UniDS: A Unified Dialogue System for Chit-Chat and Task-oriented Dialogues

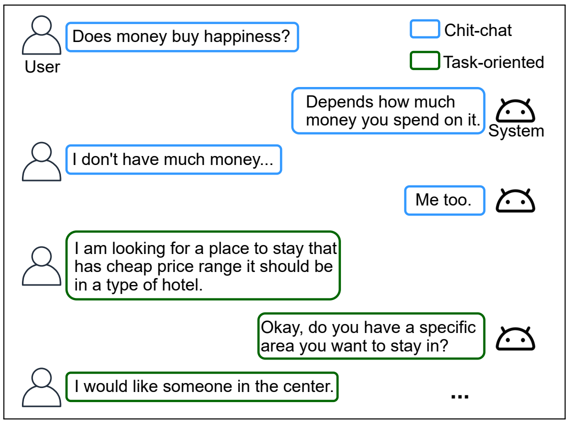
Xinyan Zhao[1][*], **Bin He**[2], **Yasheng Wang**[2], **Yitong Li**[2], **Fei Mi**[2], **Yajiao Liu**[2],
Xin Jiang[2], **Qun Liu**[2], **Huanhuan Chen**[1]
[1] University of Science and Technology of China
[2] Huawei Noah's Ark Lab
sa516458@mail.ustc.edu.cn,
{hebin.nlp, wangyasheng, liyitong3, mifei2, yajiao.liu, Jiang.Xin, qun.liu}@huawei.com,
hchen@ustc.edu.cn

Abstract

With the advances in deep learning, tremendous progress has been made with chit-chat dialogue systems and task-oriented dialogue systems. However, these two systems are often tackled separately in current methods. To achieve more natural interaction with humans, dialogue systems need to be capable of both chatting and accomplishing tasks. To this end, we propose a **uni**fied **d**ialogue **s**ystem (**UniDS**) with the two aforementioned skills. In particular, we design a unified dialogue data schema, compatible for both chit-chat and task-oriented dialogues. Besides, we propose a two-stage training method to train UniDS based on the unified dialogue data schema. UniDS does not need to adding extra parameters to existing chit-chat dialogue systems. Experimental results demonstrate that the proposed UniDS works comparably well as the state-of-the-art chit-chat dialogue systems and task-oriented dialogue systems. More importantly, UniDS achieves better robustness than pure dialogue systems and satisfactory switch ability between two types of dialogues. This work demonstrates the feasibility and potential of building a general dialogue system.

1 Introduction

Dialogue system is an important tool to achieve intelligent user interaction, and it is actively studied by NLP and other communities. Current research of dialogue systems focus on task-oriented dialogue (TOD) systems (Hosseini-Asl et al., 2020; Peng et al., 2020; Yang et al., 2021), achieving functional goals, and chit-chat dialogue systems aiming at entertainment (Zhou et al., 2018; Zhang et al., 2020; Zhao et al., 2020; Roller et al., 2021). Different methods are devised for these two types of dialogue systems separately. However, a more suitable way for users would be to have one dialogue agent that is able to handle both chit-chat and TOD in one conversation. As illustrated in Figure 1, users may have communication-oriented needs (e.g. chatting about money and happiness) and task-oriented needs (e.g. hotel reservation) when interacting with a dialogue agent. Furthermore, inputs of dialogue systems are often interfered by background noise, such as voice from other people or devices, collected by the preceding automatic speech recognition (ASR) module. Therefore, the chit-chat ability may also improve the robustness of a task-oriented dialog system (Zhao et al., 2017).



Does money buy happiness?

Depends how much money you spend on it.

I don't have much money...

Me too.

I am looking for a place to stay that has cheap price range it should be in a type of hotel.

Okay, do you have a specific area you want to stay in?

I would like someone in the center.

Figure 1: Illustration of users being interested to chit-chat with the dialogue system before booking a hotel.

As shown in Table1, there are many differences between chit-chat and task-oriented dialogues. Creating a single model for different tasks without performance degradation is challenging (Kaiser et al., 2017). Some works attempt to model different dialogue skills via different experts or adapters (Madotto et al., 2020; Lin et al., 2021). However, these methods increase the number of parameters and hard to achieve satisfactory performance on both types of dialogues. Besides, previous work lack the exploration of the ability to switch between different types of dialogues.

This work proposes a auto-regressive language model based dialogue system to handle chit-chat

---

[*] This work was done during an internship at Huawei Noah's Ark Lab.

Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 13 - 22
May 26, 2022 ©2022 Association for Computational Linguistics

	Diversity	**Purpose**	**Turns**	**Mainstream method**
Chit-chat	Strong	Entertainment	Long	End-to-end method
Task-oriented dialogue	Weak	Completing tasks	Short	Pipeline method*

Table 1: Differences between chit-chat and task-oriented dialogues. *: The model will predict belief state and system act before giving a response, to this end, the training set needs to be annotated with belief state and system act.

and TOD in a unified framework (UniDS). Specifically, since chit-chat data do not have explicit belief state and agent action, to unify chit-chat and task-oriented dialogues format, we device belief state and agent act for chit-chat dialogues as task-oriented dialogues. On the other hand, because of the diversity of chit-chat, chit-chat dialogue systems need more training data than task-oriented dialogue systems, e.g., 147,116,725 dialogues for DialoGPT (Radford et al., 2019) and 8,438 dialogues for UBAR (Yang et al., 2021). To overcome this difference, we propose to train UniDS in a two-stage way. A chit-chat model is first trained with huge chit-chat dialogues, and then we train UniDS from the chit-chat dialogue system with mixed dialogues based on our proposed unified dialogue data schema.

We evaluate UniDS using a public task-oriented dialogue dataset MultiWOZ and a chit-chat dataset extracted from Reddit[1] through both automatic and human evaluations. UniDS achieves comparable performance compared to the state-of-the-art chit-chat dialogue system DialoGPT, and TOD system UBAR. In addition, we empirically show that UniDS is more robust to noise in task-oriented dialogues, and UniDS shows a desirable ability to switch between the two types of dialogues.

The contributions of this work are summarised as follows:

- To the best of our knowledge, this is the first work presenting a unified dialogue system to jointly handle chit-chat and task-oriented dialogues in an end-to-end way.
- We design a *unified dialogue data schema* for chit-chat and TOD, allowing the training and inference of dialogue systems to be performed in a unified manner.
- To tackle the gap between chit-chat dialogue systems and task-oriented dialogue systems in the requirement of training data, a two-stage training method is proposed to train UniDS.
- Extensive empirical results show that UniDS performs comparably to state-of-the-art chit-chat dialogue systems and task-oriented dialogue systems. Moreover, UniDS achieves better robustness to dialog noise and satisfactory switch ability between two types of dialogues.

2 Related Work

With the development of large-scale language models, chit-chat dialogue systems achieve remarkable success. Based on GPT-2 (Radford et al., 2019), DialoGPT (Zhang et al., 2020) is further trained on large-scale dialogues extracted from Reddit. DialoGPT could generate more relevant, contentful, and fluent responses than previous methods. Afterwards, larger pre-train LM based chit-chat dialogue systems (Adiwardana et al., 2020; Bao et al., 2020; Roller et al., 2021) are proposed and achieve even better performance. In the area of task-oriented dialogue systems, recent research (Hosseini-Asl et al., 2020; Peng et al., 2020; Yang et al., 2021) concatenated elements in a dialogue into one sequence and utilized pre-train LM to generate the belief state, system act, and response in an end-to-end way and achieved promising results.

There are several works related to the unified dialogue system. Zhao et al. (2017) insert one turn chit-chat dialogue into task-oriented dialogues to train a model with better out-of-domain recovery ability. Attention over Parameters (AoP) (Madotto et al., 2020) utilizes different decoders for different dialogue skills (e.g., hotel booking, restaurant booking, chit). However, the performance of AoP can be improved and it largely increases parameters comparing with models that handle a single type of dialogues. ACCENTOR (Sun et al., 2021) adds chit-chat utterance at the beginning or end of task-oriented responses to make the conversation more engaging, but ACCENTOR is unable to have a chit-chat with users. Unlike the above works, UniDS does not add extra parameters to existing dialogue models, and UniDS could alternatively handle chit-chat and task-oriented dialogues in a

---

[1] https://www.reddit.com/

seamless way.

3 Unified Dialogue System

3.1 Architecture of UniDS

As illustrated in Figure 2, we formulate unified dialogue system as an auto-regressive language model. A dialogue session at turn t has the following components: user input U_t, belief state B_t, database search result D_t, system act A_t, and response R_t. Each component consists of tokens from a fixed vocabulary. For turn t, the dialogue context C_t is the concatenation of all the components of the previous dialogues as well as the user input at turn t: $C_t = [U_0, B_0, D_0, A_0, R_0, \cdots, R_{t-1}, U_t]$. Given the dialogue context C_t, UniDS first generates the belief state B_t:

$$B_t = \text{UniDS}(C_t)\,, \tag{1}$$

and use it to search the database to get the search result D_t. Then, UniDS generates the system act A_t conditioned on the updated context by extending C_t with B_t and D_t:

$$A_t = \text{UniDS}([C_t, B_t, D_t])\,. \tag{2}$$

Lastly, the response R_t is generated conditioned on the concatenation of all previous components:

$$R_t = \text{UniDS}([C_t, B_t, D_t, A_t])\,. \tag{3}$$

3.2 Unified Dialogue Data Schema

In the widely adopted task-oriented dialogue system pipeline, a dialogue session consists of a user input utterance, a belief state that represents the user intention, a database search result, a system act, and a system response (Young et al., 2013; Yang et al., 2021). However, due to the diversity of chit-chat and the cost of manual annotation, chit-chat dialogue systems do not assume the existence of the belief state nor system act (Bao et al., 2020; Zhang et al., 2020). The inconsistency of data format between chit-chat and TOD hinders the implementation of a unified model. To tackle this problem, we design a data schema with belief state, database result representation and system act for chit-chat. Table 2 illustrates such unified data schema with examples. The following sections explain each component in detail.

3.2.1 Belief state

The unified belief state is represented in the form of "<domain> slot [value]". A belief state could have several domains, each containing several slot-value pairs. As we can observe, extracting belief state of TOD may need to copy some words from the user utterance. To make UniDS keep this copy mechanism, for chit-chat, nouns in the user utterance U_t are extracted as the slot or value of belief state.

3.2.2 DB result

We use a special token to represent the number of matched entities under the constraints of the belief state in the current turn.

3.2.3 System act

System acts are represented as "<domain> <act> [slot]" for TOD. The meaning of "<domain>" is the same as in belief states. "[act]" denotes the type of action the system needs to perform. Following the "domain-act" pair, slots are optional. For chit-chat, token "<chit_act>" denotes the dialogue system will chat with the user.

Therefore, a processed dialogue sequence X_t at turn t for either TOD or chit-chat can be both represented as:

$$X_t = [C_t, B_t, D_t, A_t, R_t]. \tag{4}$$

3.3 Two-stage training method

Since the diversity of chit-chat in topics and terms, chit-chat dialogue systems need much larger training data than task-oriented dialogue systems. If directly training UniDS with the unified dialogue data which contains much more chit-chat dialogues than task-oriented dialogues, the trained model may ignore the ability to complete task-oriented dialogues. Therefore, this work proposes a two-stage method for training UniDS. As illustrated in Figure 3, we propose to first train a chit-chat dialogue model with huge chit-chat dialogues, and then we train UniDS from the chit-chat dialogue system with mixed dialogues. The mixed dialogue data is obtained by mixing chit-chat and TOD data which are pre-processed by the proposed unified data schema in the ratio of 1:1. Motivated by the recent success of applying GPT-2 for task-oriented dialogue systems (Hosseini-Asl et al., 2020; Peng et al., 2020; Yang et al., 2021) and chit-chat dialogue systems (Zhang et al., 2020), we use DialoGPT(Zhang et al., 2020) as our chit-chat model, and train UniDS from DialoGPT.

The training objective for UniDS is to maximize the joint probability of all tokens in X_t computed

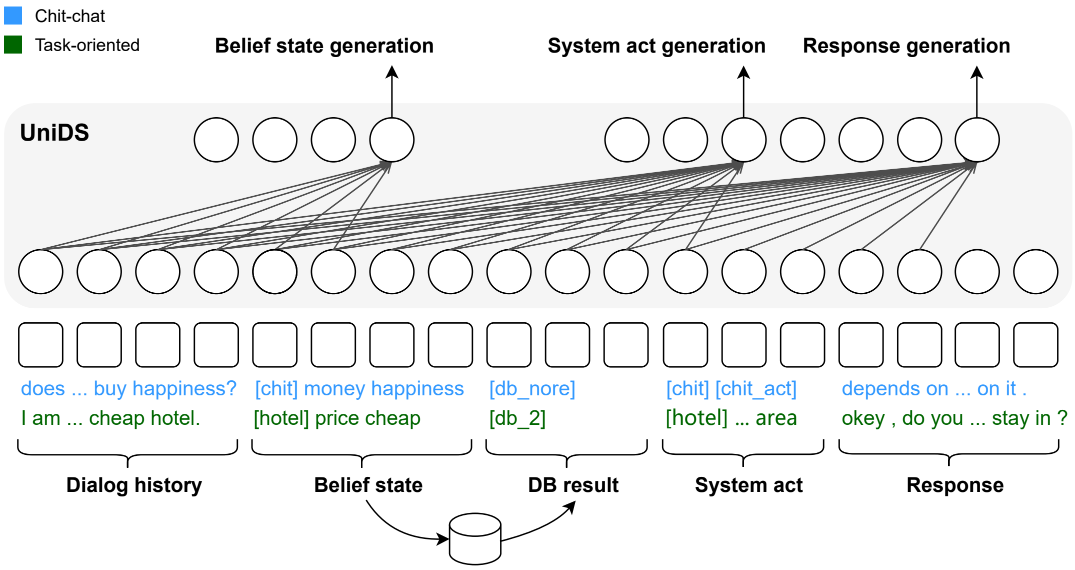


Figure 2: The architecture of UniDS.

	Unified dialogue data schema	Chit-chat example	Task-oriented example
User input	Tokenized utterance	does money buy happiness ?	i am looking for a cheap hotel .
Belief state	<domain> slot [value]	<chit> money happiness	<hotel> price cheap
DB result	A token indicated the number of candidate entities	<db_nore>	<db_2>
Act	<domain> <act> [slot]	<chit> <chit_act>	<hotel> <request> area
Response	Tokenized utterance	depends on how much money you spend on it .	do you have a specific area you want to stay in ?

Table 2: Unified dialogue data schema (where tokens inside the square bracket are optional) and examples.

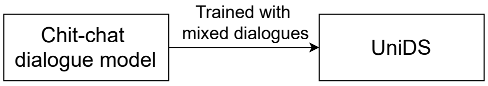


Figure 3: Training process of UniDS.

in an auto-regressive manner as:

$$\mathcal{L} = \sum_{i=1}^{N} -\log P(\boldsymbol{x}_i|\boldsymbol{x}_{<i}) \,, \tag{5}$$

where $\boldsymbol{x}_i$ is a token of X_t, and $\boldsymbol{x}_{<i}$ are the preceding tokens.

4 Experiment

4.1 Datasets

4.1.1 Task-oriented Dialogue Dataset

For task-oriented dialogues, we adopt the publicly multi-domain task-oriented MultiWOZ (Budzianowski et al., 2018), which consists of $10,438$ dialogues spinning over seven domains (*taxi, attraction, police, restaurant, train, hotel, hospital*).[2] The train/validation/test sets of MultiWOZ have $8438/1000/1000$ dialogues, respectively. Each dialogue contains 1 to 3 domains.

4.1.2 Chit-chat Dataset

We derived open-domain chit-chat dialogue from Reddit dump[3]. To avoid overlapping, the chit-chat training set and test set are extracted from the Reddit posts in 2017 and 2018 respectively. To ensure the generation quality, we conduct a careful data cleaning. A conversation will be filtered when (1) there is a URL in the utterance; (2) there is an utterance longer than 200 words or less than 2 words; (3) the dialogue contains "[removed]" or "[deleted]" tokens; (4) the number of utterances in the dialogue is less than 4; (5) the dialogue contains offensive words. Finally, we sample $8,438$ dialogues for training which is the same size as the training set of MultiWOZ. The validation set and test set contain $6,000$ dialogues and $8,320$ dialogues, respectively.

4.2 Baselines

For chit-chat dialogue, we compare UniDS with **DialoGPT** (Zhang et al., 2020). For fair comparisons,

---

[2] We use MultiWOZ 2.0.

[3] https://files.pushshift.io/reddit/comments/

Model	# of para.	Task-oriented Dialogue				Chit-chat			
		Inform	Success	BLEU	Combined	BLEU	Dist-1	Dist-2	AvgLen
UBAR*	82M	91.5	77.4	17.0	101.5	-	-	-	-
PPTOD	∼220M	89.20	79.40	18.62	102.92	-	-	-	-
UBAR-12L	117M	**89.40**	75.10	16.93	99.18	-	-	-	-
DialoGPT-12L	117M	-	-	-	-	0.27	**6**	**32**	14.00
UniDS-12L	117M	87.10	**77.00**	**18.01**	**100.06**	**0.35**	**6**	30	12.00
UBAR-24L	345M	89.40	75.50	16.86	99.31	-	-	-	-
DialoGPT-24L	345M	-	-	-	-	0.43	**7**	**36**	12.28
UniDS-24L	345M	**90.30**	**80.50**	**18.72**	**104.12**	**0.45**	6	35	14.62

Table 3: Automatic evaluations of UniDS with two model sizes over two types of dialogue datasets. All results are reported in percentage, except Combined and AvgLen. Best results are in **bold**. *: Results reported in original paper (Yang et al., 2021) is not obtained by end-to-end evaluation. This result is reported by authors of UBAR in https://github.com/TonyNemo/UBAR-MultiWOZ/issues/3.

we further fine-tune a 12-layer DialoGPT and a 24-layer DialoGPT with our chit-chat dialogue training set, which we refer to as DialoGPT-12L and DialoGPT-24L, respectively.

For TOD, we consider the state-of-the-art end-to-end TOD system **UBAR** (Yang et al., 2021) and **PPTOD**(Su et al., 2021). For a fair comparison with UniDS, we also fine-tune UBAR from 12 layers DialoGPT and 24 layers DialoGPT with MultiWOZ dataset, the fine-tuned models are denoted as UBAR-12L and UBAR-24L, respectively.

4.3 Implementation Details

UniDS and other baselines are implemented based on HuggingFace's Transformers (Wolf et al., 2019). The max sequence length is 1024 and sequences longer than 1024 are truncated from the head. We use the AdamW optimizer (Loshchilov and Hutter, 2019) and greedy decoding method for inference. All models are trained on a single Tesla V100, and we perform a hyper-parameter search on batch size and learning rate. The best model and hyper-parameter are selected through the performance on the validation set of MultiWOZ only.

As shown in Table1, chit-chat dialogues need to attract users to talk more, while TOD needs to complete tasks as soon as possible. Therefore, a model trained with the mixed dialogue data tends to talk long turns instead of efficiently completing the task. Since entity recommendation acts are important for dialogue system to complete tasks efficiently, we use a weighted cross-entropy loss as the training objective of UniDS. We assign larger weights to tokens about entity recommendation actions. We empirically set the weight of entity recommendation actions in loss function to 2^4, weights of other tokens are set to 1 by default.

4.4 Evaluation Metrics

For chit-chat dialogues, the BLEU score (Papineni et al., 2002) and the average length of the generated responses are reported. Because of the diversity of chit-chat, BLEU may be difficult to reflect the quality of chit-chat responses, we also report distinct-1 and distinct-2 (Li et al., 2016) of generated dialogues, which is defined as the rate of distinct uni- and bi-grams in the generated sentences. We also conduct a human evaluation on 50 randomly sampled test dialogues for two 24 layers models. Three judges evaluate them in terms of relevance, informativeness, and how human-like the response is with a 3-point Likert-like scale (Joshi et al., 2015).

For TOD, we follow UBAR to use the following automatic metrics: **Inform** refers to the rate of the entities provided by a model are correct; **success** measures the rate of a model has answered all the requested information; and **BLEU** to measure the fluency of generated responses. A **combined** score is computed as $(\text{Inform} + \text{Success}) \times 0.5 + \text{BLEU}$ to measure the overall response quality.

4.5 Overall results

Table 3 presents the overall comparison results of automatic evaluation. The first block shows the results of UBAR. The following two blocks are various baselines trained on 12 or 24 layers DialoGPT respectively. From these results, we have the following observations.

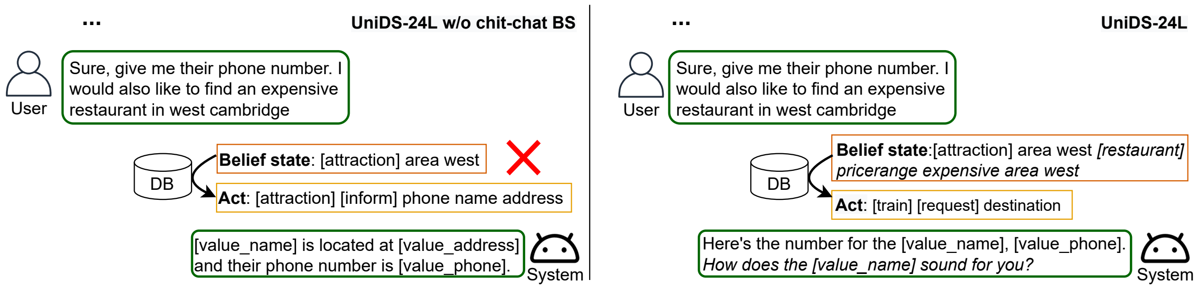
i) For the chit-chat task, UniDS achieves comparable performance with DialoGPT. For the BLEU score, UniDS outperforms DialoGPT with 12L and 24L. On other metrics, UniDS is comparable with DialoGPT. This demonstrates that UniDS can still keep strong chit-

---

[4] The appendix gives discussions for other values of weight, but does not affect the overall conclusion.

Model	Task-oriented Dialogue				Chit-chat			
	Inform	Success	BLEU	Combined	BLEU	Dist-1	Dist-2	AvgLen
UniDS-12L	87.10	77.00	18.01	100.06	0.35	6	30	12.00
w/o chit-chat BS	83.90	72.80	18.15	96.50	0.37	5	29	14.67
w/o weighted loss	81.70	71.20	17.93	94.38	0.33	6	32	14.29
UniDS-24L	90.30	80.50	18.72	104.12	0.45	6	35	14.62
w/o chit-chat BS	86.90	78.50	18.71	101.41	0.49	6	33	15.29
w/o weighted loss	85.60	76.50	18.96	100.01	0.44	6	34	14.85

Table 4: Ablation studies of automatic evaluations for UniDS.



Sure, give me their phone number. I would also like to find an expensive restaurant in west cambridge

Belief state: [attraction] area west

Act: [attraction] [inform] phone name address

[value_name] is located at [value_address] and their phone number is [value_phone].

Sure, give me their phone number. I would also like to find an expensive restaurant in west cambridge

Belief state:[attraction] area west *[restaurant] pricerange expensive area west*

Act: [train] [request] destination

Here's the number for the [value_name], [value_phone]. *How does the [value_name] sound for you?*

Figure 4: TOD examples from UniDS w/o chit-chat BS and UniDS. UniDS w/o chit-chat BS does not extract the user intent of searching restaurants, but UniDS extracts this intent successfully (highlighted in italics).

	DialoGPT-24L (Win %)	Neutral (%)	UniDS-24L (Win %)
Relevance	25.33	**42.67**	32.00
Informativeness	29.33	33.33	**37.34**
Human-like	26.67	**43.33**	30.00

Table 5: Win rate [%] between the UniDS-24L and DialoGPT-24L using three human evaluation metrics on chit-chat dialogues. "Neutral" means the generated responses of DialoGPT-24L and UniDS-24L are considered to have equal quality.

chat ability even after training with the mixed dialogue data.

ii) For the TOD task, UniDS achieves better performance than UBAR for the same parameter size. For both 12L and 24L DialoGPT, UniDS improves the BLEU score and the Combined score compared with UBAR. We believe this is because combining chit-chat dialogues for training helps the model to generate more fluent responses.

Furthermore, we also provide the human evaluation results in Table 5. UniDS is compared to DialoGPT regarding three dimensions for chit-chat dialogues. We could see that UniDS consistently wins the majority cases for all three aspects, including relevance, informativeness, and human-like.

4.6 Analysis

4.6.1 Ablation Study

In this experiment (c.f. Table 4), we compare two simplified versions of UniDS to understand the effects of different components. For comparison, we report the performance of 1) removing slots in belief state of chit-chat, denoted as "UniDS w/o chit-chat BS", and 2) replacing the weighted cross-entropy loss with a standard cross-entropy loss, denoted as "UniDS w/o weighted loss". Next, we elaborate our observations w.r.t. these two components.

w/o chit-chat BS: When removing the belief state of chit-chat dialogues, the performances of both UniDS-12L and UniDS-24L drop w.r.t. inform, success, and combined score for TOD. We believe the reason is that the process of extracting the belief state needs to copy some keywords from the user utterance, and even extracting nouns as belief state for chit-chat is helpful for UniDS to learn this copy mechanism in the TOD task. Taking the case in Figure 4 as an example, UniDS w/o chit-chat BS (left) fails to extract the user's interest in searching restaurants, while UniDS (right) extracts the restaurant slot successfully. As a result, UniDS could recommend the right entities. Furthermore, removing chit-chat BS does not degrade the performance of chit-chat.

UniDS	Inf.	Succ.	BLEU	Comb.	Switch-1	Switch-2
12L	84.60	72.00	11.72	90.02	65.8	99.5 (+33.7)
24L	85.30	75.70	12.44	92.94	64.4	99.2 (+34.8)

Table 6: Switching performance of UniDS when having 2 turns chit-chat dialogues before task-orientated dialogues. Numbers in brackets indicates the exactly switching rate at the 2nd turn.

UniDS	BLEU	Dist-1	Dist-2	AvgLen	Switch-1	Switch-2
12L	0.22	4	19	14.15	31.8	98.9 (+67.1)
24L	0.34	6	31	16.18	37.0	96.6 (+59.6)

Table 7: Switching performance of UniDS when prepending 2 turns task-oriented dialogues before chit-chat.

w/o weighted loss: When replacing the weighted cross-entropy loss in UniDS with standard cross-entropy loss, we observe a notable drop w.r.t. inform, success, and combined score in task-oriented metrics. These results demonstrate that giving more attention to entity recommendation acts helps the task completion capability. Moreover, dropping the weight loss does not affect the performance of chit-chat much.

Overall, we contend both "chit-chat BS" and "weighted loss" are beneficial for task-oriented dialogues without degrading the chit-chat capability.

4.6.2 Analysis of Switching Ability

In real-world scenarios, it is common and natural for users to switch between chit-chat and task-oriented dialogues. Therefore, we investigate the switch ability of UniDS in this subsection. To simulate the scenario of dialogue switching, we consider two setups: (1) having two turns of chit-chat dialogues before the start of a task-oriented dialogue and (2) pre-pending two turns of task-oriented dialogues at the beginning of a chit-chat dialogue. To evaluate the model's ability to switch between two types of dialogues, we propose a metric, called **Switch-n**, which is defined as the rate of a model switches its response type within the first n turns after a user switches the type of input. Additionally, we also report the model performance *after* the switching.

Tables 6 and 7 present the results of the two switching setups, and we have the following observations:

(i) It is not surprising that adding switching tasks for both chit-chat and TOD degrades the performance of UniDS, as the added 2 turns of switching utterances introduce irrelevant content, which distracts the model. However, focusing on the switching task, we observe that for almost 98% of cases, UniDS can success in dialogue task switching, from chit-chat to TOD and vice versa, within the first two turns (Switch-1 and Switch-2). This demonstrates UniDS has a good ability to switch between two types of dialogue tasks.

User@Turn t:
Will the Red Tesla get me to the restaurant for my 16:15 reservation?
Response generated by UniDS:
Yes, your taxi will pick you up at all saints church and arrive at tang chinese by 16:15. Is there anything else I can do for you today?

User@Turn $t+1$:
Ok sounds great. That is all I need. Thank you!
Response generated by UniDS:
Thank you for using our services.

User@Turn $t+2$:
How's your new year's resolution holding up?
Response generated by UniDS:
It's holding up just fine. Thanks for inquiring with us.

Table 8: Example of UniDS when switching from the task-oriented dialogue to *chit-chat*. UniDS gives a chatty response and thanks the user for using its services. Dialogue history is omitted.

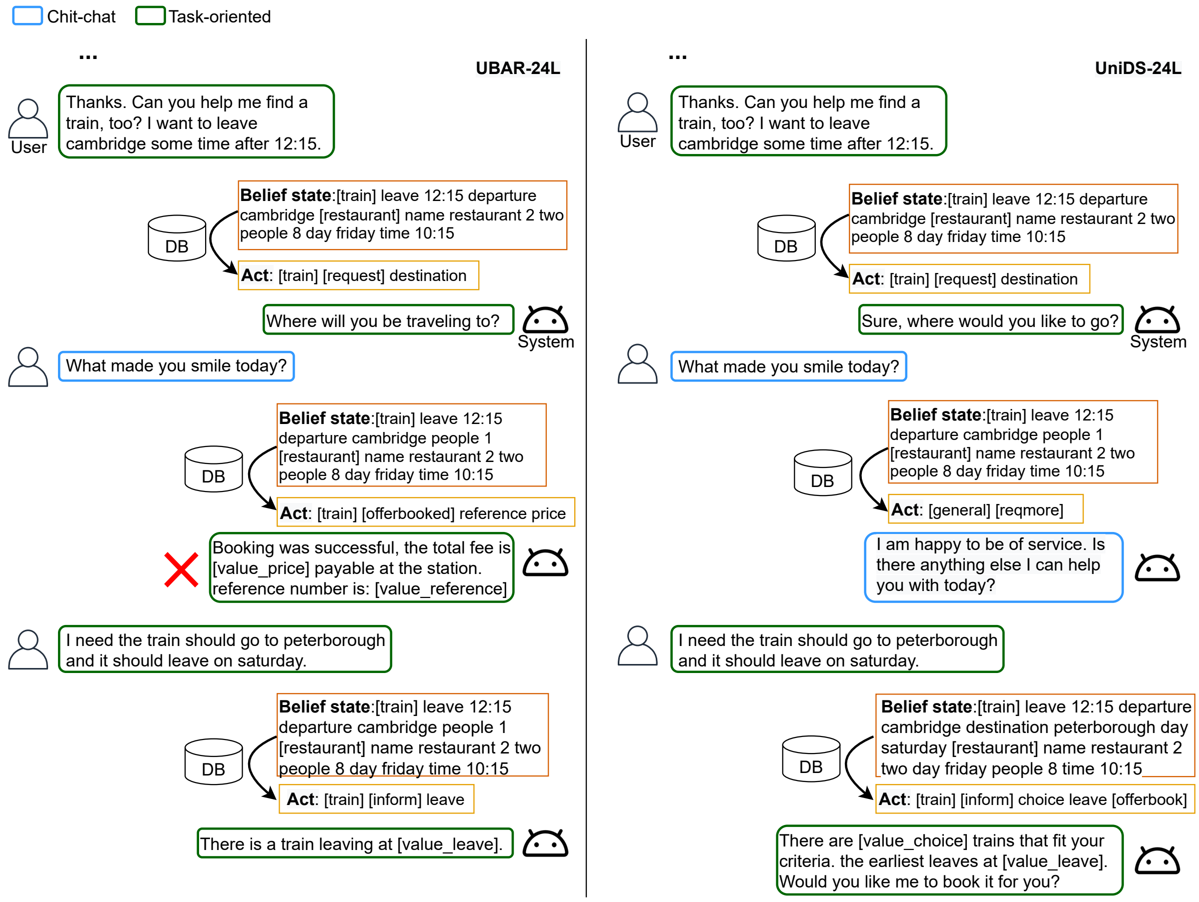
Model	Base	1 turn	2 turns
UBAR-12L	99.18	93.76 (-5.42)	88.14 (-11.04)
UniDS-12L	100.06	96.13 (-3.93)	91.42 (-8.64)
UBAR-24L	99.31	93.08 (-6.23)	88.67 (-10.64)
UniDS-24L	104.12	100.71 (-3.41)	95.68 (-8.44)

Table 9: Combined score over TOD dataset for robustness test by inserting 1 and 2 turns of task-irrelavant utterances. Full results are presented in Appendix.

(ii) When switching from task-oriented dialogues to chit-chat dialogues, the value of Switch-1 is relatively low, this may because our model tends to confirm user intents or give a transitional response rather than switch to chit-chat mode immediately. As the case shown in Table 8, when the user switches from TOD to chit-chat, UniDS gives a chatty response and thanks the user for using its services.

4.6.3 Robustness Study

Many real-world dialogue systems need real-time speech recognition to interact with users, which is easily interfered by background noise from the background environment (e.g. other people and devices). Therefore, we analyze the robustness of UniDS and UBAR by inserting several turns of



Thanks. Can you help me find a train, too? I want to leave cambridge some time after 12:15.

Belief state:[train] leave 12:15 departure cambridge [restaurant] name restaurant 2 two people 8 day friday time 10:15

Act: [train] [request] destination

Where will you be traveling to?

What made you smile today?

Belief state:[train] leave 12:15 departure cambridge people 1 [restaurant] name restaurant 2 two people 8 day friday time 10:15

Act: [train] [offerbooked] reference price

Booking was successful, the total fee is [value_price] payable at the station. reference number is: [value_reference]

I need the train should go to peterborough and it should leave on saturday.

Belief state:[train] leave 12:15 departure cambridge people 1 [restaurant] name restaurant 2 two people 8 day friday time 10:15

Act: [train] [inform] leave

There is a train leaving at [value_leave].

Thanks. Can you help me find a train, too? I want to leave cambridge some time after 12:15.

Belief state:[train] leave 12:15 departure cambridge [restaurant] name restaurant 2 two people 8 day friday time 10:15

Act: [train] [request] destination

Sure, where would you like to go?

What made you smile today?

Belief state:[train] leave 12:15 departure cambridge people 1 [restaurant] name restaurant 2 two people 8 day friday time 10:15

Act: [general] [reqmore]

I am happy to be of service. Is there anything else I can help you with today?

I need the train should go to peterborough and it should leave on saturday.

Belief state:[train] leave 12:15 departure cambridge destination peterborough day saturday [restaurant] name restaurant 2 two day friday people 8 time 10:15

Act: [train] [inform] choice leave [offerbook]

There are [value_choice] trains that fit your criteria. the earliest leaves at [value_leave]. Would you like me to book it for you?

Figure 5: Examples of UBAR-DialoGPT-24L and UniDS-24L when inserting a task-irrelevant utterance in a task-oriented dialogue. UBAR-DialoGPT reserves a train for the user randomly, which makes the task failed because the user intent is incomplete; while UniDS keeps the previous belief state and gives a chatty response. When the user returns to the TOD, UniDS could continue with the task.

irrelevant chit-chat utterances into the TOD, and we evaluate the model performance against such noise.

As observed in Table 9, both UniDS and UBAR drops on the combined score when only one turn of chit-chat dialogue is inserted. However, UniDS drop less than UBAR (about 4 vs. 6 points). Similarly, when two turns of chit-chat are inserted into TOD, UniDS drops about 8 points, and UBAR drops about 11 points on the combined score. These results demonstrate that UniDS has stronger robustness to such task-irrelevant noise than UBAR. We present an interesting case in Figure 5. When giving a task-irrelevant utterance, UBAR-24L reserves a train for the user randomly, which makes the task failed because the user intent is incomplete, while UniDS keeps the previous belief state and gives a chatty response. When the user returns to the TOD, UniDS can continue with the task.

5 Conclusion

This paper proposes a unified dialogue system (UniDS) to jointly handle both chit-chat and task-oriented dialogues in an end-to-end framework. Specifically, we propose a unified dialogue data schema for both chit-chat and task-oriented dialogues, and a two-stage method to train UniDS. To our best knowledge, this is the first study towards an end-to-end unified dialogue system.

Experiments show that UniDS performs comparably with state-of-the-art chit-chat dialogue systems and task-oriented dialogue systems without adding extra parameters to current chit-chat dialogue systems. More importantly, the proposed UniDS achieves good switch ability and shows better robustness than pure task-oriented dialogue systems. Although question answering (QA) is not considered in the proposed UniDS, as an initial attempt, our explorations may inspire future studies towards building a general dialogue system.

6 Ethical Considerations

We notice that some chit-chat utterances generated by the proposed UniDS may be unethical, biased or offensive. Toxic output is one of the main issues of current state-of-the-art dialogue models trained on large naturally-occurring datasets. We look forward to furthering progress in the detection and control of toxic outputs.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020. PLATO-2: towards building an open-domain chatbot via curriculum learning. *CoRR*, abs/2006.16779.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 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.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

A. Joshi, Saket Kale, Satish Chandel, and D. Pal. 2015. Likert scale: Explored and explained. *British Journal of Applied Science and Technology*, 7:396–403.

Lukasz Kaiser, Aidan N. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One model to learn them all. *CoRR*, abs/1706.05137.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119.

Zhaojiang Lin, Andrea Madotto, Yejin Bang, and Pascale Fung. 2021. The adapter-bot: All-in-one controllable conversational model. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 16081–16083.

I. Loshchilov and F. Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

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

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. SOLOIST: few-shot task-oriented dialog with A single pre-trained auto-regressive model. *CoRR*, abs/2005.05298.

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

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 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, EACL 2021, Online, April 19 - 23, 2021*, pages 300–325.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2021. Multi-task pre-training for plug-and-play task-oriented dialogue system. *CoRR*, abs/2109.14739.

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 *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1570–1583.

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

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. UBAR: towards fully end-to-end task-oriented dialog system with GPT-2. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third*

Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 14230–14238.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. 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, ACL 2020, Online, July 5-10, 2020*, pages 270–278.

Tiancheng Zhao, Allen Lu, Kyusong Lee, and Maxine Eskenazi. 2017. Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pages 27–36.

Xinyan Zhao, Feng Xiao, Haoming Zhong, Jun Yao, and Huanhuan Chen. 2020. Condition aware and revise transformer for question answering. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 2377–2387.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. 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 730–739.