CGoDial: A Large-Scale Benchmark for Chinese Goal-oriented Dialog Evaluation

Yinpei Dai, Wanwei He, Bowen Li, Yuchuan Wu, Zheng Cao,
Zhongqi An, Jian Sun, Yongbin Li*
Alibaba Group, Beijing, China
{yinpei.dyp,hewanwei.hww,shengxiu.wyc,zhengzhi.cz,
zhongqi.azq,jian.sun,shuide.lyb}@alibaba-inc.com

Abstract

Practical dialog systems need to deal with various knowledge sources, noisy user expressions, and the shortage of annotated data. To better solve the above problems, we propose **CGoDial**[1], a new challenging and comprehensive **C**hinese benchmark for multi-domain **Go**al-oriented **Dial**og evaluation. It contains 96,763 dialog sessions, and 574,949 dialog turns totally, covering three datasets with different knowledge sources: 1) a slot-based dialog (SBD) dataset with table-formed knowledge, 2) a flow-based dialog (FBD) dataset with tree-formed knowledge, and a retrieval-based dialog (RBD) dataset with candidate-formed knowledge. To bridge the gap between academic benchmarks and spoken dialog scenarios, we either collect data from real conversations or add spoken features to existing datasets via crowdsourcing. The proposed experimental settings include the combinations of training with either the entire training set or a few-shot training set, and testing with either the standard test set or a hard test subset, which can assess model capabilities in terms of general prediction, fast adaptability and reliable robustness.

1 Introduction

Goal-oriented dialog systems converse with users naturally, helping them fulfill specific goals such as restaurant booking, hotel reservation, and flight search. Many popular datasets have been introduced to facilitate the dialog research, ranging from simple single-domain (Wen et al., 2017) to more difficult multi-domain dialogs (Budzianowski et al., 2018). Similar to the widely studied GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a) for general natural language understanding, large-scale benchmarks have been specifically designed for dialog systems. For instance, Mehri et al. (2020) proposed the DialoGLUE, a collection of seven popular English dialog datasets, to assess the natural language understanding ability in a few-shot setting. More recently, Peng et al. (2021b) introduced the RADDLE, another large-scale benchmark that includes a diagnostic checklist facilitating detailed robustness analysis for pre-trained language models (Radford et al.; Peng et al., 2021a).

However, compared to the abundance of English dialog benchmarks, the resource of Chinese dialog datasets in goal-oriented dialog remains quite limited, let alone the large-scale benchmark suitable for the dialog evaluation of Chinese pre-trained language models. Moreover, current Chinese dialog datasets, such as CrossWOZ (Zhu et al., 2020) and RiSAWOZ (Quan et al., 2020), mainly focus on the schema-guided dialog as in the classical MultiWOZ dataset (Budzianowski et al., 2018), where conversations are grounded on a table-formed ontology. This kind of ontology usually defines a set of possible slot-value pairs to be recognized during the dialog for database query (Young et al., 2013). In real-world applications, however, more types of goal-oriented dialogs are involved. For instance, in Xie et al. (2022) the conversational process is grounded on a pre-defined tree-formed ontology, where the whole process is constrained in a dialog flow. Also, in some scenarios, the conversation has been simplified into a multi-turn response selection problem so that the dialog system is able to transfer to new domains quickly and reliably (Henderson et al., 2019a, 2020; Dai et al., 2020).

These observations motivate us to construct CGoDial, a new large-scale Chinese benchmark for goal-oriented dialog, targeting the evaluation of model adaptability and robustness in real situations. Concretely, we consider three types of goal-oriented dialogs with different knowledge database sources. The first is the slot-based dialog (SBD), where a system often asks a user for required attributes to constrain the search in a table-formed

*Corresponding author
[1]Dataset available at https://github.com/AlibabaResearch/DAMO-ConvAI/cgodial.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4097 - 4111
December 7-11, 2022

database and then provides entities and extra information to the user. To build SBD, we adapt the existing Chinese dataset RiSAWOZ (Quan et al., 2020) by supplementing it using crowd-sourcing for more difficult variations such as noisy expression augmentation, external knowledge utilization, and out-of-scope management. 44,800 new dialogs spanning 12 domains are collected for SBD. The second is the flow-based dialog (FBD), where the system guides the user to fulfill specific tasks based on tree-structured dialog flow. This flow is a new form of knowledge database that defines causal constraints of conversations. For example, when a user comes to withdraw money in a housing insurance consulting scenario, the system must first determine the user's identity before moving to the next step. Therefore, compared with SBD, FBD has a strict order to request user information We collect new dialogs between human customers and an online customer service dialog agent from real businesses, which to the best of our knowledge, is the first Chinese FBD dataset. After data desensitization, we have 6,786 dialogs and 45,601 turns spanning four different domains in the end. The third is the retrieval-based dialog (RBD), like bAbI-dialog (Bordes and Weston, 2017), where the dialog system learns to select the correct response from a candidate response set. We build RBD by adapting the existing ECD (Zhang et al., 2018) dataset and adding more noisy user expressions like ASR errors to raise the overall difficulty. We choose 56,000 non-ambiguous complex dialog examples from the original dataset.

To fully assess dialog model capabilities in general prediction, fast adaptability, and reliable robustness. For all the above types of dialog datasets, we propose four different experimental settings, including the combinations of training with either the full training set or a few-shot training set, and testing with either the standard test set or a hard test subset. Extensive experiments have been conducted on CGoDial with various Chinese pre-trained language model baselines for dialog modeling, such as Chinese-T5 (Zhao et al., 2019) and CDial-GPT (Wang et al., 2020). To further facilitate the research of dialog pre-training, we also release a new UniLM-based dialog model pre-trained on large-scale human-to-human dialog corpora, which achieves the best results on all tasks.

Therefore, the contributions of this paper are three-fold:

- A challenging and comprehensive Chinese dialogue benchmark consisting of three different types of goal-oriented dialogs.
- Standardized evaluation measures that facilitate the study of robustness and adaptability.
- Competitive baselines across different dialog tasks. Both the datasets and codes will be open-sourced to push forward the research in goal-oriented dialog.

2 Related Work

2.1 Goal-oriented Dialog Benchmarks.

The introduction of benchmarks brings forward the research of goal-oriented dialog. Several recent tendencies have penetrated the development:

From single-domain to multi-domain. The earliest dataset such as bAbI-dialog (Bordes and Weston, 2017) and WOZ (Wen et al., 2017) focus on the single-domain dialog. In the following, extended-bAbI (Dai et al., 2020), MultiWOZ (Budzianowski et al., 2018) and Schema-Guided Dialog (Mosig et al., 2020) are proposed to solve more difficult tasks in multi-domain dialog.

From slot-based to flow-based. The classical MultiWOZ dataset is grounded on the table-formed ontology, so the dialog understanding problem can be put forward as a multi-turn slot-filling task. However, in many real applications, there are fixed order constraints to collect the slot information from users, STAR (Mosig et al., 2020) and ABCD (Chen et al., 2021) are proposed recently to cover this property by exerting causal constraints on the dialog flows.

From simplified tasks to real scenarios. Most previous dialog benchmarks (Rojas-Barahona et al., 2018; Dai et al., 2018; Rastogi et al., 2020; Quan et al., 2020; Zhu et al., 2020; Zhang et al., 2022b) focus on text-in text-out dialogs but neglect spoken characteristics in real problems. RADDLE (Peng et al., 2021b) is the first to consider evaluating model robustness by adding various ASR noises to the original MultiWOZ but is not publicly available now. Shafran et al. (2022) extends the same idea to propose a speech-aware dialog systems technology challenge. NSD (Wu et al., 2021) aims to discover unknown or out-of-domain slot types for dialogue systems. EmoWOZ (Feng et al., 2021) recognizes the critical role of emotion and provides a new

emotion-annotated corpus of goal-oriented dialogs based on MultiWOZ. SSTOD (Zhang et al., 2022a) proposes a novel sub-slot filling task that is crucial in string-formed information collection like phone numbers and people names.

From single-modal to multi-modal. One type of multi-modal dialog is to build a system that can help users search for target objects via generating textual responses and object pictures. Common datasets are MMD (Saha et al., 2018), MMConv (Liao et al., 2021) and JDDC 2.0 (Zhao et al., 2021), another type is to deal with immersive and situated multi-modal scenarios, such as SIMMC (Crook et al., 2019), SIMMC2.0 (Kottur et al., 2021) and TEACH (Padmakumar et al., 2022), where the visual input of the agent's surrounding environment needs to be used for conversations; and 5) From monolingual to multi-lingual, such as BiTOD (Lin et al., 2021), GlobalWOZ (Ding et al., 2022), Multi^{2}WOZ (Hung et al., 2022) and AllWOZ (Zuo et al., 2021).

Compared with previous work, our CGoDial focuses on single-modal and is the first Chinese benchmark that covers flow-based dialog and considers real scenarios.

2.2 Dialog Data Collection

Goal-oriented dialog construction can be broadly divided into three categories according to the data collection scheme.

Machine-to-machine (M2M) scheme creates data via dialog simulation (Bordes and Weston, 2017; Rastogi et al., 2020; Kottur et al., 2021), given manually designed utterance templates and dialogue process. Crowd-sourcing is then used to paraphrase the dialog with more varied language expressions. This approach can collect very large-scale data at low cost (Shah et al., 2018) but often lacks noisy conditions and flexible processes that appear in human conversations (Black et al., 2011).

Human-to-machine (H2M) scheme collects dialogs based on an existing dialog system. Standard datasets include the Let's Go Bus Information System (Raux et al., 2005) and the second and third Dialog State Tracking Challenges (Henderson et al., 2014a,b). This method is closest to the real applications but requires a well-deployed dialog agent prepared ahead.

Human-to-human (H2H) scheme is also a cheap way to collect dialog data as there are many available dialog corpora on the internet like Twitter (Ritter et al., 2010) and Reddit (Henderson et al., 2019b). However, building annotated dialogs that meet the requirements of specific tasks relies on the costly Wizard-of-Oz framework. CrossWOZ (Zhu et al., 2020) and MultiWOZ (Budzianowski et al., 2018) lie in this line.

In our CGoDial benchmark, slot-based dialog and retrieval-based dialog are collected by H2H, and flow-based dialog is collected by H2M.

3 CGoDial Benchmark

In this section, we elaborate on the three datasets in our CGoDial benchmark, in terms of task definition and dataset construction. To give an overall impression, Figure 1 illustrate typical examples for each dataset and Table 1 summarizes detailed statistics. Appendix A.1 illustrates more dataset distributions of dialog length and turn length.

3.1 Slot-based Dialog Dataset

The slot-based dialog (SBD) is tasked to search and provide entities from a table-formed database that meet the requirement of a user through natural interactions, such as reserving a restaurant for the user. The database specifies an ontology about the semantic scope of the system can process, which is represented as a set of slot-value pairs to maintain a dialog state at each turn. For example, in the classical MultiWOZ dataset (Budzianowski et al., 2018), if a user says *"I would like a cheap hotel in the west."*, the system needs to extract the dialog state as *'pricerange=cheap, area=west'*. Then the system uses the dialog state to query the database and decides dialog actions based on the searching results. For example, the system can decide to request more unfilled slots like *"what type of food do you like"* to constrain the search if there are too many returned entities. After the entity is fixed, the user may ask for more information about the entity like *"what is the phone number and address of the nados restaurant?"*, and the system needs to provide the information of the requested slots *'phone_number, address'*.

3.1.1 Task Definition

To fully evaluate the model capability, we include tasks for dialog understanding, policy, and generation. For the dialog understanding, we choose the dialog state tracking (DST) task (Henderson et al., 2014a), which tracks the slot-values conveyed by the user during the dialog. We use `JGA` to denote

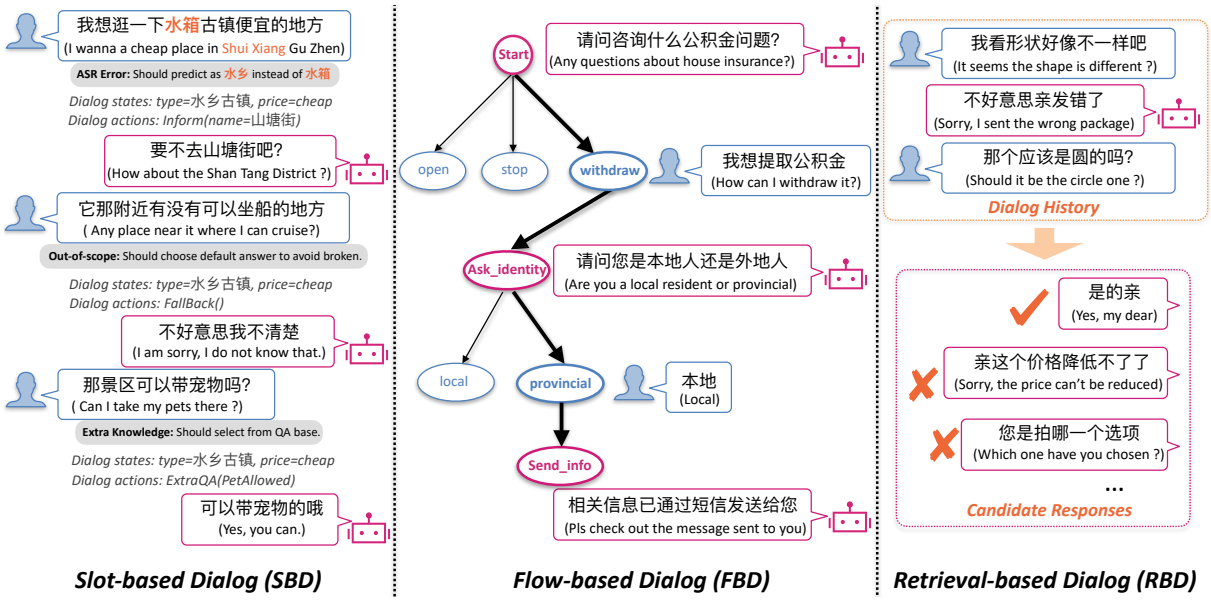

Figure 1: Three dialog examples of SBD, FBD and RBD in CGoDial, respectively.

Dataset		Train	Valid	Test-std	Test-hard
Slot-Based Dialog (SBD)	#dialog	40,000	2,400	2,400	600
	#turn	403,740	32,464	37,144	9,286
Flow-Based Dialog (FBD)	#dialog	3,950	1,450	1,386	985
	#turn	15,558	5,167	5,624	3,629
Retrieval-Based Dialog (RBD)	#dialog	39,589	4,627	961	701
	#turn	50,000	5,000	1,000	719

Table 1: Statistics of CGoDial. We maintain two test sets for evaluation. One is the standard test set (denoted as test-std), and another is a robust test set that is a hard subset chosen from test-std (denoted as test-hard).

the joint goal accuracy as in Heck et al. (2020) for DST, which is the proportion of turns that all slot-value pairs have been correctly predicted. For the dialog policy, since there is the one-to-many property (Zhang et al., 2020a) in policy planning and we do not have user simulators (Ultes et al., 2017) to calculate the online task success rate, we adopt a similar metric in Budzianowski et al. (2018) and use `Succ` to denote turn accuracy of dialogs where all user requested slots[2] have been correctly answered and the searched entity follows the oracle dialog state. For the dialog generation, we follow the common practice to compute the average BLEU scores (Papineni et al., 2002) for all turns. Specifically, we choose BLEU-4 in SBD. Inspire by Mehri et al. (2019), we also calculate a combined score `Comb` to measure the overall performance for all tasks, which is the geometric mean of above metrics: $\texttt{Comb} = \sqrt[3]{\texttt{JGA} \times \texttt{Succ} \times \texttt{BLEU}}$. Different from the original combined score that adds all scores linearly, the geometric mean should be more reasonable in the aspect of dimensional calculation.

3.1.2 Dataset Construction

We build our SBD dataset based on the existing Chinese dataset RiSAWOZ (Quan et al., 2020). Since RiSAWOZ is purely text-in-text-out and has limited language variation, it can not reflect the difficulty in realistic spoken dialog system applications. We add three important common features in real scenarios to fill this gap in the current Chinese dialog benchmarks.

Feature 1: External knowledge. Users often ask for some new knowledge that is not covered in the current ontology. Inspired by the work of (Kim et al., 2020), we expand the coverage of dialog systems by incorporating external unstructured knowledge sources to tackle users' unseen requests. In our SBD, the external knowledge is represented as question-answer (QA) pairs that can be utilized

[2]Except regular requested slots in RiSAWOZ, external QA pairs and OOS utterances are also treated as specially requested slots to predict together. Details are given in the next section.

during the conversation. The system must decide whether the current turn should select a proper answer from the external resources or generate the answer by itself. To construct this new part of dialog data, we first ask the crowd-sourcing people to make up new relevant QA pairs given the RiSAWOZ dialogs and the ontology. After manually selecting 150 basic QA pairs, such as *"Can I take my pets?"* and *"Are there any artistic shows in the scenic spot?"*, we ask the crowd-sourcing people to paraphrase each QA pair into nine similar pairs, and acquire a total of 1,500 QA pairs. Then we insert QA pairs into the original dialog randomly as the external knowledge. Specifically, each dialog is inserted at least one QA pair with a probability of 0.5. We also hold 200 QA pairs out of the training or validation sets, and only add them into the test set for generalization evaluation. We treat the question selection problem as specially requested slots detection problem like *'IsPetAllowed'*, and count the accuracy as a part of `Succ`.

Feature 2: Out-of-scope (OOS) utterances. One of the most common problems in practical dialog systems is that users can talk about something beyond the semantic scope that the system can process (e.g., meaningless sentences like *'Uh huh... well it should have...'* and irrelevant questions *'where can I cruise'*). Thus, practical dialog systems require robust detection of OOS situations to avoid conversational breakdowns and properly handle unseen user behaviors. To simulate the feature, for each dialog in RiSAWOZ, we ask the human annotators to insert plausible OOS turns based on the given context and all external QA pairs. Each dialog is inserted at most two OOS utterances with the probability of 0.6. The corresponding default answer for OOS is *"I am sorry, I do not know that"*. Like QA prediction, we treat the OOS detection as a special requested slot *'OOS'* detection, and count the accuracy as a part of `Succ`.

Feature 3: Spoken noise. To mimic the spoken language phenomena in real applications, we add speech errors by crowd-sourcing. More concretely, we ask the people to read out all utterances four times in modified RiSAWOZ, allowing them to vary the expression subtly under the same core semantics. For example, the utterance *'I want a cheap hotel'* can be read as *'I need a cheap hotel please, I am in a hurry'*. After that, we use the off-the-shelf ASR tool[3] to transcribe all audios into noisy texts. Therefore, each dialog in RiSAWOZ is augmented into four dialogs with varied expressions and ASR noises. We carefully clean the texts to obtain our final SBD dataset. The comparison between SBD and RiSAWOZ is shown in Table 2.

Dataset	RiSAWOZ (Quan et al., 2020)	CGoDial-SBD
#Domain	12	12
#Dialog	11,600	44,800
#Turns	75,991	473,348
Extra knowledge	No	Yes
OOS detection	No	Yes
Spoken feature	No	Yes

Table 2: The comparison of between SBD and the RiSAWOZ dataset.

3.2 Flow-based Dialog Dataset

Flow-based dialog (FBD) is quite common in industrial dialog products, such as such as Microsoft BotFramework [4], Google DialogFlow [5], Salesforce Converse [6] and Alibaba Intelligent Robot [7], due to its easy-to-use and drag-and-drop characteristics for dialog developers. However, currently, there is no available Chinese dataset for FBD, which motivates us to collect a new FBD dataset to facilitate the research.

3.2.1 Task Definition

The flow-based dialog (FBD) typically contains an explicit grounded dialog flow to instruct the ongoing conversation between users and systems. The flow is a decision tree with user nodes and system nodes alternately. The user node (a.k.a., intention node) specifies the classes that what kind of utterance the user would say, the system node (a.k.a., reply node) specifies the system response. Each user node points to only one system node, but each system node can point to multiple user nodes. Usually, the flow is handcrafted by experienced dialog composers and specifies the strict order of the dialog process (Mosig et al., 2020). For example, in a banking transaction business (Chen et al., 2021), if a user wants to withdraw the deposit, a bank clerk should first determine the user's identity by confirming his residency, then it would go into detailed

[3]https://www.alibabacloud.com/help/product/30413.htm
[4]https://learn.microsoft.com/en-us/composer/
[5]https://cloud.google.com/dialogflow
[6]https://github.com/salesforce/Converse
[7]https://www.alibabacloud.com/product/bot

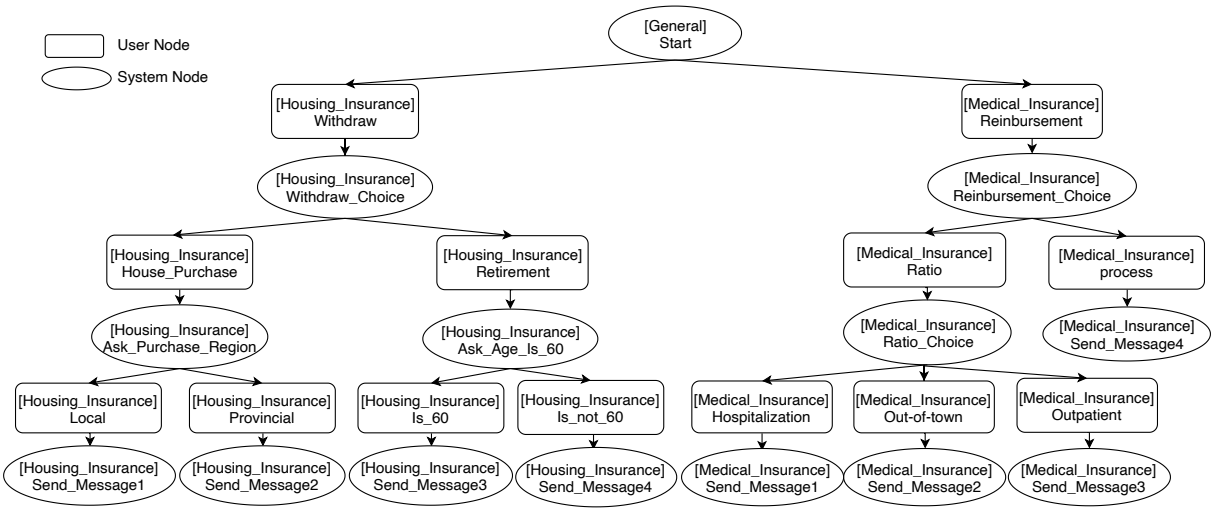

Figure 2: A part of the multi-domain FBD dialog flow from the social insurance business (There are 284 nodes in total, we only show 27 nodes here). The names on the nodes are direct translations from Chinese. The words in the square brackets [..] are the *domain* names, while the other words are the node names. Each user node can be viewed as a user intention. Each system node contains a predefined response.

business. Although the dialog process in FBD is not as flexible as in SBD, it has more advantages in specifying certain rules in the conversation.

For FBD, the main task is to predict the correct user node on the flow that the current user utterance is talking about. Therefore it can be formulated as a node classification problem given the dialog context and the flow structure. All the responses on the system nodes are pre-defined, so both dialog policy and dialog generation tasks are not crucial. Inspired by Rajendran et al. (2019), two essential metrics are used here: 1) `TurnACC` measures the percentage of user nodes that are correctly chosen, 2) `DialACC` measures the percentage of dialogs where every user node is correctly chosen. The former can be viewed as an ability of dialog understanding, while the latter tracks task completion.

3.2.2 Dataset Construction

We first manually design four different dialog flows for daily businesses, including social insurance, residency management, housing insurance, and electronic toll collection. Please refer to the Figure 2 for a detailed dialog flow from social insurance. We then deploy four different dialog agents for the businesses using the Alibaba Intelligent Robot. All the agents are then used to collect human-to-machine spoken dialogs from real conversations. After collecting the online conversational logs, we use the same ASR tool to transcribe the spoken audios into noisy texts to construct the dataset. We then ask two business experts to examine the dialogs and rectify mispredicted dialog data.

Detailed statistics of FBD are given in the Table 3. As shown in the table, we propose a new FBD dataset with 6.78k turns spanning four different businesses based on specific business trees. The system will guide the user to finish the goal according to the schema described in the tree. The dialog progresses via traversing the tree, where at each turn, the agent needs to select the correct user node from all possible successors by understanding the user's utterance, and give the response saved on the next corresponding system node.

3.3 Retrieval-based Dialog Dataset

3.3.1 Task Definition

Learning end-to-end retrieval-based dialog is a crucial direction in dialog research (Bordes and Weston, 2017; Dai et al., 2020; Tao et al., 2021), which is also very common in real applications (Williams et al., 2017; Henderson et al., 2019c). The retrieval-based dialog (RBD) in our benchmark is similar to the well-studied bAbI-dialog (Bordes and Weston, 2017). Given the dialog history, RBD aims to select the correct response from a candidate set of responses, learning goal-oriented dialog in an end-to-end manner. Since RBD is a retrieval problem, we leverage the standard IR metrics to evaluate model performance, including recall@1 (`R@1`) and mean reciprocal rank (`MRR`).

Dataset	STAR (Mosig et al., 2020)	CGodial-FBD				
		Social Insurance	Residency Management	Housing Insurance	Electronic Toll Collection	*Total*
#Domain	13	6	10	9	12	37
#Dialog	5,820	4,510	1,036	454	786	6,786
#Turn	127,833	17,510	4,488	1,316	3,035	26,349
#Node	297	284	260	146	98	788
Spoken feature	No	Yes	Yes	Yes	Yes	–

Table 3: The detailed comparison of between FBD and STAR dataset.

Model	Full-train & Test-std				Few-train & Test-std				Full-train & Test-hard			
	JGA	Succ	BLEU	Comb	JGA	Succ	BLEU	Comb	JGA	Succ	BLEU	Comb
Chinese-T5	49.4	48.3	**26.2**	75.1	2.16	5.42	8.27	12.1	27.43	26.17	**22.49**	49.3
CDial-GPT	**53.7**	49.6	24.4	76.1	**5.58**	5.49	8.10	13.6	33.15	28.00	19.28	49.9
Our PCM	51.6	**52.6**	25.6	**77.6**	5.17	**5.62**	**8.89**	**14.3**	**33.49**	**32.83**	21.63	**54.8**

Table 4: Results on slot-based dialog (SBD) dataset.

3.3.2 Dataset Construction

We adapt from the existing text-in-text-out E-commerce Dialogue Corpus (Zhang et al., 2018) to construct our RBD dataset. Since the original corpus is too large to be incorporated as a part of our dialog benchmark, we run five BERT baselines with different random seeds to sort the data according to the average predict accuracy and choose a proportion (100k turns) of difficult dialog examples as our initial RBD. Then we ask the crowd-sourcing people to vote whether the dialog turn has ambiguous response selection; if not, we ask them to read out all the utterances to add spoken features through the same procedure in SBD construction to acquire the dataset.

3.4 Data Quality Control

Crowd-sourcing brings noisy data and annotations. To guarantee the quality of collected dialog, for all datasets (i.e., SBD, FBD and RBD) in CGoDial, we ask another three crowd-sourcing people to vote whether the data is recognizable and its annotation is nonambiguous. Then we recollect the dialog samples that at least two people vote against (around 20k turns) until they meet our standard.

4 Experiments

4.1 Baselines

Pre-trained conversation models gain increasing research interest in dialog communities (Zhang et al., 2020b; Wu et al., 2020; He et al., 2022c,a,b). Our CGoDial benchmark targets evaluating the Chinese pre-trained conversation models (PCMs). However, the number of published Chinese PCMs is quite limited. For SBD, we use the Chinese-T5 (Zhao et al., 2019) and CDial-GPT (Wang et al., 2020) as base models, and use MinTL (Lin et al., 2020), a T5-based dialog model, and UBAR (Yang et al., 2021), a GPT-based dialog model, as the downstream dialog models respectively. Note that both MinTL and UBAR are specifically designed for datasets like MultiWOZ, which is similar to our SBD. For FBD and RBD, since all tasks are classification tasks, we choose two Chinese BERT-like pre-trained language models, StructBERT (Wang et al., 2019c) and Roberta-wwm (Cui et al., 2019) as baselines. StructBERT incorporates language structures into pre-training for deep language understanding. Roberta-wwm uses the whole word masking for MLM training.

4.2 Methods

In this work, we propose a new Chinese PCM by pre-training a UniLM on a sizeable open-domain Chinese dialog corpus for ten epochs. The corpus comprises nearly 100 million conversations collected from online textual dialog forums. Since the total number is vast for the textual dialogs, we apply the artificial ASR augmentation (Ma et al., 2020) rather than crowd-sourcing to add spoken errors. In this way, our PCM can achieve better robustness than other pre-trained models only trained on plain texts.

For SBD data, all models take the dialog history as input and output the dialog state, the dialog action, and the response for evaluation as in Yang et al. (2021). For FBD data, all models take the dia-

Model	Full-train & Test-std		Few-train & Test-std		Full-train & Test-hard		Few-train & Test-hard	
	TurnACC	DialACC	TurnACC	DialACC	TurnACC	DialACC	TurnACC	DialACC
StructBERT	77.13	55.45	63.79	40.44	55.40	13.34	50.60	13.09
Roberta-wwm	78.01	55.68	67.28	44.24	55.02	12.69	53.16	14.58
Our PCM	**78.97**	**58.80**	**69.72**	**46.91**	**57.73**	**15.03**	**56.37**	**16.88**

Table 5: Results on flow-based dialog (FBD) dataset.

Model	Full-train & Test-std		Few-train & Test-std		Full-train & Test-hard		Few-train & Test-hard	
	R@1	MRR	R@1	MRR	R@1	MRR	R@1	MRR
StructBERT	35	54.82	15.6	33.69	9.60	37.16	8.48	26.62
Roberta-wwm	35.3	56.16	17.9	39.30	18.08	43.56	10.85	32.27
Our PCM	**39.1**	**59.54**	**27.7**	**48.02**	**23.50**	**48.60**	**17.80**	**39.7**

Table 6: Results on retrieval-based dialog (RBD) dataset.

log history as input and give the node prediction by adding a linear layer to the pooled [CLS] representation for classification. We leave the utilization of tree structure in future work. For the RBD data, we concatenate the dialog history and each candidate response iteratively to predict a binary label $1/0$ by adding a linear layer to the pooled [CLS] representation, indicating whether the candidate is true or false. In the testing phase, we rank the candidates according to their predicted probability.

4.3 Settings

To measure how well the model performs on complicated dialogs, we pick hard test data from the three datasets in CGoDial to form a specific subset for robustness evaluation. We first run the three baselines (Chinese-T5, Roberta-wwm, and our PCM) to make predictions on all test data, and choose the dialog data with at least one baseline that gives the wrong answer. Then we ask crowdsourcing people to score the data ranging from 1 to 5 according to several criteria: 1) whether it is hard to understand; 2) whether it contains spoken features (e.g., ASR errors) on keywords in utterances. Finally, we choose data scored more than 2.5 as our hard test set.

The sizes of all the pre-trained models are the base scale (12 layers and the 768 hidden embedding dimensions). For SBD, we use the same hyperparameters in the MinTL model for Chinsese-T5 , and the UBAR model for CDial-GPT and our PCM. For FBD and SBD, we use AdamW optimizer for optimization with an initial learning rate of 1e-5. The warm-up proportion is 10%. The batch size is set to 128 and the maximum input length is 512. The hidden size of the output classification head is 128. Best model checkpoints are chosen based on the validation within 20 epochs.

4.4 Results

To evaluate the ability of few-shot adaptation (Geng et al., 2019, 2020) and robustness (Peng et al., 2021b), for all datasets in CGoDial, we employ several evaluation settings: 1) *full-train&test-std*, i.e., using all training data and evaluating on the standard test set. 2) *few-train&test-std*, i.e., using only 10% training set and evaluating on the standard test set. 3) *full-train&test-hard*, i.e., using *full-train* and evaluating on the hard test subset. 4) *Few-train&test-hard*, i.e., using *few-train* and evaluating on the hard test subset.

Table 4 shows the results on slot-based dialog dataset. As we can see, our PCM achieves the best combined scores on different settings, especially in the test-hard setting. Concretely, our model outperforms other baselines on almost all the JGA and Succ metrics, indicating that it has good ability in dialog understanding and policy. Especially in the robust setting, our model obtains 4.83 points improvement (28.00→32.83) in Succ and 4.9 points improvement (49.9→54.8) when evaluating on the test-hard set. We conjecture that our large-scale and harder training corpus makes our PCM perform more reliably in difficult spoken cases. However, our model performs slightly worse on the BLEU metric than Chinese-T5, possibly because UniLM is more suitable for language understanding tasks. When training in the few-shot setting, all models degrades drastically, showing a large space to improve the model few-shot learning ability in spoken task-oriented dialog tasks. Since the results of *few-shot&test-hard* are too low, we neglect this hybrid setting.

Table 5 and 6 show the results on flow-based

Dataset	metric	after augmented	before augmented
SBD	Comb	77.64	90.32
FBD	DialACC	58.8	–
RBD	R@1	39.14	48.59

Table 7: Comparsion of CGoDial and original data. FBD does not have comparison since it is collected from real spoken dialogs originally.

dialog and retrieval-based dialog datasets, respectively. In all settings, our PCM shows its superior performance to a large margin, demonstrating the ability of our PCM on classification dialog tasks in CGoDial. Particularly, on the *few-shot&test-hard* setting, our PCM obtains 2.3 points improvement (14.59→16.88) in DialACC on FBD, and 6.95 points improvement (10.85→17.80) in R@1 on RBD. However, all tasks are still far from an acceptable performance for real applications, which indicates the challenge and difficulty of our benchmark.

In addition, to show that our augmented CGoDial datasets do have more difficulty in language variation and ASR error than non-augmented original dialog data, we run PCM on the *full-train&test-std* setting for each dataset and demonstrate the difference in Table 7.

5 Conclusion

We propose CGoDial, a new large-scale Chinese goal-oriented dialog benchmark featuring three different types of datasets: slot-based dialog, flow-based dialog, and retrieval-based dialog, which are used for comprehensive dialog evaluation. We also propose three different experimental settings: standard training, limited data use, and robustness testing, to assess the dialog model from the aspects of general prediction, fast adaptability, and reliable robustness, respectively. Results from several competitive baselines show the challenge of CGoDial, which is worthy of follow-up research.

Ethical Considerations

The collection of our CGoDial dataset is consistent with the terms of use of any sources and the original authors' intellectual property and privacy rights. The new dataset (FBD) and adapted dataset (SBD, RBD) are collected with the ALIDUTY platform[8], and each crowd-sourcing person requires up to 10 minutes to complete. The source of annotators are mainly from colledge students and professional annotators provided by the platform. The requested inputs are general language variations and speaking voices. No privacy-related information is collected during data collection. Each person was paid 0.1-0.2 USD for a single turn dialog data, which is higher than the minimum wage requirements in our area. The platform also hires professional reviewers to review the collected data to ensure no ethical concerns, e.g., toxic language and hate speech.

[8]Unfortunately, the platform is closed now.

Limitations

From the aspect of the dataset construction, our CGoDial has two main limitations: 1) lacking human-to-human spoken dialogs in real situations. This kind of dialog data has many issues in terms of privacy; thus, it is not easy to open source. 2) For SBD and RBD, we re-build the dataset on other existing textual corpora; therefore, the original datasets largely restrict the final dialog-level flexibility and pattern variation.

From the aspect of the proposed method, we have not used tree structure in the FBD; however, the structural information should be crucial to improve the final overall performance.

References

Alan W Black, Susanne Burger, Alistair Conkie, Helen Hastie, Simon Keizer, Oliver Lemon, Nicolas Merigaud, Gabriel Parent, Gabriel Schubiner, Blaise Thomson, et al. 2011. Spoken dialog challenge 2010: Comparison of live and control test results. In *Proceedings of the SIGDIAL 2011 Conference*, pages 2–7.

Antoine Bordes and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. *ICLR*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, 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*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. 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, Online. Association for Computational Linguistics.

Paul A Crook, Shivani Poddar, Ankita De, Semir Shafi, David Whitney, Alborz Geramifard, and Rajen Subba. 2019. Simmc: Situated interactive multi-modal conversational data collection and evaluation platform. *ASRU*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Yinpei Dai, Hangyu Li, Chengguang Tang, Yongbin Li, Jian Sun, and Xiaodan Zhu. 2020. Learning low-resource end-to-end goal-oriented dialog for fast and reliable system deployment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 609–618, Online. Association for Computational Linguistics.

Yinpei Dai, Zhijian Ou, Dawei Ren, and Pengfei Yu. 2018. Tracking of enriched dialog states for flexible conversational information access. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6139–6143. IEEE.

Bosheng Ding, Junjie Hu, Lidong Bing, Sharifah Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022. Globalwoz: Globalizing multiwoz to develop multilingual task-oriented dialogue systems. *ACL*.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gašić. 2021. Emowoz: A large-scale corpus and labelling scheme for emotion in task-oriented dialogue systems. *arXiv preprint arXiv:2109.04919*.

Ruiying Geng, Binhua Li, Yongbin Li, Jian Sun, and Xiaodan Zhu. 2020. Dynamic memory induction networks for few-shot text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1087–1094, Online. Association for Computational Linguistics.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction networks for few-shot text classification. 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 3904–3913, Hong Kong, China. 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. *arXiv preprint arXiv:2209.06638*.

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. 2022c. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10749–10757.

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 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 35–44, 1st virtual meeting. Association for Computational Linguistics.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019a. A repository of conversational datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Matthew Henderson, Paweł Budzianowski, Inigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, et al. 2019b. A repository of conversational datasets. *arXiv preprint arXiv:1904.06472*.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014a. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014b. The third dialog state tracking challenge. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 324–329. IEEE.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019c. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5392–5404, Florence, Italy. Association for Computational Linguistics.

Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. Multi2woz: A robust multilingual dataset and conversational pretraining for task-oriented dialog. *arXiv preprint arXiv:2205.10400*.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289, 1st virtual meeting. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. Mmconv: An environment for multimodal conversational search across multiple domains. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 675–684.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405, Online. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2106.02787*.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. CharBERT: Character-aware pre-trained language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 39–50, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *arXiv preprint arXiv:2009.13570*.

Shikib Mehri, Tejas Srinivasan, and Maxine Eskenazi. 2019. Structured fusion networks for dialog. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 165–177, Stockholm, Sweden. Association for Computational Linguistics.

Johannes EM Mosig, Shikib Mehri, and Thomas Kober. 2020. Star: A schema-guided dialog dataset for transfer learning. *arXiv preprint arXiv:2010.11853*.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. *AAAI*.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Baolin Peng, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao. 2021b. RADDLE: An evaluation benchmark and analysis platform for robust task-oriented dialog systems. 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 4418–4429, Online. Association for Computational Linguistics.

Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. 2020. RiSAWOZ: A large-scale multi-domain Wizard-of-Oz dataset with rich semantic annotations for task-oriented dialogue modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 930–940, Online. Association for Computational Linguistics.

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

Janarthanan Rajendran, Jatin Ganhotra, and Lazaros C. Polymenakos. 2019. Learning end-to-end goal-oriented dialog with maximal user task success and minimal human agent use. *Transactions of the Association for Computational Linguistics*, 7:375–386.

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

Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. 2005. Let's go public! taking a spoken dialog system to the real world. In *in Proc. of Interspeech 2005*. Citeseer.

Alan Ritter, Colin Cherry, and William B Dolan. 2010. Unsupervised modeling of twitter conversations. In *NAACL*.

Lina Rojas-Barahona, Bo-Hsiang Tseng, Yinpei Dai, Clare Mansfield, Osman Ramadan, Stefan Ultes, Michael Crawford, and Milica Gasic. 2018. Deep learning for language understanding of mental health

concepts derived from cognitive behavioural therapy. *arXiv preprint arXiv:1809.00640*.

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

Izhak Shafran, Mingqiu Wang, Abhinav Rastogi, and Hagen Soltau. 2022. Speech-aware dialog systems technology challenge. *DSTC workshop*.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Chongyang Tao, Jiazhan Feng, Rui Yan, Wei Wu, and Daxin Jiang. 2021. A survey on response selection for retrieval-based dialogues. IJCAI.

Stefan Ultes, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Dongho Kim, Iñigo Casanueva, Paweł Budzianowski, Nikola Mrkšić, Tsung-Hsien Wen, Milica Gašić, and Steve Young. 2017. PyDial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78, Vancouver, Canada. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019b. Glue: A multi-task benchmark and analysis platform for natural language understanding. *ICLR*.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019c. Structbert: incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 91–103. Springer.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 665–677, Vancouver, Canada. Association for Computational Linguistics.

Chien-Sheng Wu, Steven C.H. 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, Online. Association for Computational Linguistics.

Yanan Wu, Zhiyuan Zeng, Keqing He, Hong Xu, Yuanmeng Yan, Huixing Jiang, and Weiran Xu. 2021. Novel slot detection: A benchmark for discovering unknown slot types in the task-oriented dialogue system. 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 3484–3494, Online. Association for Computational Linguistics.

Tian Xie, Xinyi Yang, Angela S Lin, Feihong Wu, Kazuma Hashimoto, Jin Qu, Young Mo Kang, Wenpeng Yin, Huan Wang, Semih Yavuz, et al. 2022. Converse–a tree-based modular task-oriented dialogue system. *arXiv preprint arXiv:2203.12187*.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog systems with gpt-2. *AAAI*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Sai Zhang, Yuwei Hu, Yuchuan Wu, Jiaman Wu, Yongbin Li, Jian Sun, Caixia Yuan, and Xiaojie Wang. 2022a. A slot is not built in one utterance: Spoken language dialogs with sub-slots. *arXiv preprint arXiv:2203.10759*.

Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020a. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9604–9611.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill 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, Online. Association for Computational Linguistics.

Zhenyu Zhang, Bowen Yu, Haiyang Yu, Tingwen Liu, Cheng Fu, Jingyang Li, Chengguang Tang, Jian Sun, and Yongbin Li. 2022b. Layout-aware information extraction for document-grounded dialogue: Dataset, method and demonstration. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7252–7260.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nan Zhao, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. The jddc 2.0 corpus: A large-scale multimodal multi-turn chinese dialogue dataset for e-commerce customer service. *arXiv preprint arXiv:2109.12913*.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. UER: An open-source toolkit for pre-training models. 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): System Demonstrations*, pages 241–246, Hong Kong, China. Association for Computational Linguistics.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.

Lei Zuo, Kun Qian, Bowen Yang, and Zhou Yu. 2021. Allwoz: Towards multilingual task-oriented dialog systems for all. *arXiv preprint arXiv:2112.08333*.

A Appendix A.1

The dataset distributions of dialog length and turn length are shown in the following figures.

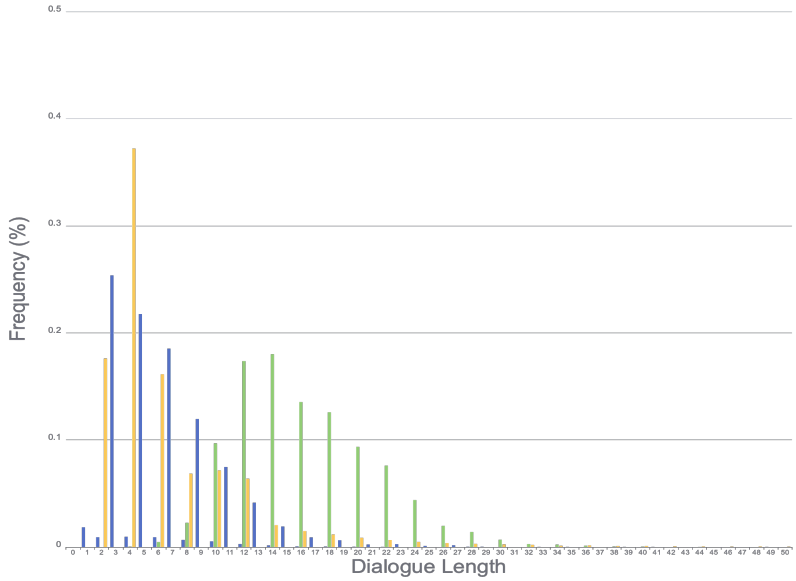

Figure 3: The distribution of the number of turns in the three kinds of dialog in CGoDial: flow-based, slot-based and retrieval-based dialog.

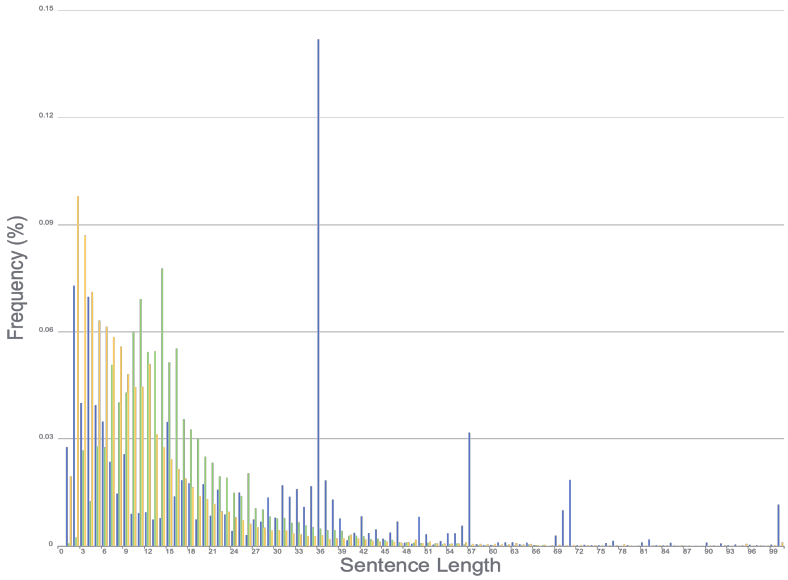

Figure 4: The distribution of the length of turn in the three kinds of dialog in CGoDial: flow-based, slot-based and retrieval-based dialog. The reason for some peaks in flow-based dialog distribution is due to the template response.