MultiWOZ 2.4: A Multi-Domain Task-Oriented Dialogue Dataset with Essential Annotation Corrections to Improve State Tracking Evaluation

Anonymous ACL submission

Abstract

The MultiWOZ 2.0 dataset was released in 2018. It consists of more than 10,000 task-oriented dialogues spanning 7 domains, and has greatly stimulated the research of task-oriented dialogue systems. However, there is substantial noise in the state annotations, which hinders a proper evaluation of dialogue state tracking models. To tackle this issue, massive efforts have been devoted to correcting the annotations, resulting in three improved versions of this dataset (i.e., MultiWOZ 2.1-2.3). Even so, there are still lots of incorrect and inconsistent annotations. This work introduces MultiWOZ 2.4, in which we refine all annotations in the validation set and test set on top of MultiWOZ 2.1. The annotations in the training set remain unchanged to encourage robust and noise-resilient model training. We further benchmark nine state-of-the-art dialogue state tracking models. All these models achieve much higher performance on MultiWOZ 2.4 than on MultiWOZ 2.1.

1 Introduction

Task-oriented dialogue systems serve as personal assistants. They play an important role in helping users accomplish numerous tasks such as hotel booking, restaurant reservation, and map navigation. An essential module in task-oriented dialogue systems is the dialogue state tracker, which aims to keep track of users' intentions at each turn of a conversation (Mrkšić et al., 2017). The state information is then leveraged to determine the next system action and generate the next system response.

In recent years, tremendous advances have been made in the research of task-oriented dialogue systems, attributed to a number of publicly available dialogue datasets like DSTC2 (Henderson et al., 2014), FRAMES (El Asri et al., 2017), WOZ (Wen et al., 2017), M2M (Shah et al., 2018), MultiWOZ 2.0 (Budzianowski et al., 2018), SGD (Rastogi et al., 2020), CrossWOZ (Zhu et al., 2020), RiSAWOZ (Quan et al., 2020), and TreeDST (Cheng et al., 2020). Among them, MultiWOZ 2.0 is the first large-scale dataset spanning multiple domains and thus has attracted the most attention. Specifically, MultiWOZ 2.0 contains about 10,000 dialogues spanning 7 domains including *attraction*, *bus*, *hospital*, *hotel*, *restaurant*, *taxi*, and *train*.

However, substantial noise has been found in the dialogue state annotations of MultiWOZ 2.0 (Eric et al., 2020). To remedy this issue, Eric et al. (2020) fixed 32% of dialogue state annotations across 40% of the dialogue turns, resulting in an improved version MultiWOZ 2.1. Despite the significant improvement on annotation quality, MultiWOZ 2.1 still severely suffers from incorrect and inconsistent annotations (Zhang et al., 2020; Hosseini-Asl et al., 2020). The state-of-the-art joint goal accuracy (Zhong et al., 2018) for dialogue state tracking on MultiWOZ 2.1 is merely around 60% (Li et al., 2021). Even worse, the noise in the validation set and test set makes it relatively challenging to assess model performance properly and adequately. To reduce the impact of noise, different preprocessing strategies have been utilized by existing models. For example, TRADE (Wu et al., 2019) fixes some general annotation errors. SimpleTOD (Hosseini-Asl et al., 2020) cleans partial noisy annotations in the test set. TripPy (Heck et al., 2020) constructs a label map to handle value variants. These preprocessing strategies, albeit helpful, lead to an unfair performance comparison. In view of this, we argue that it is valuable to further refine the annotations of MultiWOZ 2.1.

As a matter of fact, massive efforts have already been made to further improve the annotation quality of MultiWOZ 2.1, resulting in MultiWOZ 2.2 (Zang et al., 2020) and MultiWOZ 2.3 (Han et al., 2020b). Nonetheless, they both have some limitations. More concretely, MultiWOZ 2.2 allows the presence of multiple values in the dialogue state. But it doesn't cover all the value variants.

Domain	Slot
attraction	area, name, type
bus	arriveby, book people, day, departure, destination, leaveat
hospital	department
hotel	area, book day, book people, book stay, internet, name, parking, pricerange, stars, type
restaurant	area, book day, book people, book time, food, name, pricerange
taxi	arriveby, departure, destination, leaveat
train	arriveby, book people, day, departure, destination, leaveat

Table 1: The predefined slots within each domain.

This incompleteness brings about serious inconsistencies. MultiWOZ 2.3 focuses on dialogue act annotations. The noise on dialogue state annotations has not been fully resolved.

In this work, we introduce MultiWOZ 2.4, an updated version on top of MultiWOZ 2.1, to improve dialogue state tracking evaluation. Specifically, we identify and fix all the incorrect and inconsistent annotations in the validation set and test set. This refinement results in changes to the state annotations of more than 40% of turns over 65% of dialogues. Since our main purpose is to improve the correctness and fairness of model evaluation, the annotations in the training set remain unchanged. Even so, the empirical study shows that much better performance can be achieved on MultiWOZ 2.4 than on all the previous versions (i.e., MultiWOZ 2.0-2.3). Furthermore, a noisy training set motivates us to design robust and noise-resilient training mechanisms, e.g., data augmentation (Summerville et al., 2020) and noisy label learning (Han et al., 2020a). Considering that collecting noise-free large multi-domain dialogue datasets is costly and labor-intensive, we believe that training robust dialogue state tracking models from noisy training data will be of great interest to both industry and academia.

2 Annotation Refinement

In the MultiWOZ 2.0 dataset, the dialogue state is represented as a set of predefined slots (refer to Table 1 for all the slots of each domain) and their corresponding values. The slot values are extracted from the dialogue context. For example, *attraction-area=centre* means that the slot is *attraction-area* and its value is *centre*. Since a dialogue may involve multiple domains and each domain also has multiple slots, it is impractical to ensure that the state annotations obtained via a crowdsourcing process are consistent and noise-free. Even though MultiWOZ 2.1 has tried to correct the annotation errors, the refining process was based on crowdsourcing as well. Therefore, MultiWOZ 2.1 still suffers from incorrect and inconsistent annotations.

2.1 Annotation Error Types

We identify and fix ten types of annotation errors (including inconsistent annotations) in the validation set and test set of MultiWOZ 2.1. Figure 1 shows an example for each error type.

- **Context Mismatch:** The slot has been annotated, however, its value is inconsistent with the one mentioned in the dialogue context.
- **Mis-Annotation:** The slot is not annotated, even though its value has been mentioned.
- **Not Mentioned:** The slot has been annotated, however, its value has not been mentioned in the dialogue context at all.
- **Multiple Values:** The slot should have multiple values, but not all values are included.
- **Typo:** The slot has been correctly annotated, except that its value includes a typo.
- **Implicit Time Processing:** This relates to the slots that take time as the value. Instead of copying the time specified in the dialogue context, the value has been implicitly processed (e.g., adding or subtracting 15 min).
- **Slot Mismatch:** The extracted value is correct, but it has been matched to a wrong slot.
- **Incomplete Value:** The slot value is a substring or an abbreviation of its full shape (e.g., “Thurs" vs. “Thursday").
- **Delayed Annotation:** The slot has been annotated several turns later than its value first mentioned in the dialogue context.
- **Unnecessary Annotation:** These unnecessary annotations are not incorrect but they exacerbate inconsistencies as different annotators have different opinions on whether to annotate these slots or not. In general, the values of these slots are mentioned by the system

Error Type	Conversation	MultiWOZ 2.1	MultiWOZ 2.4
Context Mismatch	**Usr:** Hello, I would like to book a taxi from restaurant 2 two to the museum of classical archaeology.	taxi-destination=museum of archaelogy and anthropology	taxi-destination=museum of classical archaeology
Mis-Annotation	**Usr:** I need a place to dine in the centre of town.	rest.-area=none	rest.-area=centre
Not Mentioned	**Usr:** I am planning a trip in Cambridge.	hotel-internet=dontcare	hotel-internet=none
Multiple Values	**Usr:** Something classy nearby for dinner, preferably Italian or Indian cuisine?	rest.-food=Indian	rest.-food=Indian|Italian
Typo	**Usr:** I am looking for a restaurant that serves Portuguese food.	rest.-food=Portugese	rest.-food=Portuguese
Implicit Time Processing	**Usr:** I need a train leaving after 10:00.	train-leaveat=10:15	train-leaveat=10:00
Slot Mismatch	**Usr:** Can you please help me find a place to go in town in the same area as the hotel? Preferably a college.	attraction-name=college attraction-type=none	attraction-name=none attraction-type=college
Incomplete Value	**Sys:** I recommend Charlie Chan. Would you like a table? **Usr:** Yes. Monday, 8 people, 10:30.	rest.-name=Charlie	rest.-name=Charlie Chan
Delayed Annotation	**Usr:** Please recommend one and book it for 6 people. **Sys:** I would recommend express by holiday inn Cambridge. From what day should I book? **Usr:** Starting Saturday. I need 5 nights for 6 people.	hotel-book people=none hotel-book people=6	hotel-book people=6 hotel-book people=6
Unnecessary Annotation	**Usr:** I am looking for a museum. **Sys:** The Broughton house gallery is a museum in the centre. **Usr:** That sounds good. Could I get their phone number?	attraction-area=centre	attraction-area=none

Figure 1: Examples of each error type. For each example, only the problematic slots are presented. "rest." is short for restaurant. Note that in the state annotations of the MultiWOZ dataset, a slot is represented as the concatenation of the domain name and the slot name to include the domain information.

to respond to previous user requests or provide supplementary information. We found that in most dialogues, these slots are not annotated. Hence, we remove these annotations. However, the `name`-related slots are an exception. If the user requests more information (e.g., *address* and *postcode*) about the recommended "name", the slots will be annotated.

2.2 Annotation Refinement Procedure

In the validation set and test set of MutliWOZ 2.1, there are 2,000 dialogues with more than 14,000 dialogue turns. And there are 5 domains with a total of 30 slots (the *bus* domain and *hospital* domain only occur in the training set). To guarantee that the refined annotations are as correct and consistent as possible, we decided to rectify the annotations by ourselves rather than crowd-workers. However, if we check the annotations of all 30 slots at each turn, the workload is too heavy. To ease the burden, we instead only checked the annotations of turn-active slots. A slot being turn-active indicates that its value is determined by the dialogue context of current turn and is not inherited from previous turns. The average number of turn-active slots in the original annotations and in the refined annotations is 1.16 and 1.18, respectively. The full dialogue state is then obtained by accumulating all turn-active states from the first turn to current turn.

We also observed that some slot values are mentioned in different forms, such as "concert hall" vs. "concerthall" and "guest house" vs. "guest houses". The `name`-related slot values may have a word `the` at the beginning, e.g., "Peking restaurant" vs. "the Peking restaurant". We normalized these variants by selecting the one with the most frequency. In addition, all `time`-related slot values have been updated to the 24:00 format. We performed the above refining process twice to reduce mistakes and it took us one month to finish this task.

2.3 Statistics on Refined Annotations

Table 2 shows the count and percentage of slot values changed in MultiWOZ 2.4 compared with MultiWOZ 2.1. Note that `none` and `dontcare` are regarded as two special values. As can be seen, most slot values remain unchanged. This is because a dialogue only has a few active slots and all the other slots always take the value `none`. Table 3 further reports the ratio of refined slots, turns and dialogues. Here, the ratio of refined slots is computed on the basis of refined turns. It is shown that the corrected states relate to more than 40% of turns over 65% of dialogues. On average, the annotations of 1.53 ($30 \times 5.10\%$) slots at each refined turn have been rectified. We then report the value vocabulary size (i.e., the number of candidate values) of each slot and its value change ratio in

Refinement Type	Count	Ratio(%)
no change	433,064	97.92
none→value	3,230	0.73
valueA/dontcare→valueB	1,506	0.34
value/dontcare→none	2,846	0.64
none/value→dontcare	1,614	0.36

Table 2: The count and ratio of slot values changed in MultiWOZ 2.4 compared with MultiWOZ 2.1. "*none*" and "*dontcare*" are regarded as two special values.

Dataset	Slot(%)	Turn(%)	Dialogue(%)
val	5.03	42.32	67.10
test	5.17	39.29	63.56
total	5.10	40.81	65.33

Table 3: The ratio of refined states, turns and dialogues.

Table 4. For some slots, the value vocabulary size decreases due to value normalization and error correction. For some slots, the value vocabulary size increases mainly because a few labels that contain multiple values have been additionally introduced. Table 4 also indicates that the value change ratio of the name-related slots is the highest. Since these slots usually have "longer" values, the annotators are more likely to make incomplete and inconsistent annotations.

3 Benchmark Evaluation

In this part, we present some benchmark results.

3.1 Benchmark Models

In recent years, many neural dialogue state tracking models have been proposed based on the MultiWOZ dataset. These models can be roughly divided into two categories: predefined ontology-based methods and open vocabulary-based methods. The ontology-based methods perform classification by scoring all possible slot-value pairs in the ontology and selecting the value with the highest score as the prediction. By contrast, the open vocabulary-based methods directly generate or extract slot values from the dialogue context. We benchmark the performance of our refined dataset on both types of methods, including SUMBT (Lee et al., 2019)[1], CHAN (Shan et al., 2020)[2], STAR (Ye et al., 2021)[3], TRADE (Wu et al., 2019)[4], PIN (Chen et al., 2020)[5], SOM-DST (Kim et al., 2020)[6], SimpleTOD (Hosseini-Asl et al., 2020)[7], SAVN (Wang et al., 2020)[8], and TripPy (Heck et al., 2020)[9]. Among them, the first three are ontology-based

Slot	2.1	2.4	Val(%)	Test(%)
attraction-area	7	8	1.97	1.93
attraction-name	106	92	5.17	5.08
attraction-type	17	23	4.62	3.77
hotel-area	7	8	3.92	3.99
hotel-book day	8	8	0.33	0.52
hotel-book people	9	9	0.68	0.53
hotel-book stay	6	7	0.42	0.42
hotel-internet	5	4	2.32	2.24
hotel-name	48	46	6.28	3.95
hotel-parking	5	4	2.54	2.35
hotel-pricerange	6	6	1.76	2.06
hotel-stars	8	10	1.52	1.44
hotel-type	5	4	5.06	4.78
rest.-area	7	8	2.18	2.38
rest.-book day	8	11	0.35	0.27
rest.-book people	9	9	0.37	0.45
rest.-book time	59	62	0.56	0.46
rest.-food	89	93	2.58	2.28
rest.-name	135	121	7.61	5.40
rest.-pricerange	5	7	1.51	2.05
taxi-arriveby	62	61	0.41	0.56
taxi-departure	177	172	0.80	0.86
taxi-destination	185	181	1.13	0.60
taxi-leaveat	92	89	0.84	0.45
train-arriveby	109	73	1.40	2.86
train-book people	11	12	1.22	1.76
train-day	8	9	0.31	0.24
train-departure	19	15	0.71	1.10
train-destination	20	17	0.71	1.00
train-leaveat	128	96	4.64	5.12

Table 4: The slot value vocabulary size counted on the validation set and test set of MultiWOZ 2.1 and MultiWOZ 2.4, respectively, and the slot-specific value change ratio. "rest." is the abbreviation of restaurant.

[1] https://github.com/SKTBrain/SUMBT
[2] https://github.com/smartyfh/CHAN-DST
[3] https://github.com/smartyfh/DST-STAR
[4] https://github.com/jasonwu0731/trade-dst
[5] https://github.com/BDBC-KG-NLP/PIN_EMNLP2020
[6] https://github.com/clovaai/som-dst
[7] https://github.com/salesforce/simpletod
[8] https://github.com/wyxlzsq/savn
[9] https://gitlab.cs.uni-duesseldorf.de/general/dsml/trippy-public

	Model	Joint Goal Accuracy (%)			Slot Accuracy (%)	
		MWZ 2.1 Test	MWZ 2.4 Test	MWZ 2.4 Val	MWZ 2.1 Test	MWZ 2.4 Test
predefined ontology	SUMBT	49.01	61.86 (+12.85)	62.31	96.76	97.90
	CHAN	53.38	68.25 (+14.87)	68.23	97.39	98.52
	STAR	**56.36**	**73.62** (+17.26)	74.59	**97.59**	**98.85**
open vocabulary	TRADE	45.60	55.05 (+9.45)	57.01	96.55	97.62
	PIN	48.40	58.92 (+10.52)	60.37	97.02	98.02
	SOM-DST	51.24	**66.78** (+15.54)	68.77	97.15	**98.38**
	SimpleTOD	51.75	57.18 (+5.43)	55.02	96.78	96.97
	SAVN	54.86	60.55 (+5.69)	61.91	**97.55**	98.05
	TripPy	**55.18**	59.62 (+4.44)	60.06	97.48	97.94

Table 5: Joint goal accuracy and slot accuracy of different models on MultiWOZ 2.1 and MultiWOZ 2.4.

Dataset	SUMBT (%)	TRADE (%)
MultiWOZ 2.0	48.81	48.62
MultiWOZ 2.1	49.01	45.60
MultiWOZ 2.2	49.70	46.60
MultiWOZ 2.3	52.90	49.20
MultiWOZ 2.3-cof	54.60	49.90
MultiWOZ 2.4	**61.86**	**55.05**

Table 6: Comparison of test set joint goal accuracy on different versions of the MultiWOZ dataset.

Domain	SOM-DST (%)		STAR (%)	
	2.1	2.4	2.1	2.4
attraction	69.83	83.22	70.95	84.45
hotel	49.53	64.52	52.99	69.10
restaurant	65.72	77.67	69.17	84.20
taxi	59.96	54.76	66.67	73.63
train	70.36	82.73	75.10	90.36

Table 7: Comparison of domain-specific test set joint goal accuracy on MultiWOZ 2.1 and MultiWOZ 2.4.

approaches. The rest are open vocabulary-based methods. For all these models, we employ the default hyperparameter settings to retrain them on MultiWOZ 2.4.

3.2 Results Analysis

We adopt the joint goal accuracy (Zhong et al., 2018) and slot accuracy as the evaluation metrics. The joint goal accuracy is defined as the ratio of dialogue turns in which each slot value has been correctly predicted. The slot accuracy is defined as the average accuracy of all slots. The detailed results are presented in Table 5. As can be observed, all models achieve much higher performance on MultiWOZ 2.4. The ontology-based models demonstrate the highest performance promotion, mainly benefiting from the improved ontology. SAVN and TripPy show the least performance increase, because they have already utilized some value normalization techniques to tackle label variants in MultiWOZ 2.1. SimpleTOD also shows the least performance improvement. The reason may be that SimpleTOD generates state values directly while other methods such as TRADE leverage the copy mechanism (See et al., 2017) to assist in the generation process. We then report the joint goal accuracy of SUMBT and TRADE on different versions of the dataset in Table 6, in which MultiWOZ 2.3-cof means MultiWOZ 2.3 with co-reference applied. As can be seen, both methods perform better on MultiWOZ 2.4 than on all the previous versions. We also include the domain-specific accuracy of SOM-DST and STAR in Table 7, which shows that except SOM-DST in the *taxi* domain, both methods demonstrate higher performance in each domain.

3.3 Per-Slot (Slot-Specific) Accuracy

In the previous subsection, we have presented the joint goal accuracy and average slot accuracy of nine state-of-the-art dialogue state tracking models. The results have strongly verified the quality of our refined annotations. Here, we further report the per-slot (slot-specific) accuracy of SUMBT on different versions of the MultiWOZ dataset. The slot-specific accuracy is defined as the ratio of dialogue turns in which the value of a particular slot has been correctly predicted. The results are shown in Table 8, from which we can observe that the majority of slots (21 out of 30) demonstrate higher accuracies on MultiWOZ 2.4. Even though MultiWOZ 2.3-cof additionally introduces the co-reference annotations as a kind of auxiliary information, it still

Slot	MultiWOZ 2.1	MultiWOZ 2.2	MultiWOZ 2.3	MultiWOZ 2.3-cof	MultiWOZ 2.4
attraction-area	95.94	95.97	96.28	**96.80**	96.38
attraction-name	93.64	93.92	95.28	94.59	**96.38**
attraction-type	96.76	97.12	96.53	96.91	**98.24**
hotel-area	94.33	94.44	94.65	95.02	**96.16**
hotel-book day	98.87	99.06	99.04	99.32	**99.52**
hotel-book people	98.66	98.72	98.93	99.17	**99.19**
hotel-book stay	99.23	99.50	99.70	99.70	**99.88**
hotel-internet	97.02	97.02	97.45	97.56	**97.96**
hotel-name	94.67	93.76	94.71	94.71	**96.92**
hotel-parking	97.04	97.19	97.90	98.34	**98.68**
hotel-pricerange	96.00	96.23	95.90	96.40	**96.59**
hotel-stars	97.88	97.95	97.99	98.09	**99.16**
hotel-type	94.67	94.22	**95.92**	95.65	94.75
restaurant-area	96.30	95.47	95.52	96.05	**97.52**
restaurant-book day	98.90	98.91	98.83	**99.66**	98.59
restaurant-book people	98.91	98.98	99.17	99.21	**99.31**
restaurant-book time	99.43	99.24	99.31	**99.46**	99.28
restaurant-food	97.69	97.61	97.49	97.64	**98.71**
restaurant-name	92.71	93.18	95.10	94.91	**96.01**
restaurant-pricerange	95.36	95.65	95.75	96.26	**96.59**
taxi-arriveby	98.36	98.03	98.18	**98.45**	98.17
taxi-departure	96.13	96.35	96.15	**97.49**	96.55
taxi-destination	95.70	95.50	95.56	**97.59**	95.68
taxi-leaveat	98.91	98.96	**99.04**	99.02	98.72
train-arriveby	96.40	96.40	96.54	96.76	**98.85**
train-book people	97.26	97.04	97.29	97.67	**98.62**
train-day	98.63	98.60	99.04	**99.38**	98.94
train-departure	98.43	98.40	97.56	97.50	**99.32**
train-destination	98.55	98.30	97.96	97.86	**99.43**
train-leaveat	93.64	94.14	93.98	93.96	**96.96**

Table 8: Per-slot (slot-specific) accuracy (%) of SUMBT on different versions of the dataset. The results on MultiWOZ 2.1-2.3 and MultiWOZ 2.3-cof are from (Han et al., 2020b). It is shown that most slots demonstrate stronger performance on MultiWOZ 2.4 than on all the other versions.

only shows the best performance in 7 slots. Compared with MultiWOZ 2.1, SUMBT has achieved higher slot-specific accuracies in 26 slots on MultiWOZ 2.4. These results confirm again the utility and validity of our refined version MultiWOZ 2.4.

3.4 Case Study

In order to understand more intuitively why the refined annotations can boost the performance, we showcase several dialogues from the test set in Table 9, where we include the annotations of MultiWOZ 2.1 and MultiWOZ 2.4 and also the predictions of SOM-DST and STAR. It is easy to check that the annotations of MultiWOZ 2.1 are incorrect, while the annotations of MultiWOZ 2.4 are consistent with the dialogue context. As can be observed, for the first four dialogues, the predictions of both SOM-DST and STAR are the same as the annotations of MultiWOZ 2.4. For the last dialogue, the prediction of STAR is consistent with the annotation of MultiWOZ 2.4, while the predicted slot value of SOM-DST is different from the annotations of both MultiWOZ 2.1 and MultiWOZ 2.4. Since MultiWOZ 2.1 and MultiWOZ 2.4 share the same training set, we can conclude that the refined annotations help us evaluate the model performance more properly and the performance obtained on MultiWOZ 2.4 will also be higher.

Dialogue ID	Dialogue Context, Groundtruth Annotations, and Predictions of SOM-DST and STAR			
PMUL1931	**Sys:** We have 6 different guest houses that fit your criteria. Do you have a specific price range in mind? **Usr:** No, it does not matter.			
	MultiWOZ 2.1	MultiWOZ 2.4	SOM-DST	STAR
	hotel-pricerange none	*hotel-pricerange dontcare*	*hotel-pricerange dontcare*	*hotel-pricerange dontcare*
PMUL3158	**Usr:** I want to find a place in town to visit called jesus green outdoor pool.			
	MultiWOZ 2.1	MultiWOZ 2.4	SOM-DST	STAR
	attraction-type swimming pool	*attraction-type none*	*attraction-type none*	*attraction-type none*
MUL1489	**Sys:** Ok, you are all set for cote on Friday, table for 8 at 17:30. Can I help with anything else? **Usr:** Can I have the reference number for the reservation please? **Sys:** Booking was unsuccessful. Can you try another time slot? **Usr:** What about 16:30?			
	MultiWOZ 2.1	MultiWOZ 2.4	SOM-DST	STAR
	restaurant-book time 17:30	*restaurant-book time 16:30*	*restaurant-book time 16:30*	*restaurant-book time 16:30*
PMUL0550	**Sys:** I recommend Charlie Chan. Would you like to reserve a table? **Usr:** Yes. Monday, 8 people, 10:30.			
	MultiWOZ 2.1	MultiWOZ 2.4	SOM-DST	STAR
	restaurant-name Charlie	*restaurant-name Charlie Chan*	*restaurant-name Charlie Chan*	*restaurant-name Charlie Chan*
MUL1697	**Sys:** I am sorry none of them have booking available for that time, another time maybe? **Usr:** Is 09:45 an available time?			
	MultiWOZ 2.1	MultiWOZ 2.4	SOM-DST	STAR
	restaurant-book time 21:45	*restaurant-book time 09:45*	*restaurant-book time 10:45*	*restaurant-book time 09:45*

Table 9: Examples of test set dialogues in which the annotations of MultiWOZ 2.1 are incorrect but the predictions of SOM-DST and STAR are correct (except the prediction of SOM-DST in the last example), as the predicted slot values are consistent with the annotations of MultiWOZ 2.4. Note that only the problematic slots are presented.

4 Discussion

Recall that in MultiWOZ 2.4, we only refined the annotations of the validation set and test set. The annotations in the training set remain unchanged (the same as MultiWOZ 2.1). As a result, all the benchmark models are retrained on the original noisy training set. The only difference is that we use the cleaned validation set to choose the best model and then report the results on the cleaned test set. Even so, we have shown in our empirical study that all the benchmark models obtain better performance on MultiWOZ 2.4 than on all the previous versions. Considering that all the previous refined versions also corrected the (partial) annotation errors in the training set, the superiority of MultiWOZ 2.4 indicates that existing versions haven't fully resolved the incorrect and inconsistent annotations. The cleaned validation set and test set of MultiWOZ 2.4 can more appropriately reflect the true performance of existing models. In addition, the refined validation set and test set can also be combined with the training set of MultiWOZ 2.3 and thus even higher performance of existing methods can be expected, as MultiWOZ 2.3 has the cleanest training set by far.

On the other hand, it is well-understood that deep (neural) models are data-hungry. However, it is costly and labor-intensive to collect high-quality large-scale datasets, especially dialogue datasets that involve multiple domains and multiple turns. The dataset composed of a large noisy training set and a small clean validation set and test set is more common in practice. In this regard, our refined dataset is a better reflection of the realistic situation we encounter in our daily life. Moreover, a noisy training set may motivate us to design more robust and noise-resilient training paradigms. As a matter of fact, noisy label learning (Han et al., 2020a; Song et al., 2020) has been widely studied in the

machine learning community to train robust models from noisy training data. Numerous advanced techniques have been investigated as well. We hope to see these techniques can also be applied to the study of dialogue systems and thus accelerate the development of conversational AI.

5 Conclusion

In this work, we introduce MultiWOZ 2.4, an updated version of MultiWOZ 2.1, by rectifying all the annotation errors in the validation set and test set. We keep the annotations in the training set as is to encourage robust and noise-resilient model training. We further benchmark nine state-of-the-art dialogue state tracking models on MultiWOZ 2.4 to facilitate future research. All the chosen benchmark models have demonstrated much better performance on MultiWOZ 2.4 than on MultiWOZ 2.1, verifying the quality of the refined annotations.

Potential Impacts

We believe that our refined dataset MultiWOZ 2.4 would have substantial impacts in the academia. At first, the cleaned validation set and test set can help us evaluate the performance of dialogue state tracking models more properly and fairly, which undoubtedly is beneficial to the research of task-oriented dialogue systems. In addition, MultiWOZ 2.4 may also serve as a potential dataset to assist the research of noisy label learning in the machine learning community, especially given that many existing noisy label learning methods rely on synthesized noisy data to test their effectiveness.

References

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.

Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. 2020. Parallel interactive networks for multi-domain dialogue state generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1921–1931, Online. Association for Computational Linguistics.

Jianpeng Cheng, Devang Agrawal, Héctor Martínez Alonso, Shruti Bhargava, Joris Driesen, Federico Flego, Dain Kaplan, Dimitri Kartsaklis, Lin Li, Dhivya Piraviperumal, Jason D. Williams, Hong Yu, Diarmuid Ó Séaghdha, and Anders Johannsen. 2020. Conversational semantic parsing for dialog state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8107–8117, Online. Association for Computational Linguistics.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany. Association for Computational Linguistics.

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

Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W Tsang, James T Kwok, and Masashi Sugiyama. 2020a. A survey of label-noise representation learning: Past, present and future. *arXiv preprint arXiv:2011.04406*.

Ting Han, Ximing Liu, Ryuichi Takanobu, Yixin Lian, Chongxuan Huang, Wei Peng, and Minlie Huang. 2020b. Multiwoz 2.3: A multi-domain task-oriented dataset enhanced with annotation corrections and co-reference annotation. *arXiv preprint arXiv:2010.05594*.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. TripPy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 35–44, 1st virtual meeting. Association for Computational Linguistics.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. 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, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2020. Efficient dialogue state tracking by selectively overwriting memory. In *Proceedings*

of the 58th Annual Meeting of the Association for Computational Linguistics, pages 567–582, Online. Association for Computational Linguistics.

Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. SUMBT: Slot-utterance matching for universal and scalable belief tracking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5478–5483, Florence, Italy. Association for Computational Linguistics.

Shiyang Li, Semih Yavuz, Kazuma Hashimoto, Jia Li, Tong Niu, Nazneen Rajani, Xifeng Yan, Yingbo Zhou, and Caiming Xiong. 2021. Coco: Controllable counterfactuals for evaluating dialogue state trackers. *arXiv preprint arXiv:2010.12850*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

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

Yong Shan, Zekang Li, Jinchao Zhang, Fandong Meng, Yang Feng, Cheng Niu, and Jie Zhou. 2020. A contextual hierarchical attention network with adaptive objective for dialogue state tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6322–6333, Online. Association for Computational Linguistics.

Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. 2020. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*.

Adam Summerville, Jordan Hashemi, James Ryan, and William Ferguson. 2020. How to tame your data: Data augmentation for dialog state tracking. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 32–37, Online. Association for Computational Linguistics.

Yexiang Wang, Yi Guo, and Siqi Zhu. 2020. Slot attention with value normalization for multi-domain dialogue state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3019–3028, Online. Association for Computational Linguistics.

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

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

Fanghua Ye, Jarana Manotumruksa, Qiang Zhang, Shenghui Li, and Emine Yilmaz. 2021. Slot self-attentive dialogue state tracking. In *Proceedings of the Web Conference 2021*, pages 1598–1608.

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

Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wang, Philip Yu, Richard Socher, and Caiming Xiong. 2020. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 154–167, Barcelona, Spain (Online). Association for Computational Linguistics.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1458–1467, Melbourne, Australia. 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.