Out-of-Distribution Generalization Challenge in Dialog State Tracking

Jiasheng Ye[*] **Yawen Ouyang**[*] **Zhen Wu**[†] **Xinyu Dai**
National Key Laboratory for Novel Software Technology, Nanjing University
Collaborative Innovation Center of Novel Software Technology and Industrialization, China
{yejiasheng, ouyangyw}@smail.nju.edu.cn, {wuz, daixinyu}@nju.edu.cn

Abstract

Dialog State Tracking (DST) is a core component for multi-turn Task-Oriented Dialog (TOD) systems to understand the dialogs. DST models need to generalize to Out-of-Distribution (OOD) utterances due to the open environments dialog systems face. Unfortunately, utterances in TOD are multi-labeled, and most of them appear in specific contexts (i.e., the dialog histories). Both characteristics make them different from the conventional focus of OOD generalization research and remain unexplored. In this paper, we formally define OOD utterances in TOD and evaluate the generalizability of existing competitive DST models on the OOD utterances. Our experimental result shows that the performance of all models drops considerably in dialogs with OOD utterances, indicating an OOD generalization challenge in DST. [3]

1 Introduction

In multi-turn Task-Oriented Dialog (TOD) systems, Dialog State Tracking (DST) is responsible for extracting key information from dialog histories to help the systems make responses [Williams et al., 2016]. Deployed in open environments, it has to deal with a wide variety of user inputs. On the contrary, training data for the system are limited, and sampling bias may exist in the data collection process [Shen et al., 2021]. As a result, the distribution of actual user utterances is typically different from the training data [Williams, 2013]. Therefore, it is essential for DST models to generalize to the Out-of-Distribution (OOD) utterances [Shen et al., 2021].

The utterances in a multi-turn TOD, however, are of a complex data type that cannot be trivially tackled by existing works on OOD generalization. On one hand, the semantics of an utterance can be made up of multiple dialog acts. For example, the user in Dialog MUL2395 of MultiWOZ 2.3 [Han et al., 2021] says *"Yes, I also love Turkish food. Is there something in the center that's expensive? Also, what type of attraction is All Saints Church?"* Here, the speaker informs the preferred cuisine, as well as the location and price range of the restaurant, and requests for the type of an attraction all in one utterance. On the other hand, the utterances in a dialog depend on their dialog histories. For instance, a user is more likely to discuss his or her preferred cuisine under the topic related to restaurants, especially after the system asks about it. Nevertheless, existing works on OOD generalization focus on individual data instances with single labels [Shen et al., 2021]. As a result, they are insufficient to be applied to the OOD generalization problem in DST.

In this paper, we extend the research area of OOD generalization to multi-turn TOD, DST task in particular. In specific, we first formally define OOD utterances in multi-turn TOD. Based on the

[*] Equal contributions

[†] Corresponding author

[3] Our code is available at https://github.com/yegcjs/DST_OOD

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

definition, we propose a procedure to identify OOD utterances according to their labels. We then further evaluate the OOD generalizability of three competitive DST models which rely on three different tracking strategies. The experiment result shows that the performance of all three evaluated models drops considerably in dialogs with OOD utterances, confirming the challenge of the OOD generalization problem in DST task.

2 Background

Dialog State Tracking (DST) aims to summarize the dialog in the form of a dialog state to provide sufficient information for the dialog system to take action [Williams et al., 2016]. Formally, given a dialog up to the T-th turn $(U_0^{(s)}, U_0^{(u)}, \ldots, U_T^{(s)}, U_T^{(u)})$, where $U_t^{(s)}$ and $U_t^{(u)}$ are the system and user utterance in the t-th turn, respectively. Given a set of predefined domain-slot pairs, DST aims to produce slot values for each domain-slot pair according to the dialog. For example, given a set of domain-slot pairs $\{$(*restaurant*, *food*), (*restaurant*, *price*)$\}$ and a dialog about a user finding a moderately priced Chinese restaurant, DST outputs *Chinese* for (*restaurant*, *food*) and *moderate* for (*restaurant*, *price*).

Generally, the semantics in dialogs can be represented as dialog acts, which are triplets of domain, intent, and slot [Traum and Hinkelman, 1992, Su, 2018]. The semantics of an utterance usually composes of multiple (*domain*, *intent*, *slot*) triplets, and therefore represented as a set of triplets. For the sample utterance presented in Paragraph 2 of Section 1, its semantics is represented as $\{$(*attraction*, *inform*, *name*), (*restaurant*, *inform*, *food*), (*restaurant*, *inform*, *price*), (*restaurant*, *inform*, *area*), (*attraction*, *request*, *type*)$\}$.

3 Definition and Identification of OOD Utterances in Multi-turn TOD

Conceptually, a data instance is out of distribution if it is drawn from a distribution different from the training data [Ye et al., 2022, Shen et al., 2021]. However, the actual distribution of training data is unknown. Therefore, deciding whether an utterance is out of distribution according to this conceptual definition is intractable. This limitation motivates us to devise an operational definition based on the semantic labels of the utterances, which is described in the next paragraph.

Recall that the semantics of an utterance in multi-turn TOD composes of multiple dialog acts, and the occurrence of an utterance depends on its context. Hereby, we consider two cases where an utterance in multi-turn TOD is OOD. For the first case, the semantics of the utterance itself is unseen in the training data. As for the other case, the given utterance semantics is unseen after its context, even though the utterance semantics itself exists in training data. In this paper, we name the two cases as **non-contextual OOD utterances** and **contextual OOD utterances**, respectively.

The operational definition naturally leads to the procedure of identifying whether an utterance is OOD and its type. The pseudocode of the full procedure is in Appendix B. Let $S(U)$ represent the semantics (i.e., the set of dialog acts) of the utterance U. The procedure can be divided into two parts: preprocess and identification.

1. **Preprocess** For each dialog in the training set with the semantics of the utterances as $(S(U_0^{(s)}), S(U_0^{(u)}), \ldots, S(U_T^{(s)}), S(U_T^{(u)}))$, and any t in $\{0, 1, \ldots, T\}$, we add $S(U_t^{(u)})$ to the training response set of corresponding dialog history $R(S(U_0^{(s)}), S(U_0^{(u)}), .., S(U_t^{(s)}))$, where $(S(U_0^{(s)}), S(U_0^{(u)}), .., S(U_t^{(s)}))$ is semantics sequence of the dialog history up to t-th turn.
2. **Identification** Given an user utterance $U_t^{(u)}$ and its dialog history $(U_0^{(s)}, U_0^{(u)}, \ldots, U_t^{(s)})$, we can decide its type with following steps. First, if the semantics of the user utterance belongs to the training response set of its dialog history (i.e., $S(U_t^{(u)}) \in R(S(U_0^{(s)}), S(U_0^{(u)}), \ldots, S(U_t^{(s)}))$), the user utterance is in-distribution(ID). Otherwise, we further check whether it is a contextual or non-contextual OOD utterance by whether there exists a training response set that contains the semantics of the user utterance. Formally, an OOD utterance $U_t^{(u)}$ is a contextual OOD if $S(U_t^{(u)}) \in \bigcup R(\cdot)$. Otherwise, it is a non-contextual OOD.

A noteworthy corollary from the definition is that any utterance is OOD if it comes after a dialog history with OOD utterances since the corresponding training response set of the dialog history is empty.

4 Experiments

With the definition of OOD utterances in multi-turn TOD, we conduct experiments to evaluate the OOD generalizability of DST models. We particularly focus on the following research questions. **RQ1**: Do existing competitive DST models generalize well to OOD utterances? **RQ2**: Is the generalization difficulty of the two types of OOD utterances different? **RQ3**: After a turn with an OOD utterance, DST models have to handle dialogs with OOD utterances in dialog history. Do OOD utterances affect the performance of DST models in their future turns?

For the first question, we compare the performance of the DST models on dialogs whose latest utterances are OOD and ID, respectively. To answer the second question, we categorize the dialogs to be tested based on the types of user utterances in the latest turns. And to the third question, we further categorize the dialogs into two types according to whether there exist OOD utterances in the dialog history. In all, we divide all the dialogs into five categories (Table 1).

Before presenting our findings, we first introduce the tested DST models (Section 4.1), datasets (Section 4.2), and evaluation metrics (Section 4.3).

4.1 DST Models

There are three typical ways for recently proposed DST models to obtain slot values [Balaraman et al., 2021]. One can generate the slot values directly, copy the value from utterances or other known information, or apply a mixture of both strategies. We select three recent competitive models in our experiment to represent three kinds of strategies, which are **SimpleTOD** [Hosseini-Asl et al., 2020], **TripPy** [Heck et al., 2020] and **TRADE** [Wu et al., 2019], respectively.

4.2 Datasets

MultiWOZ 2.3 [Han et al., 2021] is a revised version of MultiWOZ 2.0 [Budzianowski et al., 2018]. It corrects the annotation mistakes on domains, intents, and slots of the utterances. We observe that there already exist many dialogs with OOD utterances in the original test set of MultiWOZ 2.3. So we first categorize the dialogs in the original MultiWOZ 2.3 test set and evaluate the performance of DST models on them.

MultiWOZ OOD We construct another test set by modifying the original MultiWOZ 2.3 test set because the MultiWOZ 2.3 test set is insufficient to support further analyses. When comparing different types of utterances (RQ1 and RQ2), we expect consistent dialog histories in different categories. When comparing dialog histories with and without OOD utterances (RQ3), we expect consistent latest user utterances. As the split of MultiWOZ 2.3 test set does not satisfy the two above properties, we instead modify the original test set to construct a new test set. The details of the construction process can be found in Appendix C.

4.3 Metrics

We evaluate the performance of DST models with two metrics: Joint Goal Accuracy (JGA) and Turn-level State Accuracy (TSA).

Joint Goal Accuracy [Henderson et al., 2014] We follow the convention in the dialog state tracking task and apply JGA as a metric in our evaluation. JGA measures whether the predicted state is identical to the gold ones. It strictly considers a prediction correct if and only if values in all the slots are correctly predicted.

Turn-level State Accuracy Rastogi et al. [2020] and LI et al. [2020] propose to compute the state accuracy on turn level besides the traditional JGA. During computing TSA, only slots with different values from the previous turn are involved. In this way, TSA ignores the errors accumulated from dialog histories and accurately measures the performance on the latest turn [Rastogi et al., 2020].

Table 1: TSA/JGA of different DST models on different types of dialogs

Dataset	DST Model		Non-contextual OOD	Contextual OOD	ID
MultiWOZ 2.3	SimpleTOD	History w/o OOD	62.50/65.63	76.40/71.68	88.41/87.84
		History w OOD	48.63/39.73	71.49/42.58	-
	TripPy	History w/o OOD	81.25/84.38	83.22/78.50	91.64/91.09
		History w/ OOD	56.16/36.99	80.28/49.51	-
	TRADE	History w/o OOD	62.50/65.63	71.12/67.44	87.09/86.05
		History w/ OOD	42.47/26.71	63.71/33.88	-
MultiWOZ OOD	SimpleTOD	History w/o OOD	17.21/15.58	47.09/42.79	86.28/84.88
		History w/ OOD	11.16/5.58	33.96/15.35	-
	TripPy	History w/o OOD	46.28/42.33	67.33/62.68	91.63/89.77
		History w/ OOD	38.60/17.91	57.21/31.74	-
	TRADE	History w/o OOD	33.02/28.37	56.51/51.28	85.12/81.62
		History w/ OOD	18.88/8.60	37.44/17.55	-

4.4 Results and Discussions

Full experimental results are on Table 1.

4.4.1 DST Models Generalize to OOD Utterances Poorly (RQ1)

To study whether the selected competitive DST models generalize well on dialogs with OOD utterances, we compare the model performance on dialogs with and without OOD utterances. In particular, we focus on dialogs without OOD utterances in their history. By doing so, we avoid the potential influence of OOD utterances in dialog history.

As demonstrated in Table 1, the performance of all three models drops considerably on dialogs with OOD utterances compared to those without. Moreover, we find that in dialogs without OOD utterances, all the evaluated models achieve JGA over 86% in the original MultiWOZ 2.3 test set and 81% in MultiWOZ OOD. Nevertheless, all the models perform disappointingly in dialogs with OOD utterances. This suggests that improving the generalizability to OOD utterances is critical for the existing DST models to further improve their overall performance.

4.4.2 Non-contextual OOD Utterances are More Difficult to Generalize (RQ2)

To investigate the relative OOD generalization difficulty, we compare the model performance on dialogs whose latest utterances are of different types. Similar to the discussion on RQ1 (4.4.1), we compare dialogs without OOD utterances in history for fairness.

Both in terms of TSA and JGA, all three models perform worse in non-contextual OOD utterances. This is not surprising as utterances with the same semantics as a contextual OOD utterance are seen in the training set. As a result, it is easier to generalize.

4.4.3 OOD Utterances Even Hurt DST Performance in Their Future Turns (RQ3)

OOD utterances sometimes exist in dialog histories. Whether DST performance suffers from OOD utterances in previous turns remains unclear. To answer this question, we compare DST performance in dialogs with and without OOD utterances in history for each type of current user utterance. In this setting, comparison on JGA is extremely unfair because it accumulates a hugely different amount of errors from dialog histories with and without OOD utterances. Therefore, we only compare TSA.

As shown in Table 1, for all types of latest user utterances and all of the three evaluated models, TSA is lower in dialogs with OOD utterances in history. This suggests that OOD utterances even hurt the DST performance in their future turns.

5 Conclusion and Future Work

In this paper, we extend the scope of OOD generalization to multi-turn TOD, DST task in particular. We first formally define OOD utterances in multi-turn TOD, and experiment to evaluate the generalizability of three competitive DST models to them. The results show that the performance of all the models drops considerably in dialogs with OOD utterances, especially those non-contextual ones. Moreover, we find that OOD utterances affect DST performance not only in their turn but also in their future turns.

We verified the OOD generalization challenge in DST task, while how to improve the OOD generalizability of DST methods remains unsolved. Moreover, it would be interesting to further adapt our definition of OOD utterances in multi-turn TOD to more tasks besides DST. We expect that this work can draw the community's attention to the OOD generalization problem in multi-turn TOD.

Acknoledgement

We would like to thank the anonymous reviewers for their comments. This work is supported by NSFC Projects (No. 62206126, 61936012 and 61976114).

References

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

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

Jason D Williams. Multi-domain learning and generalization in dialog state tracking. In *Proceedings of the SIGDIAL 2013 Conference*, pages 433–441, 2013.

Ting Han, Ximing Liu, Ryuichi Takanabu, Yixin Lian, Chongxuan Huang, Dazhen Wan, Wei Peng, and Minlie Huang. Multiwoz 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 206–218. Springer, 2021.

David R Traum and Elizabeth A Hinkelman. Conversation acts in task-oriented spoken dialogue. *Computational intelligence*, 8(3):575–599, 1992.

Pei-Hao Su. *Reinforcement Learning and Reward Estimation for Dialogue Policy Optimisation*. PhD thesis, University of Cambridge, 2018.

Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7947–7958, 2022.

Vevake Balaraman, Seyedmostafa Sheikhalishahi, and Bernardo Magnini. Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 239–251, 2021.

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

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gašic. Trippy: A triple copy strategy for value independent neural dialog state tracking. In *21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 35, 2020.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 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, 2019.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 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, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL `https://aclanthology.org/D18-1547`.

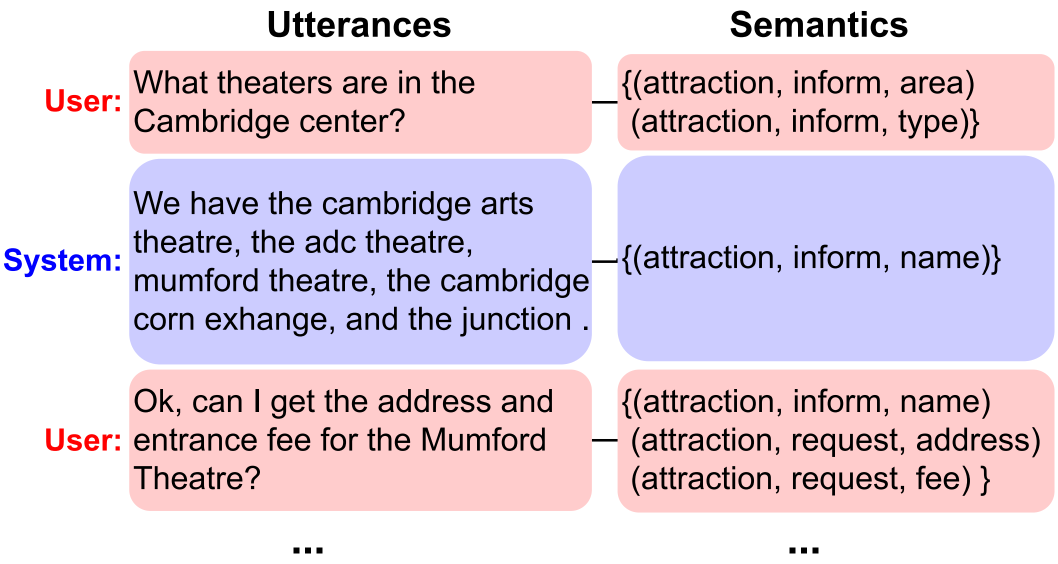
Matthew Henderson, Blaise Thomson, and Jason D Williams. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272, 2014.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Schema-guided dialogue state tracking task at dstc8. *arXiv preprint arXiv:2002.01359*, 2020.

SHIYANG LI, Semih Yavuz, Kazuma Hashimoto, Jia Li, Tong Niu, Nazneen Rajani, Xifeng Yan, Yingbo Zhou, and Caiming Xiong. Coco: Controllable counterfactuals for evaluating dialogue state trackers. In *International Conference on Learning Representations*, 2020.

A An Example of Multi-turn TOD Dialogs and Semantics Labels of Utterances in it

An example of dialogs in multi-turn TOD is in figure 1. The user and the system take turns to utter and each utterance is labeled with its dialog acts.



What theaters are in the Cambridge center? — {(attraction, inform, area) (attraction, inform, type)}

We have the cambridge arts theatre, the adc theatre, mumford theatre, the cambridge corn exhange, and the junction . — {(attraction, inform, name)}

Ok, can I get the address and entrance fee for the Mumford Theatre? — {(attraction, inform, name) (attraction, request, address) (attraction, request, fee) }

Figure 1: An example of dialog TOD with dialog acts labeled. This example is extracted from PMUL2292 in the training set of MultiWOZ 2.3. Each utterance (left) is labeled with a set of dialog acts (right).

B Pseudocode for Identifying OOD Utterances with Semantics Labels

The full procedure include two parts: preprocess (Algorithm 1) and identification (Algorithm 2. We first set up training response sets during preprocessing. And then we apply the identification procedure to decide whether an utterance is OOD by looking up the training response sets.

Algorithm 1 Preprocess

```
Input: dialogs in the training set $\mathcal{D}_{train}$ with dialog acts labels
Output: the set of all training response sets $\{R(\cdot)\}$
Initialize all training response sets $R(\cdot)$ as $\emptyset$.
for each dialog $(U_0^{(s)}, U_0^{(u)}, \ldots, U_T^{(s)}, U_T^{(u)}) \in \mathcal{D}_{train}$ do
    for $t \in \{0, 1, \ldots, T\}$ do
        % Add the semantic label of current user utterance (i.e. $S(U_t^{(u)})$) into the training
        % response set of corresponding dialog history.
        $R(S(U_0^{(s)}), S(U_0^{(u)}), \ldots, S(U_t^{(s)})) \leftarrow R(S(U_0^{(s)}), S(U_0^{(u)}), \ldots, S(U_t^{(s)})) \cup S(U_t^{(u)})$
return $\{R(\cdot)\}$
```

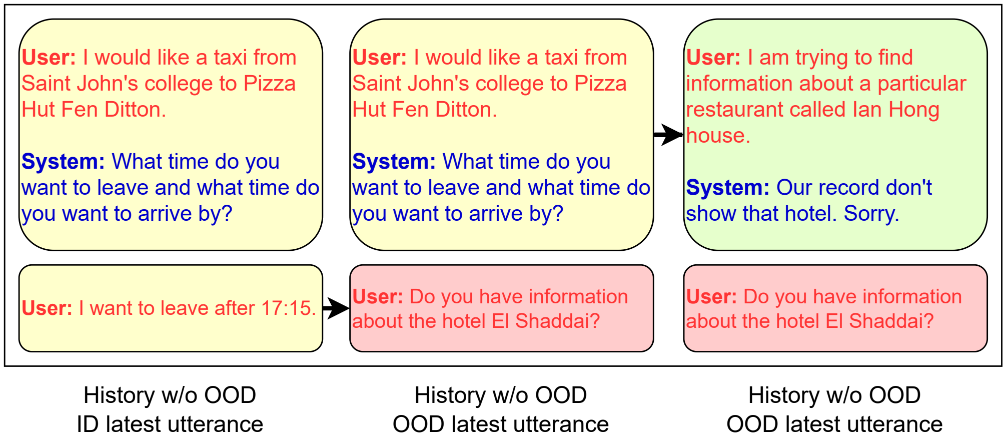
Algorithm 2 Identification

```
Input: the set of all the training response sets $\{R(\cdot)\}$,
        a segment of dialog history $(U_0^{(s)}), U_0^{(u)}, \ldots, U_t^{(s)}, U_t^{(u)})$ with labels
Output: the type of latest user utterance $U_t^{(u)}$
if $S(U_t^{(u)}) \in R(S(U_0^{(s)}), S(U_0^{(u)}), \ldots, S(U_t^{(s)}))$ then
    return ID
else
    for each training response set $R \in \{R(\cdot)\}$ do
        if $S(U_t^{(u)}) \in R$ then
            return Contextual OOD
    return Non-contextual OOD
```

C Construction of MultiWOZ OOD

To tackle the limitation of the MultiWOZ 2.3 test set (Section 4.2), we construct a new test set named MultiWOZ OOD by following steps. First, we pick out all the dialogs without OOD utterances. For each of the picked-out dialogs, we replace the last user utterance with another user utterance from the MultiWOZ 2.3 test set to make the dialog containing a dialog history without OOD utterances and an OOD latest utterance (for the analyses in Section 4.4.1 and Section 4.4.2). Finally, for each dialog constructed in the previous step, we replace its dialog history with one that contains OOD utterances (for the analyses in Section 4.4.3). In particular, the new dialog history is also from the original MultiWOZ 2.3 test set and has the same length as the original one. This step produces dialogs made up of dialog history with OOD utterances and an OOD user utterance in the latest turn. Figure 2 demonstrates an example of the data construction process.

Note that MultiWOZ OOD only serves as a test set. Before testing, training and validation of the DST models are conducted on the original split of MultiWOZ 2.3 dataset.



User: I would like a taxi from Saint John's college to Pizza Hut Fen Ditton.

System: What time do you want to leave and what time do you want to arrive by?

User: I want to leave after 17:15.

User: I would like a taxi from Saint John's college to Pizza Hut Fen Ditton.

System: What time do you want to leave and what time do you want to arrive by?

User: Do you have information about the hotel El Shaddai?

User: I am trying to find information about a particular restaurant called Ian Hong house.

System: Our record don't show that hotel. Sorry.

User: Do you have information about the hotel El Shaddai?

Figure 2: An example of the data construction process for MultiWOZ OOD. The process starts with a dialog whose history has no OOD utterances and the latest user utterance is also ID. Then the latest user utterance is replaced with an utterance that is OOD under the history. This results in a dialog whose history has no OOD utterances but the latest user utterance is OOD. In the final step, the dialog history in the previously generated dialog is further replaced with one that has OOD utterances, which results in a dialog whose history has OOD utterances and the latest user utterance is OOD as well.