_\theta = \sum_{t=1}^{K} \sum_{i=1}^{N_t} \sum_{j=1}^{M_t} \log p(Y_t^i|X_t^i, P_t^j) \quad (2)$$

Note that in Eq. 2, we propose to use multi-prompt training, which means applying multiple prompts to one single input instance: $\sum_{j=1}^{M_t} \log p(Y_t^i|X_t^i, P_t^j)$.

However it's not necessary to apply all the M_t prompts available to one single case, as the prompts $P_t^j (j = 1, 2, \cdots, M_t)$ are representation of the task t and have similar embeddings in the latent task space. Therefore a subset of P_t^j can be randomly sampled, resulting in $\tilde{P}_t^j (j = 1, 2, \cdots, \tilde{M}_t)$. The loss $\sum_{j=1}^{M_t} \log p(Y_t^i|X_t^i, P_t^j)$ can be approximated by $\frac{M_t}{\tilde{M}_t} \sum_{j=1}^{\tilde{M}_t} \log p(Y_t^i|X_t^i, \tilde{P}_t^j)$ to save calculation time.

If the ratio $\frac{M_t}{\tilde{M}_t}$ is not added, we can simply adjusting the weights of datasets or tasks in training by adjusting the number of prompts applied. It is beneficial as some tasks or datasets are deemed more important by the researchers. Adding more prompts to those tasks or datasets can make the model better focus on them.

4. Experiments

4.1. Baseline&Benchmark

To comprehensively evaluate our UniPCM, we carefully choose ten downstream datasets in six tasks, mainly evaluating the model's ability in dialog understanding, response generation, and comprehensive ability.

4.1.1. Dialog understanding

Dialog understanding is crucial for building a high-quality dialog system as it's impossible to generate high-quality responses without having a good understanding of the context. DialoGLUE (Mehri et al., 2020) is a benchmark that comprehensively evaluates the dialogue understanding ability of a dialog system, which consists of four tasks: slot filling (REST8K (Coope et al., 2020), DSTC8 (Rastogi et al., 2020)), intent prediction (BANKING77 (Casanueva et al., 2020), CLINC150 (Larson et al., 2019), HWU64 (Liu et al., 2021a)), semantic parsing (TOP (Gupta et al., 2018)), and dialog state tracking (MultiWOZ2.1 (Eric et al., 2020)). We follow the original preprocessing and evaluating scripts of Mehri et al. (2020), except that we modify the implementation to a sequence-to-sequence generation format to fit the model's training. The evaluation metrics for slot filling, intent prediction and semantic parsing are F1, accuracy and exact match respectively. For dialog state tracking task of Multiwoz, we apply our model to the SOTA generative baseline SDP-DST (Lee et al., 2021) and joint goal accuracy (JGA) is reported.Apart from T5 (Raffel et al., 2020) (we trained our model upon a T5-base model), we choose SPACE-2 (He et al., 2022a) and Flan-T5 (Chung et al., 2022) as our baselines, as SPACE-2 represent the SOTA pre-trained model targeting task-oriented dialog understanding, while Flan is the most prevalent instruction-tuning method used for multi-task training. The results of TOD-BERT (Wu et al., 2020) and the best variant of ConvBERT (Mehri et al., 2020) are also reported for comparison.

4.1.2. Response Generation

Open-domain response generation, or chit-chat, is also an important skill for building a high-quality dialog system. We evaluate our model on two classic chit-chat datasets PersonaChat (Zhang et al., 2018a) and DailyDialog (Li et al., 2017). We follow the preprocessing and evaluation scripts of FSB (Madotto et al., 2021), BLEU (Papineni et al., 2002), word-level F1 and Rouge-L (Lin, 2004) reported. We choose DialoGPT (Zhang et al., 2020b) and PPTOD (Su et al., 2022) as our baseline.

4.1.3. Comprehensive ability

We evaluate the comprehensive ability of a dialog system on the Multiwoz end to end generation task (End2End) (Budzianowski et al., 2018). In End2End task, the model needs to track the use's state, understand user's intention, decide the best responding strategies and generate coherent response, which is quite challenging. Multiple dialog

skills, such as intent prediction, dialog state tracking, policy optimization, and response generation, are necessary to complete the task. We apply our model to the SOTA method MTTOD (Lee, 2021) and use the official evaluation scripts [3] given by (Nekvinda and Dušek, 2021). We compare our results to LABES (Zhang et al., 2020a), SOLOIST (Peng et al., 2021), UBAR (Yang et al., 2021) and PPTOD (Su et al., 2022).

4.2. Implementation
4.2.1. Pre-training

We pre-train our conversation model UniPCM on the collected corpus UniPreDial. The statistics of the tasks and datasets, as well as the generated prompts, are shown in Table 4. The maximum sequence length of input context is set to 256. The batch size is set to 64 and an AdamW optimizer is employed for optimization with a constant learning rate of 2e-5. The pre-training is performed on eight 80GB Tesla-A100 GPU and takes about 72 hours.

4.2.2. Downstream tasks

For downstream tasks, we finetune UniPCM following the corresponding baseline scripts. For each few-shot and zero-shot experiment, we exclude the training data other than the few-shot data in the pre-training datasets accordingly to avoid unfair data use. During testing, we test the model with 5 random prompts sampled from all the available prompts for each testinig instance (if the prompts are used). We view the results as 5 independent experiments and the mean result of the performance is reported as the final result. The variance of the experiment is reduced as we take the mean results of 5 experiments. Moreover, to achieve high score under this testing setting, the model needs to perform well on all the available prompts. The resulting high performance proves that our model is robust to input prompts.

4.3. Main Results

We conduct our experiments on the baseline and benchmarks mentioned above. The implementation detail is shown in Sec. 4.2.

4.3.1. DialoGLUE Results

As shown in Table 1, our model UniPCM excels at few-shot setting, improving **7.14%** on average scores over the T5 baseline, achieving SOTA results on all 7 datasets of DialoGLUE and improve by **1.75%** over the previous SOTA result SPACE-2 on average scores.

For the full data setting, our model is competitive, achieving the best average scores among the strong baselines and consistently outperforms Flan-T5 on all datasets, which demonstrate the efficacy

of our pre-training methods. It is worth noticing that SPACE-2 performs quite well on this task, which is mainly because its TOD targeted modelling, which makes the model restricted to understanding task in TOD datasets.

4.3.2. MultiWOZ2.0 End2End Results

As shown in Table 2, our model UniPCM improves over the previous SOTA model MTTOD in both full data and few shot scenarios by 1.6 and 2.2 on combined score respectively. The model's improvements mainly fall in the Inform and Success, implying that the pre-training improves the model's dialog understanding and decision-making ability. Meanwhile, the few-shot improvements are not so remarkable as in DialoGLUE datasets, probably resulting from the delexicalization preprocessing used in MultiWOZ (Zhang et al., 2020b), making the language used in this dataset slightly different from those in other datasets in pre-training.

4.3.3. Chit-chat Results

As shown in Table 3, UniPCM consistently improves over all of the baseline results in zero-shot and few-shot settings in Persona and DailyDialog. The results imply that combining open-domain chat datasets in the multi-task pre-training procedure will improve the model's ability to perform open-domain chatting. Meanwhile, the performance of PPTOD, a model trained on task-oriented dialog datasets only, does not improve over the T5 baseline on chit-chat tasks, which shows the importance of combining open-domain chit-chat tasks in pre-training.

4.4. Analysis and Ablation Study
4.4.1. Ablation study for UniPCM in few-shot setting

It has been shown in Table 1 that UniPCM excels at few-shot setting, and we want to have a full understanding of why UniPCM achieves great performance in few-shot setting. We get three main conclusions from the ablation study shown in Table 5: **(1) Using multi-prompt training in finetuning stage greatly helps the model's performance in few-shot setting**, achieving 2.98% gain. **(2) Using multi-prompt training in pre-training stage will help the model learn better in multi-task scenario.** Although using one human-written prompt in the pretraining stage will help improve the dialog understanding ability by 1.20%, by using multi-prompt training in the pre-training stage, the results improve by **3.08%**, which shows that using multi-prompt training in the pre-training stage will greatly benefit the model's performance in downstream task. **(3) PET will help in low-resource setting.** Adding PET (Schick and Schütze, 2021a,b), which

---

Setting	Model	avg	BANKING77	HWU64	CLINIC150	REST8K	DSTC8 [†]	TOP	MULTIWOZ
10-shot data	T5	76.52	76.01	81.77	88.36	85.31	74.72	76.03	51.63
	TOD-BERT *	79.96	85.99	86.74	93.07	87.62	50.19	77.77	48.54
	ConvBERT *	78.72	85.06	85.69	93.06	87.58	44.36	72.01	48.89
	SPACE-2 *	81.91	88.31	88.85	95.22	88.85	54.41	79.55	50.70
	Flan-T5	80.68	84.48	86.88	91.80	90.59	78.68	76.78	53.52
	UniPCM	**83.66**	**90.16**	**90.05**	**95.78**	**92.62**	**83.27**	**79.63**	**53.73**
Full data	T5	85.70	92.60	91.07	96.49	95.95	93.60	81.41	56.66
	TOD-BERT*	85.43	93.02	89.87	95.93	95.53	90.05	81.90	56.30
	ConvBERT*	86.17	93.44	92.38	97.11	95.44	91.20	82.08	56.56
	SPACE-2*	87.56	**94.77**	**94.33**	**97.80**	96.20	91.38	82.74	**59.51**
	Flan-T5	86.99	93.47	92.37	96.71	96.41	94.51	84.32	58.68
	UniPCM	**87.59**	94.41	93.40	97.47	**96.92**	**96.15**	**84.58**	58.76

Table 1: Results of seven datasets from the DialoGLUE benchmark in low-resource and full data setting. * denotes the model is specified for understanding task in TOD only. [†] denotes we fix a bug in the original scripts, resulting higher score in DSTC8 dataset and we exclude the dataset in the avg score for fair comparison.

MultiWOZ2.0 End2End					
Setting	Model	Inform	Success	BLEU	Combined score
Full data	LABES	68.5	58.1	18.9	82.2
	SOLOIST	82.3	72.4	13.6	90.9
	UBAR	83.4	70.3	17.6	94.4
	PPTOD	83.1	72.7	18.2	96.1
	MTTOD	85.9	76.5	19.0	100.2
	UniPCM (ours)	**88.3**	**76.8**	**19.2**	**101.8**
Few shot(10%)	MTTOD	66.8	52.8	**15.7**	75.5
	UniPCM (ours)	**68.4**	**57.2**	14.9	**77.7**

Table 2: Full data and few-shot results on Multiwoz2.0 End2End task, inform, success, BLEU and combined score are reported.

Model Configuration		Persona			DailyDialog		
Setting	Model	BLEU	F1	Rouge-L	BLEU	F1	Rouge-L
Zero-shot	T5 (baseline)	0.94	15.24	9.16	0.29	9.76	8.51
	PPTOD	0.70	13.83	10.74	0.39	10.44	10.14
	DialoGPT	0.57	9.61	11.83	0.45	15.18	18.99
	UniPCM (ours)	**1.15**	**16.45**	**18.25**	**0.85**	**17.81**	**21.04**
Few-shot(10%)	T5 (baseline)	1.76	17.18	18.14	0.53	12.62	16.44
	PPTOD	1.85	17.44	17.75	0.39	14.58	17.65
	DialoGPT	1.23	14.74	18.39	0.77	16.35	18.16
	UniPCM (ours)	**2.41**	**19.16**	**18.81**	0.81	**18.04**	**21.23**

Table 3: Few-shot and Zero-shot results on Personachat and DailyDialog dataset (task: chit-chat). BLEU, word-level F1 and Rouge-L are reported.

leverages the generated prompts to perform semi-supervised training, improves by 1.08% over the strong baseline, which shows that our generated prompts can help model better utilize unlabeled data by using PET.

4.4.2. Finetuning with multiple prompts.

Although we have shown in Table 1 that multi-prompt training will greatly improve the model's performance of finetuning in few-shot setting, it is not clear how the number of prompts available will influence the final results. From Table 6, we can see that simply applying 1 prompt will increase by 2.306% on test accuracy. Moreover, applying a small number of prompts (7) can greatly improve the test accuracy (4.643%). To manually select prompts that are deemed better by human experts will not help much (0.323%). Moreover, using a

large number of prompts (25) will improve a little over fewer prompts result (0.811%). Therefore in PET, we propose to use subsets of prompts to fine-tune the voting models, which will yield the best performance.

4.5. Automatically generated Prompts

Using the 494 keywords extracted from the Super-Instruction datasets (Wang et al., 2022), we generate 3423 prompts on 74 tasks. However, as our work mainly focus on pre-training a conversation model, we mainly evaluate the 303 prompts used in pre-training. The rest of generated prompts will be released with our codes and can be further studied.

4.5.1. Visualization of Generated Prompts

To better understand the prompts distribution in the latent space, we visualize the embeddings of the prompts generated using t-SNE visualization (Van der Maaten and Hinton, 2008). As illustrated in Figure 3, we use embeddings from language models to approximate the embeddings in the latent space, as the embeddings in the latent space are not available. The results show that our generated prompts are task-centric, yet diverse. Moreover, comparing the embeddings in our pre-trained model and T5-base model, we can see that pre-training makes the prompt embeddings of the same task cluster, meaning that the model understands the relation between tasks and prompts better after pre-training.

4.5.2. Human Evalution

We perform human evaluation to comprehensively evaluate the quality of the generated prompts. We sum up three key characteristics of good prompts: task-specific, coherency and fluency, which is defined as:

Task-specificity: whether the prompt accurately reflects the essence of the task.

Task type	Intent	Dialog state tracking	Emotion	Summary	Question answering	Generation	Response	Multiple choice	Text2sql	Grounded dialog	Total
Tasks	Intent	DST, slot filling	Emotion	Summary	DialQA, DocQA	Generation	Response, Chat	Multiple choice	Text2sql	TOD, US, KG-dial	15
Number of prompts	37	33	14	11	35	51	27	39	29	27	303
Number of datasets	22	21	7	5	12	4	23	3	2	23	122
Number of instances	1,382,413	4,382,314	171,353	449,995	460,681	198,999	16,555,894	44,992	19,059	2,959,786	26,625,486

Table 4: Statistics of tasks, datasets, and prompts in UniPreDial.

Method	avg	BANKING77	HWU64	CLINIC150	REST8K	DSTC8	TOP
T5	76.52	76.01	81.77	88.36	85.31	74.72	76.03
+ MP	79.50	83.77	85.02	91.67	88.24	79.67	76.69
+ MP + PT	80.70	83.70	85.86	92.73	90.79	80.41	78.64
+ MP + MPPT	82.58	87.92	88.74	94.76	91.55	82.87	78.75
+ MP + MPPT + PET*	83.66	90.16	90.05	95.78	92.62	83.27	79.63

Table 5: Ablation study on six datasets from the DialoGLUE benchmark in low-resource setting (10-shot data), MP means multi-prompt training in the finetuning stage, PT means pre-training, MPPT means multi-prompt training in the pre-training stage. * denotes it is the UniPCM.

Number of Prompts	0	1(avg)	7(random)	7(selected)	25
Test Acc	76.006	78.312	82.955	83.279	**83.766**

Table 6: Few-shot(10%) results on BANKING77 dataset using different numbers of prompts. For 1 prompt setting, we report the average scores of randomly selected prompt to reduce variance.

Coherency: whether the prompt can form coherent sequences with most of the inputs and outputs.

Fluency: whether the prompt itself is grammatically correct and fluent.

Experts in dialog system are asked to score 0, 1, 2 for the three metrics on the prompts generated by TAP and crowdsourcing human-written prompts randomly selected from Prompsource (Bach et al., 2022), the average scores reported in table 7. The results show that our generated prompts are superior to the crowdsourcing human-written prompts, improving the task-specificity, coherency, and fluency by **9.95%**, **10.97%**, **7.59%** respectively. Moreover, it can be shown in the results that by modeling task in TAP, the prompts generated focus on the task better, while using the input-output pairs in the automatic prompt generation procedure make the prompts generated better fit with the context, resulting in higher gain in task-specificity and coherency.

4.5.3. Results on Downstream Tasks

Besides human evaluation, we measure the quality of the prompts generated using the downstream finetuning results. A T5 model is finetuned on downstream tasks with the prompts generated using multi-prompt training. We compare our automatically generated prompts with crowdsourcing prompts from Promptsource (Bach et al., 2022). Moreover, to illustrate the importance of modeling task in TAP, we try to generate prompts without task-related information, i.e. the keywords for ablation study, which is the same as the method proposed in Gao et al. (2021). The results shown in Table 8 demonstrate the superiority of our auto-
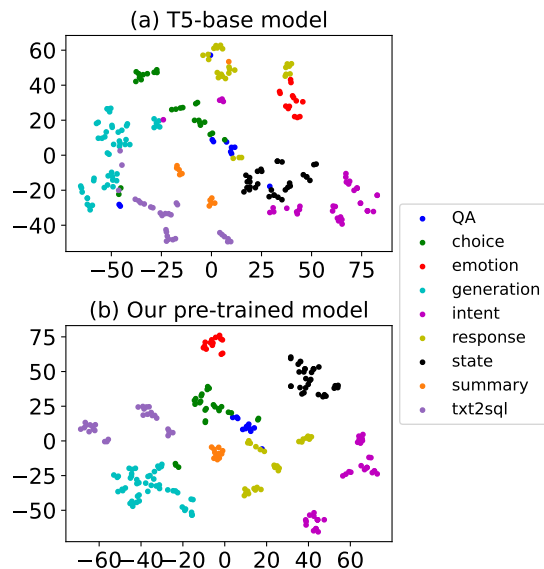

Figure 3: Prompt embeddings in the latent space using t-SNE visualization. T5-base model and our pre-trained model are used to approximate the latent space in (a) and (b) respectively.

Method	Task-specificity	Coherency	Fluency
Crowdsourcing	1.698	1.687	1.726
TAP	**1.868**	**1.871**	**1.857**
p-value	7.50e-7	7.55e-9	6.12e-5

Table 7: Human evaluation results for prompts generated by TAP and collected from Promptsource respectively. Average scores of task-specificity, coherency, and fluency is reported. p-value means the significant test result between the two methods.

matically generated prompts over human-written prompts, improving by **2.40%** on test accuracy. Meanwhile, modelling task in TAP brings an improvement of 0.89%, which shows that modeling task is beneficial for generating prompts with higher quality.

5. Conclusion and Future Work

This paper represents progress toward building high-quality dialog systems with multi-task training using automatically generated prompts. Based on a unified 'input-prompt-output' format, we generate high-quality prompts using the proposed automatic prompt generation method TAP and perform multi-task training using the proposed multi-prompt training mechanism, resulting in a powerful pre-trained conversation model UniPCM. Extensive experiments demonstrate that UniPCM is robust

Prompts	Promptsource	TAP without task	TAP
Test Acc	80.877	82.388	**83.279**

Table 8: Few-shot(10%) results on BANKING77 dataset using different sets of prompts. The size of the prompt set is set to 7 as there is only 7 prompts available from Promptsource (Bach et al., 2022)

to input prompts, capable of performing various dialog-related tasks, and has strong transfer ability, particularly in low-resource scenarios. We hope our pre-trained model UniPCM, as well as the collected datasets, will help researchers to build better dialog systems. Furthermore, since multi-task training is widely used in instruction tuning, we hope our automatic prompt generation method TAP, as well as the high-quality prompts generated, will encourage the community to further explore the limits of multi-task training.

6. Bibliographical 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*.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. 2021. Ext5: Towards extreme multi-task scaling for transfer learning. In *International Conference on Learning Representations*.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. 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 4516–4525.

Giovanni Campagna, Agata Foryciarz, Mehrad Moradshahi, and Monica Lam. 2020. Zero-shot transfer learning with synthesized data for multi-domain dialogue state tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 122–132.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional webnlg+ shared task overview and evaluation results (webnlg+ 2020). In *Proceedings of the 3rd WebNLG Workshop on Natural Language Generation from the Semantic Web (WebNLG+ 2020)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Debanjan Chaudhuri, Md Rashad Al Hasan Rony, Simon Jordan, and Jens Lehmann. 2019. Using a kg-copy network for non-goal oriented dialogues. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18*, pages 93–109. Springer.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021a. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3002–3017.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022a. Knowprompt:

Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, pages 2778–2788.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021b. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Zhi Chen, Jijia Bao, Lu Chen, Yuncong Liu, Da Ma, Bei Chen, Mengyue Wu, Su Zhu, Jian-Guang Lou, and Kai Yu. 2022b. Dialogzoo: Large-scale dialog-oriented task learning. *arXiv preprint arXiv:2205.12662*.

Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul A. Crook, and William Yang Wang. 2022c. KETOD: knowledge-enriched task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2581–2593. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Samuel Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. 2020. Spanconvert: Few-shot span extraction for dialog with pretrained conversational representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 107–121.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. Association for Computational Linguistics.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416.

Bo Dang, Danqing Ma, Shaojie Li, Xinqi Dong, Hengyi Zang, and Rui Ding. 2024. Enhancing kitchen independence: Deep learning-based object detection for visually impaired assistance. *Academic Journal of Science and Technology*, 9(2):180–184.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech & Language*, 59:123–156.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49.

Joachim Fainberg, Ben Krause, Mihai Dobre, Marco Damonte, Emmanuel Kahembwe, Daniel Duma, Bonnie Webber, and Federico Fancellu. 2018. Talking to myself: self-dialogues as data for conversational agents. *arXiv preprint arXiv:1809.06641*.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. 2019. Amazonqa: A review-based question answering task. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4996–5002. International Joint Conferences on Artificial Intelligence Organization.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *EMNLP*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Wanwei He, Yinpei Dai, Binyuan Hui, Min Yang, Zheng Cao, Jianbo Dong, Fei Huang, Luo Si, and Yongbin Li. 2022a. SPACE-2: Tree-structured semi-supervised contrastive pre-training for task-oriented dialog understanding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 553–569, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Wanwei He, Yinpei Dai, Min Yang, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022b. Unified dialog model pre-training for task-oriented dialog understanding and generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 187–200.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2021. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *arXiv preprint arXiv:2111.14592*.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. Trippy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

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.

Helan Hu, Shuzheng Si, Haozhe Zhao, Shuang Zeng, Kaikai An, Zefan Cai, and Baobao Chang. 2023. Distantly-supervised named entity recognition with uncertainty-aware teacher learning and student-student collaborative learning.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Qi Jia, Hongru Huang, and Kenny Q Zhu. 2021. Ddrel: A new dataset for interpersonal relation classification in dyadic dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13125–13133.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. 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 6280–6285, Hong Kong, China. Association for Computational Linguistics.

Endri Kacupaj, Joan Plepi, Kuldeep Singh, Harsh Thakkar, Jens Lehmann, and Maria Maleshkova. 2021. Conversational question answering over knowledge graphs with transformer and graph attention networks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 850–862.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tür. 2021. "how robust ru?": Evaluating task-oriented dialogue systems on spoken conversations. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1147–1154. IEEE.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. 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 1311–1316.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. Dialogue state tracking with a language model using schema-driven prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4937–4949, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yohan Lee. 2021. Improving end-to-end task-oriented dialog system with a simple auxiliary task. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1296–1303, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652.

Xiujun Li, Yu Wang, Siqi Sun, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv preprint arXiv:1807.11125*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Junhao Liu, Tongliang Liu, Fei Huang, and Yongbin Li. 2024. One shot learning as instruction data prospector for large language models.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020a. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ying Lin, Ling-Yan Bao, Ze-Minghui Li, Shu-Zheng Si, and Chao-Hsien Chu. 2020b. Differential privacy protection over deep learning: An investigation of its impacted factors. *Comput. Secur.*, 99:102061.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Hong Liu, Yucheng Cai, Zhijian Ou, Yi Huang, and Junlan Feng. 2022. Revisiting markovian generative architectures for efficient task-oriented dialog systems. In *Proceedings of the IEEE Spoken Language Technology Workshop 2022*.

Jingjing Liu, Panupong Pasupat, Scott Cyphers, and Jim Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8386–8390. IEEE.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021a. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pages 165–183. Springer.

Yuliang Liu, Xiangru Tang, Zefan Cai, Junjie Lu, Yichi Zhang, Yanjun Shao, Zexuan Deng, Helan Hu, Zengxian Yang, Kaikai An, Ruijun Huang, Shuzheng Si, Sheng Chen, Haozhe Zhao, Zhengliang Li, Liang Chen, Yiming Zong, Yan Wang, Tianyu Liu, Zhiwei Jiang, Baobao Chang, Yujia Qin, Wangchunshu Zhou, Yilun Zhao, Arman Cohan, and Mark Gerstein. 2023. Ml-bench: Large language models leverage open-source libraries for machine learning tasks.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021b. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13452–13460.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Kaixin Ma, Tomasz Jurczyk, and Jinho D Choi. 2018. Challenging reading comprehension on daily conversation: Passage completion on multiparty dialog. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2039–2048.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

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

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong,

Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.

Tomáš Nekvinda and Ondřej Dušek. 2021. Shades of bleu, flavours of success: The case of multiwoz. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 34–46.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics (ACL)*, pages 311–318. Association for Computational Linguistics.

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

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of The 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic.

Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (multidogo): Strategies toward curating and annotating large scale dialogue data. 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 4526–4536.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332.

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

Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 353–360.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *EACL*.

Tomáš Ko ciský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor

Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, TBD:TBD.

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1546–1556.

Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. 2018a. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Amrita Saha, Vardaan Pahuja, Mitesh Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018b. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations*.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of NAACL-HLT*, pages 3795–3805.

Pararth Shah, Dilek Hakkani-Tur, Bing Liu, and Gökhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51.

Igor Shalyminov, Alessandro Sordoni, Adam Atkinson, and Hannes Schulz. 2020. Fast domain adaptation for goal-oriented dialogue using a hybrid generative-retrieval transformer. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8039–8043. IEEE.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Shuzheng Si, Zefan Cai, Shuang Zeng, Guoqiang Feng, Jiaxing Lin, and Baobao Chang. 2023a. SANTA: Separate strategies for inaccurate and incomplete annotation noise in distantly-supervised named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3883–3896, Toronto, Canada. Association for Computational Linguistics.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023b. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2024. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. *Advances in Neural Information Processing Systems*, 36.

Shuzheng Si, Shuang Zeng, and Baobao Chang. 2022a. Mining clues from incomplete utterance: A query-enhanced network for incomplete utterance rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4839–4847, Seattle, United States. Association for Computational Linguistics.

Shuzheng Si, Shuang Zeng, Jiaxing Lin, and Baobao Chang. 2022b. SCL-RAI: Span-based

contrastive learning with retrieval augmented inference for unlabeled entity problem in NER. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2313–2318, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xingchen Song, Di Wu, Binbin Zhang, Zhendong Peng, Bo Dang, Fuping Pan, and Zhiyong Wu. 2023. ZeroPrompt: Streaming Acoustic Encoders are Zero-Shot Masked LMs. In *Proc. INTERSPEECH 2023*, pages 1648–1652.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676, Dublin, Ireland. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Derek Tam, Rakesh R Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4980–4991.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.

Yizhong Wang, Pegah Alipoormolabashi Swaroop Mishra, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Laiand Ishan Purohit, Ishani Mondaland Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. Supernaturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the Association for Computational Linguistics: EMNLP 2022*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Wei Wei, Quoc Le, Andrew Dai, and Jia Li. 2018. Airdialogue: An environment for goal-oriented dialogue research. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3844–3854.

Myers Will, Etchart Tyler, and Fulda Nancy. 2020. Conversational scaffolding: an analogy-based approach to response prioritization in open-domain dialogs. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART)*.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Chien-Sheng Wu, Steven CH Hoi, Richard Socher, and Caiming Xiong. 2020. Tod-bert: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. In *Proceedings of the*

59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4755–4766, Online. Association for Computational Linguistics.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. UBAR: Towards fully end-to-end task-oriented dialog system with GPT-2. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Zhengzhe Yang and Jinho D Choi. 2019. Friendsqa: Open-domain question answering on tv show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197.

Tom Young, Frank Xing, Vlad Pandelea, Jinjie Ni, and Erik Cambria. 2022. Fusing task-oriented and open-domain dialogues in conversational agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11622–11629.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940.

Tao Yu, Rui Zhang, Heyang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, et al. 2019a. Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases. 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 1962–1979.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921.

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, et al. 2019b. Sparc: Cross-domain semantic parsing in context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4511–4523.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020*, pages 109–117.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Yichi Zhang, Zhijian Ou, Min Hu, and Junlan Feng. 2020a. A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9207–9219, Online. Association for Computational Linguistics.

Yichi Zhang, Zhijian Ou, Min Hu, and Junlan Feng. 2020b. A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Zhuosheng Zhang, Shuohang Wang, Yichong Xu, Yuwei Fang, Wenhao Yu, Yang Liu, Hai Zhao, Chenguang Zhu, and Michael Zeng. 2018b. Task compass: Scaling multi-task pre-training with task prefix. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. Mmicl: Empowering vision-language model with multi-modal in-context learning.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021. Commonsense-focused dialogues for response generation: An empirical study. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 121–132.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

A. Theoretical Deduction

In this section, we give detailed theoretical deduction the superiority of our proposed TAP, which is helpful in both generating high-quality prompts and the model's transfer ability on unseen tasks.

A.1. Problem Setting

Given a task with input-output an instance pair (X, Y), we assume that there is some prompt P that is helpful to infer the task. Note that P may be in the form of instructions, keywords, or even just the task name itself. Also note that in real-life the input X and the prompt P may not have a strict boarder, and we separate them for the convenience of discussion, assuming that P contains the information relevant to the task, while X contains other information:

$$p(t|X, P) \approx p(t|P) \tag{3}$$
$$p(Y|X, P, t) \approx p(Y|X, t) \tag{4}$$

We use a language model to generate output Y conditioned on the input X and the prompt P, where the task t is viewed as a latent variable. The generation probability, under our latent task assumption, along with the Bayes' rule, can be written as follows:

$$
\begin{aligned}
&p(Y|X, P)\\
&= \int_t p(Y|X, P, t)p(t|X, P)dt\\
&\approx \int_t p(Y|X, P, t)p(t|P)dt \quad (Eq.(3))\\
&\propto \int_t p(Y|X, P, t)p(P|t)p(t)dt \ (Bayes'\ rule)\\
&\approx \int_t p(Y|X, t)p(P|t)p(t)dt \quad (Eq.(4))
\end{aligned}
\tag{5}
$$

A.2. Benefits of Using Multi-Prompt Training

We can further discuss the benefits of multi-prompt training, especially in the model's transfer ability on unseen test prompts. Given test prompts P^{test} and N training prompts $P^i(i = 1, 2, \cdots, N)$, by increasing the number of training prompts N, the minimal distance between the embedding of the test prompts and training prompts in the latent task space: $\min_i |emb(P^{test}) - emb(P^i)|$ will reduce and $p(Y|X, P^{test})$ (the probability of generating the correct label Y) will increase according to Eq. 5. Moreover, if the number of training prompts N is large enough, the expectation of the minimal distance will reduce to any given value above 0:

$$\lim_{N \to \infty} \mathbb{E} \min_i |emb(P^{test}) - emb(P^i)| \to 0 \tag{6}$$

Under this circumstance, the probability $p(Y|X, P^{test})$ will be optimized during training, resulting in strong performance. The details of the deduction is shown in Section A.3.

A.3. Deduction of the Consistency of Multi-prompt Training

Given the test prompt P^{test}, the distance between the embedding of the test prompts and the ith training prompts in the latent task space can be written as:

$$d^i = |emb(P^{test}) - emb(P^i)|(i = 1, 2, \cdots, N)$$

Assume that $d^i(i = 1, 2, \cdots, N)$ are independent identically distributed (i.i.d) and the supporting set of $p(d^i)$ is $d > 0$, which means the probability $P(d) = p(d^i < d) > 0$ for any distance $d > 0$. We can prove that given any $d > 0$ and any $\epsilon > 0$, there exists an N that ensures $p(\min_i(d^i) < d) > 1 - \epsilon$:

$$\forall d, \epsilon > 0 \ \exists N \ s.t. \quad p(\min_i(d^i) < d) > 1 - \epsilon \tag{7}$$

Theorem. 7 can be easily proved by calculating the probability of $p(\min_i(d^i) > d)$:

$$p(\min_i(d^i) > d) = \prod_{i=1}^{N}(1 - p(d^i < d))$$
$$= (1 - P(d))^N \quad (i.i.d) \quad (8)$$

As $P(d) > 0$, we can take $N > log_{1-P(d)}\epsilon$ according to Eq.8. Therefore the Theorem 7 is proved.

From Theorem. 7, we can deduce that if N is large enough, we can find a training prompt that satisfies $|emb(P^{test}) - emb(P^i)| < d$, therefore the distribution of $p(P^{test}|t)$ is close to $p(P^i|t)$ in the latent task space:

$$KL(p(P^{test}|t)||p(P^i|t)) < d_1 \quad (9)$$
$$KL(p(Y|X,t)p(P^{test}|t)p(t)|| $$
$$p(Y|X,t)p(P^i|t))p(t)) < d_2 \quad (10)$$
$$|\int_t p(Y|X,t)p(P^{test}|t)p(t)dt - $$
$$\int_t p(Y|X,t)p(P^i|t)p(t)dt| < d_3 \quad (11)$$

d_1, d_2, d_3 are constant that converge to zero with the increase of the number of training prompts N. Therefore the training probability of ith sample converges to test probability, proving the consistency of multi-prompt training:

$$|p(Y|X,P^i) - p(Y|X,P^{test})| \to 0 \quad (12)$$

B. Exploit Prompts for Low Resource Setting

Prompts can reduce the gap between language model pre-training and finetuning, therefore improving model's performance in downstream tasks, especially in few-shot and zero-shot settings (Gao et al., 2021; Cui et al., 2021; Chen et al., 2022a). Apart from that, pattern exploit training (PET), a self-training method leveraging multiple prompts, can greatly improve model's performance in low resource setting by perform semi-supervision training Different prompts can be used as different view for the case, and models finetuned with different prompt are used to ensemble pseudo labels on un-labeled data (Schick and Schütze, 2021a). There are a few works that improve over the original pattern exploit training: Schick and Schütze (2021b) extends PET to deal with labels that have multiple tokens, while Tam et al. (2021) proposes to provide more supervision and learn without task-specific unlabeled data. Our PET contributes in reformulating PET to apply it to generative language model. Moreover, we combine PET with our multi-task prompt pre-trained model and applied multi-prompt training in the finetuning stage of PET, improving the accuracy of the generated pseudo labels.

C. Prompts for Semi-Supervised Training: PET

To utilize numerous and diverse generated prompts, as well as the pre-trained model that performs well on those prompts, we perform PET (Schick and Schütze, 2021a) for semi-supervised training. We adapt the origin PET method to better utilize multiple prompts, as well as fitting our pretrained model.

For the generated prompts $P = P^j(j = 1, 2, \cdots, M)$, we use a partition of P, $P_1, P_2, \cdots, P_k$ to train k voting models for ensembling. The lth voting model M_l are finetuned from the pre-trained model on the annotated part of data $(X^i, Y^i)(i = 1, 2, \cdots, N_a)$ with the prompt sets P_l, the loss function as follows:

$$\mathcal{J}_\theta^l = \sum_{i=1}^{N_a}\sum_{j=1}^{|P_l|}\log p_{M_l}(Y^i|X^i, P_l^j) \quad (13)$$

To generate pseudo labels on unannotated data, we ensemble the outputs generated by voting models given all input instances and prompts:

$$\tilde{Y^i} = ensemble(\{\tilde{Y_j^i}\}), \tilde{Y_j^i} \sim p_{M_l}(\tilde{Y^i}|X^i, P_l^j) \quad (14)$$

where we use majority voting method to perform ensembling for the labels generated. Sampling is used in (14) to increase diversity of the generated label, helping us to distinguish those instances and labels that are deemed uncertain by the model. Because we finetune the voting models on a model pre-trained over all prompts and we use multi-prompt training to finetune the voting models in (15), the accuracy of the voting models is greatly improved, therefore advancing the quality of the pseudo labels generated.

The N_p pseudo labels are used to train the model, along with the annotated data, to improve the model's performance:

$$\mathcal{J}_\theta = \sum_{i=1}^{N_a}\sum_{j=1}^{M}\log p(Y^i|X^i, P^j)$$
$$+ \sum_{k=1}^{N_p}\sum_{j=1}^{M}\log p(\tilde{Y^k}|X^k, P^j) \quad (15)$$

Task	Datasets
Natural language generation	web_nlg (Castro Ferreira et al., 2020), dart (Nan et al., 2021b), e2e_nlg (Dušek et al., 2020), common_gen (Lin et al., 2020a)
Summary	dialogsum (Chen et al., 2021b), xlsum (Hasan et al., 2021), xwikis (Perez-Beltrachini and Lapata, 2021), wiki_lingua (Ladhak et al., 2020), Samsum (Gliwa et al., 2019)
Slot filling	Restaurant8k (Coope et al., 2020), TOP (Gupta et al., 2018), DSTC8 (Rastogi et al., 2020), ATIS (Hemphill et al., 1990), CrossNer (Liu et al., 2021b), FB_TOD_SF (Schuster et al., 2019), MIT-movies-eng (Liu et al., 2013), MIT-movies-eng (Liu et al., 2013), MIT-movies-trival10k (Liu et al., 2013), MIT-restaurant (Liu et al., 2013), SNIPS (Coucke et al., 2018),
Intent prediction	BANKING77 (Casanueva et al., 2020), CLINC150 (Larson et al., 2019), HWU64 (Liu et al., 2021a), FB_TOD_SF (Schuster et al., 2019), SNIPS (Coucke et al., 2018), TOP (Gupta et al., 2018), MultiWOZ2.2 (Zang et al., 2020), SGD(Rastogi et al., 2020), WOZ (Mrkšić et al., 2017), SimJoint (Shah et al., 2018), MultiWOZ_synthesis (Campagna et al., 2020), SwDA (Stolcke et al., 2000), DailyDialog (Li et al., 2017), DSTC2 (Williams et al., 2016), DSTC3 (Williams et al., 2016), InCar (Eric et al., 2017) ,PersuaGOOD (Wang et al., 2019), Frames (El Asri et al., 2017), MulDoGo (Peskov et al., 2019), BiTOD (Lin et al., 2021), MSRe2e (Li et al., 2018),
Dialog state tracking	SGD (Rastogi et al., 2020), TaskMaster1 (Byrne et al., 2019), TaskMaster2 (Byrne et al., 2019), TaskMaster3 (Byrne et al., 2019), WOZ (Mrkšić et al., 2017), KETOD (Chen et al., 2022c), MulDoGo (Peskov et al., 2019), InCar (Eric et al., 2017), SimJoint (Shah et al., 2018),MultiWOZ2.2 (Zang et al., 2020),
Multiple choice	Commensense-qa (Talmor et al., 2019), Cosmosqa (Huang et al., 2019), Meld (Poria et al., 2019)
Emotion classification	DailyDialog (Li et al., 2017), Go-emotion (Demszky et al., 2020), Meld (Poria et al., 2019), SentiHood (Saeidi et al., 2016), MAMS (Jiang et al., 2019), ASTE (Xu et al., 2021), RECCON (Poria et al., 2021)
Document-based question answering	SQuAD (Rajpurkar et al., 2016), QuAC (Choi et al., 2018), NarrativeQA (s Ko □ciský et al., 2018), Race (Lai et al., 2017)
Dialog-related question answering	DREAM (Sun et al., 2019), Molweni (Li et al., 2020), DialogRE (Yu et al., 2020), FriendsQA (Yang and Choi, 2019), DDRel (Jia et al., 2021), ReadingComprehension (Ma et al., 2018), RECCON (Poria et al., 2021), WizInt (Komeili et al., 2022)
Chit-chat & Response generation	Mutual (Cui et al., 2020), ABCD (Chen et al., 2021a), AirDialog (Wei et al., 2018), CCPE (Radlinski et al., 2019), MetalWOZ (Shalyminov et al., 2020), CMU_DoG (Zhou et al., 2018), CoQA (Reddy et al., 2019), CoSQL (Yu et al., 2019a), doc2dial (Feng et al., 2020), DSTC2-track2 (Kim et al., 2021), DSTC10-track3 (Kottur et al., 2021), MedicalDialog (Zeng et al., 2020), Self-Dialog (Fainberg et al., 2018), WOW (Dinan et al., 2018), TopicChat (Gopalakrishnan et al., 2019), Persona-Chat (Zhang et al., 2018a), MulDoGo_un (Peskov et al., 2019), CSQA (Saha et al., 2018b), AmazonQA (Gupta et al., 2019), ChitChat (Will et al., 2020), EmpatheticDialog (Rashkin et al., 2019), CommonsenseDialog (Zhou et al., 2021), ConvQuestions (Kacupaj et al., 2021), MMD (Saha et al., 2018a)
Knowledge-grounded dialog	Soccer-kgdial (Chaudhuri et al., 2019), Incar-kgdial (Chaudhuri et al., 2019), WizInt (Komeili et al., 2022), KETOD (Chen et al., 2022c)
Text to SQL	Spider (Yu et al., 2018), Sparc (Yu et al., 2019b)
Task oriented dialog & User simulation	TaskMaster1 (Byrne et al., 2019), TaskMaster2 (Byrne et al., 2019), TaskMaster3 (Byrne et al., 2019), SwDA (Stolcke et al., 2000), FusedChat(Young et al., 2022), Frames (El Asri et al., 2017), MultiWOZ2.2 (Zang et al., 2020), SGD(Rastogi et al., 2020), WOZ (Mrkšić et al., 2017), SimJoint (Shah et al., 2018), MultiWOZ_synthesis (Campagna et al., 2020), MulDoGo (Peskov et al., 2019), DailyDialog (Li et al., 2017), DSTC2 (Williams et al., 2016), DSTC3 (Williams et al., 2016), InCar (Eric et al., 2017) ,PersuaGOOD (Wang et al., 2019), BiTOD (Lin et al., 2021) MSRe2e (Li et al., 2018),

Table 9: Pre-training tasks and datasets in UniPreDial.

D. Building pre-training corpus

To perform multi-task pre-training for conversation model, we collect UniPreDial [4], which contains 122 dialog-related from 15 dialog-related tasks. The tasks in UniPreDial mainly fall into three categories: task-oriented dialog related (intent prediction, dialog state tracking and grounded dialog), open-domain chit-chat, and other dialog-related datasets.

Task-oriented dialog is extensively studied by previous researchers, resulting in abundant annotated datasets. We make full use of the annotated information as we leverage prompts to convert a turn in a dialog into multiple training instances, as shown in Figure 1.

Open-domain chit-chat datasets are important for improving the generation ability of pre-trained conversation models. We use the datasets collected in He et al. (2022b) [5] as the datasets are competitive in quality and quantity. However, instead of viewing those datasets as unannotated data for semi-supervised training for task-oriented dialog, we train the response generation task on those datasets, leveraging the coherency of open-domain chat datasets.

To extend the model's ability, we collect other datasets that can improve the model's skills. Emotion classification, summary, natural language generation, and text2sql are important skills for dialog systems in real-life scenarios, while question answering and multiple choice have similar format as dialog and will yield positive transfer in co-training (Aribandi et al., 2021).

---

[4]We collect datasets from https://hugging-face.co/datasets, https://www.parl.ai/docs/tasks.html and GitHub repositories on https://github.com/.

[5]https://github.com/AlibabaResearch/DAMO-ConvAI/tree/main/space-3