Where to Go for the Holidays: Towards Mixed-Type Dialogs for Clarification of User Goals

Zeming Liu[1*†], **Jun Xu**[2*‡], **Zeyang Lei**[2], **Haifeng Wang**[2], **Zheng-Yu Niu**[2], **Hua Wu**[2]
[1]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
[2]Baidu Inc., Beijing, China
zmliu@ir.hit.edu.cn, {xujun03, leizeyang, wanghaifeng, niuzhengyu, wu_hua}@baidu.com

Abstract

Most dialog systems posit that users have figured out clear and specific goals before starting an interaction. For example, users have determined the departure, the destination, and the travel time for booking a flight. However, in many scenarios, limited by experience and knowledge, users may know what they need, but still struggle to figure out clear and specific goals by determining all the necessary slots.

In this paper, we identify this challenge, and make a step forward by collecting a new human-to-human mixed-type dialog corpus. It contains 5k dialog sessions and 168k utterances for 4 dialog types and 5 domains. Within each session, an agent first provides user-goal-related knowledge to help figure out clear and specific goals, and then help achieve them.

Furthermore, we propose a mixed-type dialog model with a novel Prompt-based continual learning mechanism. Specifically, the mechanism enables the model to continually strengthen its ability on any specific type by utilizing existing dialog corpora effectively.

1 Introduction

One of the overarching goals of Artificial Intelligence is to build an intelligent agent that can generate coherent multi-turn dialogs to meet user needs/goals. Recently, multiple dialog agents have been launched, such as Echo and Siri. These agents usually position themselves as some kind of "do engines" that act under users' clear instructions. Specifically, they posit users have figured out clear and specific goals by determining all the necessary aspects or slots of their goals. For example, before booking a flight, a user has determined the departure, the destination and the travel time.

However, such assumption can not hold in many real-world scenarios. For example, a user wants to plan a trip to Beijing for relaxing, but he or she only has limited knowledge about Beijing. Thus it is difficult for him or her to decide which slots are needed to achieve this goal. Obviously, in this scene, the user needs additional consultant services from an agent to help figure out clear and specific goals. However, the aforementioned assumption hinders providing these services effectively.

In this paper, we make a step towards solving the challenge. In order to facilitate the study of how to help users **clarify**ing their goals, we construct a new **Dial**og corpus at Bai**du**, denoted as **DuClarifyDial**. [1] As shown in Figure 1, a user chats about "feels anxious" because of work pressure, and wants to relax himself or herself but have no clear idea about the trip. In the scenario, the agent conducts knowledge-grounded dialogs and question answering conversations to help the user learn more about goal-related knowledge, which helps figure out clear and specific goals. Finally, the user determines to visit "Wangfujing Catholic Church" and books a restaurant nearby. Specifically, in DuClarifyDial, besides basic social chitchat, an agent should help users figure out clear and specific goals by providing goal-related knowledge through coherent knowledge-grounded dialogs and question answering (QA) conversations. Then, upon request, it should also conduct task-oriented dialogs to help achieve user goals.

To this end, we first collect a human-to-human mixed-type dialog dataset. It contains 5k dialog sessions and 168k utterances for 4 dialog types and 5 domains. Specifically, each session contains at least two of following four dialog types, i.e., social chitchat, question answering, knowledge-grounded dialog, and task-oriented dialog. Furthermore, in

*Equal contribution.
† Mainly responsible for dataset collection during his internship at Baidu.
‡Corresponding author: Jun Xu.

[1] https://github.com/PaddlePaddle/Research/tree/master/NLP/ACL2022-DuClarifyDial

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1024 - 1034

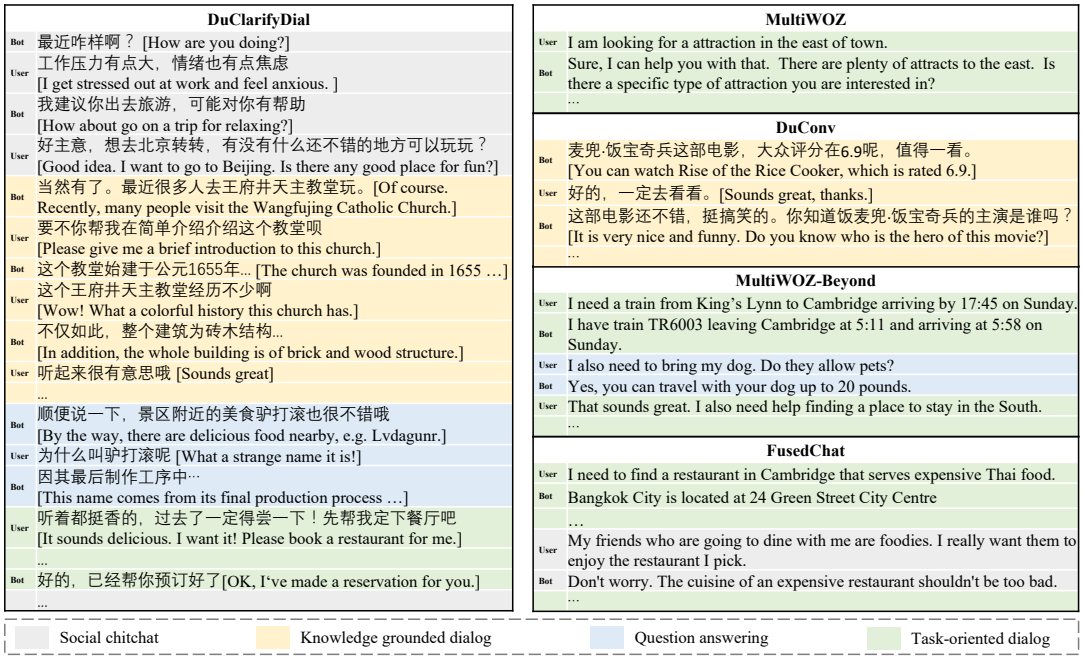

Figure 1: Dialog examples in DuClarifyDial and other dialog corpora. There are four dialog types in a single dialog session of DuClarifyDial while other dialog corpora contain one or two dialog types.

order to seamlessly blend different types of dialogs, we make efforts in both dataset collection and task definition. For *dataset collection*, we first collect human-to-human dialogs within the Wizard-of-Oz framework (Kelley, 1984). Then, we design a unified dialog state schema and dialog act schema for all types of dialogs. Here, the unification can (1) ease the dialog annotation procedures, (2) simplify dialog model design, and (3) facilitate wiser dialog management by bringing a shared dialog semantic space for different types of dialogs. Finally, we annotate dialog states and dialog acts. For *task definition*, we first unify the dialog modelling into three sub-procedures, which includes dialog state tracking, dialog act planning and response generation. Then, we define one sub-task for each sub-procedure. Besides, in order to facilitate end-to-end modelling, we also define an end-to-end dialog generation sub-task.

To facilitate model comparison, we conduct bench-marking experiments on DuClarifyDial for the aforementioned four sub-tasks. Furthermore, since DuClarifyDial is a mixed-type dialog corpus, it is straightforward to explore effective methods for utilizing existing single-type or mixed-types dialog corpora in task modelling. Specifically, we propose a novel Prompt-based continual learning mechanism to strengthen the model ability, by continually utilizing existing different types of dialog corpora. Here, we equip a pre-trained dialog model (Bao et al., 2020) with (1) different prompt texts as input and (2) type, task and domain representation in embedding layer for different dialog types. Furthermore, we train our model by two steps with continual learning mechanism: first Prompting on existing dialog corpora and then fine-tuning on DuClarifyDial.

This work makes the following contributions:

- We identify a new challenge that users have difficulties to figure out all the aspects of their goals in many real-world scenarios.
- We propose a large-scale Chinese mixed-type corpus, where each session weaves together multiple types of dialogs with natural cross-type transitions. Specifically, we design a unified dialog state (act) schema for all types of dialogs. Here, the unified organization first brings a shared semantic space for task-oriented and non-task-oriented dialogs. Then, it enables a unified dialog modelling procedures for all types of dialogs, which can facilitate more effective dialog management.
- We build benchmarking baselines on DuClarifyDial and propose a novel Prompt-based continual learning mechanism to utilize existing dialog corpora effectively.

2 Related Work

2.1 Multi-Domain Task-Oriented Dialog Datasets

Task-oriented dialog systems have continued an active research area for decades and have been consistently supported by the development of new datasets. Recently, several large-scale multi-domain task-oriented dialog datasets have emerged (Budzianowski et al., 2018; Quan et al., 2020; Rastogi et al., 2020; Zhu et al., 2020; Jin et al., 2021; Chen et al., 2021). Specifically, MultiWOZ (Budzianowski et al., 2018) is a fully-labelled collection of human-human written conversations spanning over multiple domains and topics, which contains a size of 10k dialogs. Schema (Rastogi et al., 2020) proposes a schema-guided paradigm for task-oriented dialog, which contains over 16k multi-domain conversations spanning 16 domains. CrossWOZ (Zhu et al., 2020) and RiSAWOZ (Quan et al., 2020) are Chinese cross-domain task-oriented datasets, which contains 6K and 11k dialogs respectively. ABCD (Chen et al., 2021) includes over 10K dialogs that incorporate procedural, dual-constrained actions.

Although achieved promising progress, these datasets usually posit that users have figured out clear and specific goals before staring an interaction, which is not hold in many practical scenarios. In this paper, we focus on providing additional consultant services for users, to help figure out clear and specific user goals.

2.2 Knowledge grounded Dialog Datasets

Open-domain dialog systems have attracted lots of interests in recent years. To develop more human-like dialog models, several knowledge-grounded corpora have been proposed (Wu et al., 2019b; Moon et al., 2019; Liu et al., 2020b; Zhou et al., 2020; yang Wang et al., 2021; Komeili et al., 2021; Feng et al., 2020; Yoshino and Kawahara, 2015; Tanaka et al., 2021). The main purpose on these datasets is to generate more knowledgeable dialogs. In comparison, DuClarifyDial focuses on helping figure out clear and specific user goals. Moreover, DuClarifyDial is a mixed-type dialog dataset that contains four types of dialogs.

2.3 Multi-tasking Dialogs

Recently, there are multiple efforts on developing dialog systems that can multi-task on multiple types of dialogs (Kim et al., 2020; Smith et al., 2020; Mosig et al., 2020; Madotto et al., 2020; Saha et al., 2018; Sun et al., 2021; Young et al., 2021). Specifically, Kim *et al.* (Kim et al., 2020) propose to handle out-of-API requests, by accessing unstructured domain knowledge in task-oriented dialogs. Sun *et al.* (Sun et al., 2021) and Yong *et al.* (Young et al., 2021) propose to fuse task-oriented and open-domain dialogs in conversational agents, in order to generate more engaging and interactive dialogs.

The DuClarifyDial dataset differs from these datasets in that we focus on helping figure out clear and specific user goals, rather than targeting at the out-of-API problem (Kim et al., 2020) or facilitating a more engaging and interactive dialog generation (Young et al., 2021). Furthermore, DuClarifyDial contains more types of dialogs than previous datasets. Moreover, in order to seamlessly blend different types of dialogs for efficient consulting, DuClarifyDial utilizes the same dialog state schema and dialog act schema for all types of dialogs, rather than utilizes different schema for different types of dialogs.

3 The DuClarifyDial Dataset

DuClarifyDial is designed to collect a high quality mixed-type dialog dataset for helping figure out clear and specific goals. In DuClarifyDial, one person serves as the user and the other as the wizard (agent). In order to help figure out clear and specific goals, besides social chitchat, the agent provides user-goal-related information through knowledge grounded dialogs and QA conversations, and then help achieve the goals through task-oriented dialogs.

Specifically, in order to effectively weave together multi types of dialogs for achieving this purpose, it is essential for different types of dialogs to share the same state space and action space. Thus, in Section 3.4, we utilize a unified dialog state schema and dialog act schema for the aforementioned four types of dialogs.

In the following, we will introduce the four steps of DuClarifyDial collection: (1) building knowledge base to provide goal-related information; (2) constructing dialog templates to assist dialog collection; (3) collecting conversation utterances by crowdsourcing; (4) annotating dialog states and dialog acts.

Sub-Scena.	Description
Sub-1: chitchat (Greeting)	The user says that his life was very monotonous.
Sub-2: chitchat(Help decision-making)	Bot suggests users travel. The user doesn't know where to go. Bot suggests the user go to Beijing. User consent.
Sub-3: Task-oriented dialog (Seek tourist attraction)	The user seeks for tourist attractions with high rating. Bot recommends the imperial palace, but the user has been there. Then, bot recommends users to fragrant hills. The user doesn't know fragrant hills.
Sub-4: Knowledge-grounded dialog (about fragrant hills)	The bot and user conduct an in-depth knowledge-grounded dialog about fragrant hills. Finally, the user wants to book tickets.
Sub-5: Task-oriented dialog (Book tickets)	Bot helps the user book tickets.

Table 1: An example dialog template.

3.1 Knowledge Base Construction

In order to create a knowledge base that includes five domains: hotel, attraction, restaurant, food, and movie, we collect publicly available information from the WEB. Specifically, for the hotel domain, we collect 1,133 entities and their related knowledge from two famous online accommodation reservation websites, Qunar and Ctrip. [2][3] For the attraction domain, we collect 435 entities and their related knowledge from the famous travelling website, Mafengwo. [4] For the restaurant domain, we collect 122 entities and their related knowledge from the famous shopping platform, Meituan. [5] For the food domain, we collect 1,971 entities and their related knowledge from the famous online encyclopedia, Baidu Baike. [6] Finally, for the movie domain, we collect 224 entities and their related knowledge from two famous social networking websites, Mtime and Douban. [7][8]

[2]https://www.qunar.com/
[3]https://www.ctrip.com/
[4]http://www.mafengwo.cn/
[5]https://www.meituan.com/
[6]https://baike.baidu.com/
[7]http://www.mtime.com/
[8]https://www.douban.com/

3.2 Dialog Template Construction

Based on the collected knowledge base, we generate dialog templates to guide crowdsourcing workers, which is in line with previous work(Budzianowski et al., 2018; Liu et al., 2020b). Here, each template consists of a sequence of dialog sub-scenarios, and each sub-scenario is defined by a dialog type, a dialog topic and a detailed description text. Table 1 shows an example dialog template. Specifically, in order to better imitate the real scenarios, dialog templates should introduce different interaction behaviours. For example, a user may ask for reserving a ticket during conducting an in-depth knowledge-grounded dialogs around a certain entity, e.g., an attraction. Furthermore, a user may interrupt a task-oriented dialog by chatting about some instant content in mind, and then continue the task-oriented dialog.

In order to construct dialog templates, we first utilize heuristic rules to automatically enumerate candidate sub-scenarios sequences that have natural topic transitions. Then, we utilize pre-defined templates to generate detailed descriptions for these sub-scenarios. Finally, to further ensure natural topic transitions, we manually filter out a few incoherent dialog templates, such as descriptions that contain inconsistent facts.

3.3 Dialog Collection

In order to collect high quality dialogs, we set a strict annotation procedure to guide workers to annotate dialogs based on the given templates. Specifically, the collection procedure includes three stages: (1) reliable crowdsourcing workers recruitment, (2) dialog generation, and (3) quality verification.

In the worker recruitment stage, in order to select reliable workers, we recruit 100 candidates in a famous crowdsourcing platform. [9] Then, we ask each candidate to label 10 dialog sessions based on given templates. Lastly, we employ the top-40 candidates with the highest labelling quality to serve as crowdsourcing workers.

In the dialog generation stage, we develop a labelling interface for crowdsourcing workers to converse synchronously. Then, we randomly pair up two crowdsourcing workers and set each of them a role of the user or the wizard (bot). Lastly, the two crowdsourcing workers generate dialogs with

[9]https://test.baidu.com/

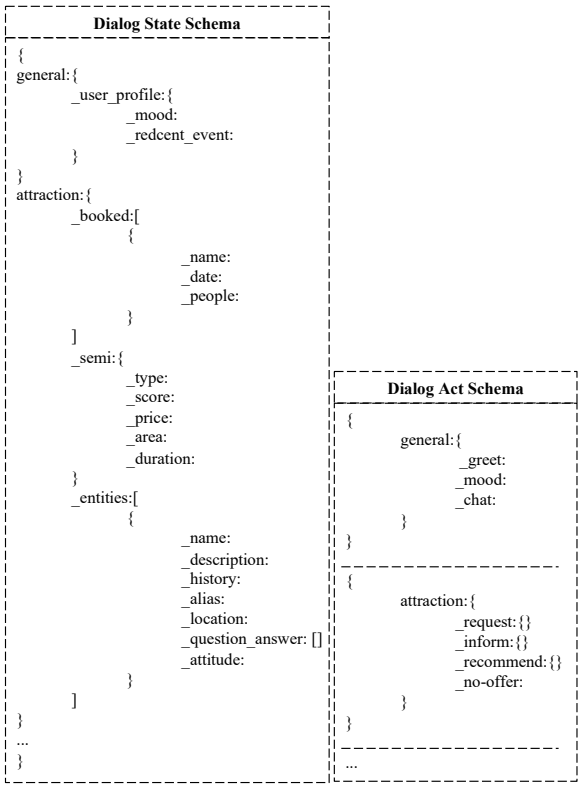

Figure 2: The unified dialog state schema and dialog act schema.

the help of the aforementioned knowledge base and dialog templates.

User Side For a given dialog template, in order to prevent information overload, we only provide a sub-scenario to the user at a time. During dialog collection, a user first reads though the detailed description to understand the provided sub-scenario. Then, based on the given sub-scenario, the user communicates with the wizard turn by turn. Finally, the user may require for another sub-scenario if he or she believes the current sub-scenario has been accomplished. Specifically, in order to diversify the corpus, we encourage the users to follow their own speaking style in communication.

Wizard Side A wizard is required to serve as a consultant, who is responsible for helping users figure out clear and specific goals. At each sub-scenario, the wizard can get access to the associated knowledge in the interface, which is extracted from the knowledge base automatically. When receiving an utterance from the user side, the wizard needs to respond appropriately.

In the quality verification stage, we manually check the collected dialogs. Specifically, if a dialog is considered as unqualified, we will ask the two crowdsourcing workers to revise the dialog until it is qualified.

3.4 Dialog Annotation

After collecting the conversation data, we recruit crowdsourcing workers to annotate dialog states and dialog acts. Specifically, in order to seamlessly blend multi types of dialogs for helping users figure out clear and specific goals, we first design a unified dialog states schema and dialog act schema for all types of dialogs, and then annotate the dialogs based on the schema.

The unified dialog state consists of a list of domain-states, as shown in Figure 2. Specifically, we add a "general" domain to store user-profile related states, e.g., user mood. The "general" domain is important, since user-profile may have a significant impact on his or her goal. For other domains, we split domain-states into three parts: (1)"_booked" for storing booked orders in this domain. Each booked order contains all the necessary information for finishing the order; (2) "_semi" for storing the important but not necessary information for an order; (3) "_entities" for storing all the mentioned entities and the mentioned specific pieces of information about these entities. Specifically, we store an "_attitude" slot in each mentioned entity to capture user interest directly. The values of the "_attitude" slot contain two types: positive and negative. Here, the "_booked" part is mainly corresponding to the task-oriented dialog, the "_entities" part is mainly corresponding to the knowledge grounded dialog and question answering dialog, and the "_semi" part corresponding to all the aforementioned three dialog types.

The unified dialog act schema consists of domains, intents, slots and values. Specifically, we add a "general" domain to store intents that are not directly related to user goals. For other domains, they usually contain four intents: "_request", "_inform", "_recommend" and "_no-offer". Specifically, the classical knowledge selection in knowledge-ground dialog is treated as an "_inform" action in this unified act schema.

Based on the unified schema, we recruit 10 crowdsourcing workers to annotate these dialog states and dialog acts. Specifically, before formal annotation, each worker must pass a labelling test. Here, we first annotate 10 dialogs manually. Then, we ask workers to annotate these dialogs. Lastly, a worker passes the test if his annotations are the same as our annotations.

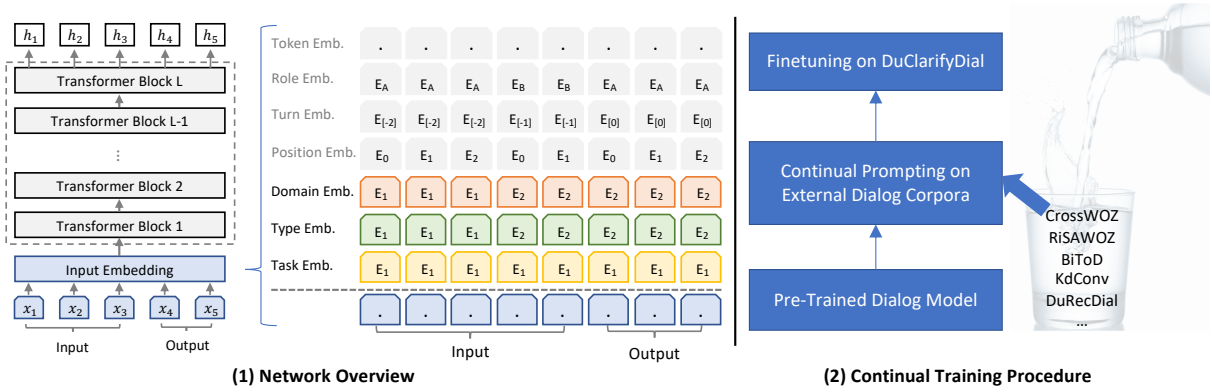

Figure 3: Overview of PLATO-MT

	Train	Dev	Test
# Dialogs	3,500	500	1,052
#Utt.	117,301	16,543	34,999
Avg. utt. per dialog	33	33	33
#Tokens	1,181,669	168,030	352,510
Avg. tokens per utt.	10	10	10
# Chitchat	3,500	500	1,052
# Know.	3,500	500	1,052
# Task	3,500	500	1,052
# QA	214	24	1,052

Table 2: Dataset statistics.

3.5 Overall Dataset

The overall collected data consists of 5,052 dialog sessions in total, with 3,000 sessions in the training set, and validation and test sets of 500 and 1,052 sessions, respectively. Overall statistics can be found in Table 2.

We conduct human evaluations for data quality. Specifically, if a dialog follows the instruction in task templates and the utterances are fluent and grammatical, it will be rated "1", otherwise "0". Then we ask three workers to judge the quality of 200 randomly sampled dialogs. Finally we obtain an average score of 0.83 on this evaluation set.

4 The Mixed-Type Dialog Model with Prompt-based Mechanism

Recently, large scale pre-trained dialog models have achieved impressive performance, both in task-oriented dialog (Heck et al., 2020; Yang et al., 2021) and open-domain chitchat (Adiwardana et al., 2020; Roller et al., 2021; Bao et al., 2020). Meanwhile, the methodologies for different types of dialogs have gradually shifted to generative and end-to-end modelling. Following these trends, we propose a pre-trained mixed-type dialog model based on (Bao et al., 2020), denoted as PLATO-MT. Furthermore, we equip our model with a novel Prompt-based continual learning mechanism to strengthen the model ability by continually utilizing external existed different types of dialog corpora.

4.1 The Prompt-based Continual Learning Mechanism

Figure 3 shows an overview of the proposed PLATO-MT model. As shown in Figure 3 (1), the model is a multi-layer transformer-based neural network. Furthermore, the inputs and outputs of all dialog sub-tasks are formalized as simple text sequences.

In order to effectively blend the abilities of mixed-type dialog in one model, we follow the "Prompt + LM Fine-tuning" strategy (Liu et al., 2021). Specifically, we design different Prompt texts as input for different dialog types. For example, for knowledge-based dialogs, the Prompt text of input is "[Knowledge] context". Here "[Knowledge]" refers to knowledge sentences used for context. Similarly, the Prompt text of QA is "[Question|Answer] context" and the Prompt text of task-oriented dialog is "[Domain|Slot|Value] context". Furthermore, we add type, task and domain embedding representation in embedding layers to further differentiate the characters of different dialog types.

Meanwhile, we train the PLATO-MT model with continuous learning mechanism, as shown in Figure 3 (2). In particular, we first carry on prompting on existing dialog corpora, such as CrossWOZ (Zhu et al., 2020), RiSAWoz (Quan et al., 2020), BiToD (Lin et al., 2021), Kdconv (Zhou et al., 2020) and DurecDial (Liu et al., 2020b). Thus we strengthen our model ability by contin-

ually utilizing external existed different types of dialog corpora. Then we finetune the prompted model on our proposed dialog corpus DuClarifyDial.

5 DuClarifyDial as a New Benchmark

We break down the mixed-type dialog modelling task into three sub-tasks: dialog state tracking, dialog act planning, and dialog-act-to-text generation. Besides, in order to facilitate end-to-end dialog modelling, we define an end-to-end dialog-context-to-text generation sub-task. For each of the four sub-tasks, we report benchmark results on the following dialog models, which have achieved promising performance in the popular MultiWOZ dataset (Budzianowski et al., 2018). Specifically, we use the original codes released by the authors.

UBAR (Yang et al., 2021) UBAR is a fully end-to-end task-oriented dialog model that takes a pre-trained model as backbone. Here, since DuClarifyDial is a Chinese dataset, we utilize a Chinese large-scale pre-trained model, ERNIE (Xiao et al., 2020), to initialize UBAR.

MinTL (Lin et al., 2020) MinTL is a strong model that utilizes effective transfer learning to plug-and-play pre-trained models. Here, instead of utilizing BART (Lewis et al., 2020) as in the original paper, we utilize the multi-lingual version, mBART (Liu et al., 2020a), for initialization.

PLATO (Bao et al., 2020) PLATO is the state-of-the-art Chinese pre-trained dialog model. We use the released parameters. [10]

PLATO-MT It is the proposed unified mixed-type dialog model with Prompt-based Continual Learning mechanism. Here, the Prompt-related parameters are random initialized.

PLATO-MT w/o Prompt It is the PLATO-MT model without Prompting. We first fine-tune it on the same set of existing dialog corpus as in PLATO-MT, and then fine-tune it on DuClarifyDial.

5.1 Dialog State Tracking

For building a successful dialog system, a robust dialog state tracking (**DST**) is considered as the first step. It takes previous dialog utterances and the recent dialog state as input, and then outputs the current dialog state.

To evaluate the performance on dialog state tracking, we utilize both slot-level metric and dialog-level metrics. For slot-level metric, we measure the slot accuracy (**Slot Acc.**). Specifically, the slot accuracy is measured by individually comparing each (domain, slot, value) triplet to its ground truth label. For dialog-level metric, besides dialog type accuracy (**Type Acc.**) and dialog domain accuracy (**Domain Acc.**), we also measure the joint goal accuracy (**Joint Acc.**) (Wu et al., 2019a). It compares the predicted dialog states to the ground truth at each turn, and the output is considered correct if and only if all the predicted values exactly match the ground truth.

Table 3 shows the evaluation results. We can see all the models achieve promising results in terms of "Type Acc." and "Domain Acc.". It indicates the effectiveness of utilizing large-scale pre-trained models as backbone. Furthermore, we notice that PLATO-MT outperforms all the baselines, especially in terms of "Slot Acc." and "Joint Acc.". It demonstrates that PLATO-MT can track dialog states effectively.

5.2 Dialog Act Planning

The dialog act planning (**DAP**) sub-task takes dialog context, current dialog state and retrieved coarse knowledge as input, and then outputs system act. Specifically, for each dialog session, we first extract all the entities in it, and then retrieve all the related knowledge about these entities to serve as the retrieved coarse knowledge.

To evaluate the performance on dialog act planning, we measure the dialog act accuracy (**Act Acc.**) and the BLEU-1/2 (Papineni et al., 2002) score.

Table 3 shows the evaluation results. We notice that PLATO-MT outperforms all the baselines, especially in terms of "Act Acc.". It demonstrates that PLATO-MT can plan appropriate dialog acts effectively.

5.3 Dialog-Act-to-Text Generation

The dialog act to text generation (**RG**) sub-task aims to transform a structured dialog act into a response. It takes dialog context and delexicalized dialog act as input, and then outputs a response.

To evaluate performance on generation, we utilize both automatic metrics and manual metrics. For automatic evaluation, we use several classical metrics, including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), CIDER (Vedantam et al., 2015) and Distinct (**Dist.**) (Li et al., 2016).

[10] https://github.com/PaddlePaddle/Knover/tree/luge-dialog/luge-dialog

Methods	Sub-Task1: DST				Sub-Task2: DAP	
	Type Acc.	Domain Acc.	Slot Acc.	Joint Acc.	Act Acc.	BLEU-1/2
UBAR	0.96	0.95	0.77	0.39	0.85	0.84/0.83
MinTL	**0.99**	0.94	0.86	0.48	0.87	0.83/0.82
PLATO	**0.99**	**0.97**	0.85	0.48	0.91	0.89/0.88
PLATO-MT	**0.99**	**0.97**	**0.88**	**0.51**	**0.93**	**0.90/0.90**
-w/o Prompt	**0.99**	**0.97**	0.87	0.49	0.92	**0.90**/0.89

Table 3: DST and DAP Results on DuClarifyDial.

Methods	Automatic Metrics				Manual Metrics			
	BLEU-1/2	METEOR	CIDER	Dist-1/2	Appr.	Info.	Hallu.	Suc.
UBAR	0.39/0.32	0.21	2.28	0.006/0.040	0.88	0.91	0.45	0.43
MinTL	0.37/0.32	0.21	2.50	0.007/0.079	0.91	0.93	0.89	0.91
PLATO	0.46/0.39	0.25	2.57	0.007/0.072	0.97	0.95	0.90	0.90
PLATO-MT	**0.50/0.43**	**0.27**	**3.00**	**0.008/0.083**	**0.99**	**0.97**	**0.93**	**0.94**
-w/o Prompt	0.46/0.40	0.26	2.84	**0.008**/0.079	0.93	0.96	0.90	0.90

Table 4: RG Results on DuClarifyDial.

For manual evaluation, we conduct evaluation on randomly sampled 50 sessions at the level of both turns and dialogs. For turn-level human evaluation, the generated responses are evaluated by three annotators in terms of appropriateness (**Appr.**) and informativeness (**Info.**). For dialog-level human evaluation, we measure hallucination (**Hallu.**) that measures information accuracy in generated responses, and dialog success (**Suc.**) that measures whether an agent helps users figure out clear goals. Specifically, if a user has not completed any order during a session, the success score is 0; Otherwise, the success score equals to the information accuracy in a session.

Table 4 shows the evaluation results. We find PLATO-MT significantly outperforms all the baselines in terms of all the metrics except "Dist-1/2" (sign test, p-value < 0.01). It indicates that PLATO-MT can generate dialogs with higher qualities.

5.4 End-to-End Dialog Generation

This end-to-end dialog generation sub-task (**E2E-DG**) takes dialog context as input, and then outputs an utterance for responding. Specifically, in the end-to-end settings, since the dialog domain and type information are not available at each turn, we do not use them as input information. Here, we consider the same set of evaluation settings as in Section5.3.

Table 5 shows the evaluation results. We find PLATO-MT significantly outperforms all the baselines in terms of all the metrics except "Dist-1/2" (sign test, p-value < 0.01). Specifically, in terms of "Hallu." and "Suc." in manual evaluation, PLATO-MT outperforms other models by a large margin. It indicates that PLATO-MT is much more competent in helping users learn about correct goal-related knowledge, which is essential for helping users figure clear and specific goals.

5.5 Ablation Study

In order to evaluate the contribution of the proposed Prompt-based continual learning mechanism, we remove the mechanism from PLATO-MT, denoted as "PLATO-MT-w/o Prompt". Here, we first fine-tune PLATO on the same set of existing dialog corpus as in PLATO-MT, and then fine-tune it on DuClarifyDial. For evaluation, we consider the same set of settings as in Section5.3.

As shown in Table 3, Table 4 and Table 5, its performance drops in terms of most metrics in all the four sub-tasks. Specifically, in manual evaluation in Table 5, we notice a sharp performance degradation in terms of "Hallu." and "Suc.". It demonstrates the Prompt-based mechanism is essential for effectively utilizing existing dialog corpora, which enables PLATO-MT can continually strengthen its ability on any specific dialog type.

Furthermore, we find that, in terms of most metrics, the mechanism gains more in the end-to-end conversation generation sub-task than in the other three sub-tasks. This is because there are no avail-

Methods	Automatic Metrics				Manual Metrics			
	BLEU-1/2	METEOR	CIDER	Dist-1/2	Appr.	Info.	Hallu.	Suc.
UBAR	0.28/0.22	0.16	1.70	0.005/0.031	0.74	0.87	0.32	0.34
MinTL	0.32/0.25	0.17	1.80	0.006/0.046	0.86	0.88	0.35	0.34
PLATO	0.32/0.25	0.16	1.28	0.005/0.034	0.78	0.88	0.36	0.36
PLATO-MT	**0.45/0.37**	**0.23**	**2.17**	**0.007/0.072**	**0.96**	**0.90**	**0.67**	**0.69**
-w/o Prompt	0.41/0.33	0.21	1.89	**0.007**/0.062	0.87	0.89	0.55	0.52

Table 5: E2E-DG Results on DuClarifyDial.

able annotated information in the end-to-end conversation generation sub-task, which makes it a more difficult task. Thus, the effect of Prompt-based continual mechanism appears relatively more significant.

6 Conclusion

In this paper, we first identify the challenge that users may struggle to figure out clear and specific goals in many real scenarios. Then, we make a step forward by collecting a new human-to-human mixed-type dialog corpus, which contains 5k dialog sessions and 168k utterances for 4 dialog types and 5 domains. Furthermore, we setup benchmarks based on the corpus. Moreover, we propose a mixed-type dialog generation model with a novel Prompt-based continual learning mechanism. Finally, experimental results demonstrate the effectiveness of the mechanism.

Ethical Considerations

We make sure that *DuClarifyDial* has been collected in a manner that is consistent with the terms of use of any sources and the intellectual property and privacy rights of the original authors of the texts. And crowd workers were treated fairly. This includes, but is not limited to, compensating them fairly and ensuring that they were able to give informed consent, which includes, but is not limited to, ensuring that they were voluntary participants who were aware of any risks of harm associated with their participation. Please see Section 3 for more details characteristics and collection process of *DuClarifyDial*.

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*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

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, Alex Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *NAACL*.

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, Online. Association for Computational Linguistics.

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.

Di Jin, Shuyang Gao, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Towards zero and

few-shot knowledge-seeking turn detection in task-orientated dialogue systems. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 281–288.

John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

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.

M. Komeili, Kurt Shuster, and J. Weston. 2021. Internet-augmented dialogue generation. *ArXiv*, abs/2107.07566.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

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*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020b. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049, Online. Association for Computational Linguistics.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, Jamin Shin, and Pascale Fung. 2020. Attention over parameters for dialogue systems. *arXiv preprint arXiv:2001.01871*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. 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*.

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.

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.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *AAAI*.

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 *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Amrita Saha, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18),*

and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 696–704. AAAI Press.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Kai Sun, Seungwhan Moon, Paul A. Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chat to enhance task-oriented dialogues. In *NAACL*.

Shohei Tanaka, Koichiro Yoshino, Katsuhito Sudoh, and Satoshi Nakamura. 2021. ARTA: Collection and classification of ambiguous requests and thoughtful actions. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 77–88, Singapore and Online. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019a. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy. Association for Computational Linguistics.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019b. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Dongling Xiao, Han Zhang, Yu-Kun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE-GEN: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3997–4003. ijcai.org.

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

yang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. In *AAAI*.

Koichiro Yoshino and Tatsuya Kawahara. 2015. Conversational system for information navigation based on pomdp with user focus tracking. *Computer Speech & Language*, 34(1):275–291.

Tom Young, Frank Xing, Vlad Pandelea, Jinjie Ni, and Erik Cambria. 2021. Fusing task-oriented and open-domain dialogues in conversational agents. *arXiv preprint arXiv:2109.04137*.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7098–7108, Online. 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.