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

Fanghua Ye
University College London
London, UK

Jarana Manotumruksa
University College London
London, UK

Emine Yilmaz
University College London
London, UK

{fanghua.ye.19, j.manotumruksa, emine.yilmaz}@ucl.ac.uk

Abstract

The MultiWOZ 2.0 dataset has greatly stimulated the research of task-oriented dialogue systems. However, its state annotations contain substantial noise, which hinders a proper evaluation of model performance. To address this issue, massive efforts were devoted to correcting the annotations. Three improved versions (i.e., MultiWOZ 2.1-2.3) have then been released. Nonetheless, there are still plenty of incorrect and inconsistent annotations. This work introduces MultiWOZ 2.4, which refines the annotations in the validation set and test set of MultiWOZ 2.1. The annotations in the training set remain unchanged (same as MultiWOZ 2.1) to elicit robust and noise-resilient model training. We benchmark eight state-of-the-art dialogue state tracking models on MultiWOZ 2.4. All of them demonstrate much higher performance than on MultiWOZ 2.1[1].

1 Introduction

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.

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 in 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.

Massive efforts have 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., 2021). However, they both have some limitations. More concretely, MultiWOZ 2.2 allows the presence of multiple values in the dialogue state. But it does not cover all the value variants. 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 incorrect and inconsistent annotations in the validation set and test set, and fix them meticulously. This refinement results in changes to the state annotations of more than 41% 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, our empirical study shows that much better performance can be achieved on MultiWOZ 2.4 than on all the previous versions.

[1]MultiWOZ 2.4 is released to the public at https://github.com/smartyfh/MultiWOZ2.4.

Proceedings of the SIGdial 2022 Conference, pages 351–360
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022

Error Type	Conversation Example	MultiWOZ 2.1	MultiWOZ 2.4
(I) 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
	Usr: I am looking for a restaurant that serves Portuguese food.	rest.-food=Portugese	rest.-food=Portuguese
(II) Missing Annotation	**Usr:** I need a place to dine in the centre of town.	rest.-area=none	rest.-area=centre
	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
(III) Not Mentioned	**Usr:** I am planning a trip in Cambridge.	hotel-internet=dontcare	hotel-internet=none
(IV) 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
	Usr: Something classy nearby for dinner, preferably Italian or Indian cuisine?	rest.-food=Indian	rest.-food=Indian|Italian
(V) Implicit Time Processing	**Usr:** I need a train leaving after 10:00.	train-leaveat=10:15	train-leaveat=10:00
(VI) 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. Only the problematic slots are presented. "rest." is short for restaurant.

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., 2020). 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 MultiWOZ 2.0 & 2.1, the dialogue state is represented as a series of *slot-value* pairs. For example, *attraction-area=centre* means that the slot is *attraction-area* and its value is *centre*. Considering that MultiWOZ 2.1 has significantly improved the annotation quality of MultiWOZ 2.0, we choose to continue the refinement on the basis of MultiWOZ 2.1. Another choice is to perform the refinement on top of MultiWOZ 2.2. However, as mentioned earlier, MultiWOZ 2.2 allows each slot to have multiple value variants. This relaxation increases the difficulty of annotating. It is challenging to include all the value variants. New value variants may also emerge as time goes by. Even worse, some value variants are ambiguous and invalid. For instance, "Peking" can be a shared variant of "Peking University" and "Peking restaurant". Hence, it is an ambiguous value variant. Besides, the benchmark evaluation on MultiWOZ 2.2 shows no evident performance improvements over MultiWOZ 2.1 (Zang et al., 2020). In light of these, MultiWOZ 2.1 is a better basis for our refinement.

2.1 Annotation Error Types

The main goal of dialogue state tracking is to track what has been uttered by a user. Thus, it is generally assumed that the dialogue state should mainly rely on user utterances[2]. Based on this assumption, we identify and fix six types of annotation errors in the validation set and test set of MultiWOZ 2.1. Figure 1 shows examples for each error type.

Context Mismatch: The slot value is inconsistent with the one mentioned in the dialogue context. We also include values with typos in this error type.

Missing Annotation: The slot is unlabelled, even though its value has been mentioned. In some cases, the annotations are delayed to later turns.

Not Mentioned: The slot has been annotated, however, its value has not been mentioned at all.

Incomplete Value: The slot value is a substring or an abbreviation of its full shape (e.g., "Thurs" vs. "Thursday"). In some cases, the slot should have multiple values, but not all values are included.

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 15 min)[3].

[2] If the user requirements cannot be satisfied (e.g., a restaurant asked by the user does not exist), the system should still track the "wrong" requirements as the dialogue state and then ask a clarification question (Doğan et al., 2022) to the user.

[3] The value is implicitly processed when the time is after or before a certain point. Albeit reasonable, it is hard to decide the exact time offset. Thus, we copy the specified time directly.

Refinement Type	Count	Ratio(%)
no change	432,972	97.90
none→value	3,230	0.73
valueA/dontcare→valueB	1,598	0.36
value/dontcare→none	2,846	0.64
none/value→dontcare	1,614	0.36

Table 1: The count and ratio of slot values changed in MultiWOZ 2.4 compared with MultiWOZ 2.1.

Unnecessary Annotation: These unnecessary annotations 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 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

The validation set and test set of MutliWOZ 2.1 contain 2,000 dialogues with more than 14,000 dialogue turns. These dialogues span over 5 domains with a total of 30 slots. 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 means 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 highest 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.

Dataset	Slot(%)	Turn(%)	Dialogue(%)
val	5.04	42.61	67.40
test	5.17	39.74	64.16
total	5.10	41.17	65.78

Table 2: The ratio of refined slots, turns and dialogues.

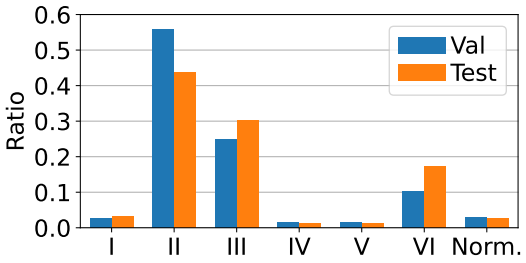

Figure 2: The ratio of different error types. "Norm." refers to values normalized based on their frequency.

2.3 Statistics on Refined Annotations

Table 1 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 the other slots always take the value none. Table 2 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 41% 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.

Figure 2 illustrates the distribution of different error types. We also treat unnormalized values (cf. §2.2) as a special type of errors. Figure 2 shows that "*Missing Annotation*" and "*Not Mentioned*" are the two most frequent error types. It also shows that more than 10% of errors are related to "*Unnecessary Annotation*", while the other types of errors only account for a relatively small proportion.

3 Benchmark Evaluation

3.1 Benchmark Models

Existing neural dialogue state tracking 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

	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
	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**	64.75 (+9.57)	64.27	97.48	98.33

Table 3: 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 4: 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 5: Comparison of domain-specific test set joint goal accuracy.

context. We benchmark the performance of our refined dataset on both types of methods, including SUMBT (Lee et al., 2019), STAR (Ye et al., 2021), TRADE (Wu et al., 2019), PIN (Chen et al., 2020), SOM-DST (Kim et al., 2020), SimpleTOD (Hosseini-Asl et al., 2020), SAVN (Wang et al., 2020), and TripPy (Heck et al., 2020).

3.2 Benchmark Results

We adopt joint goal accuracy (Zhong et al., 2018) and slot accuracy as evaluation metrics. The joint goal accuracy is defined as the ratio of dialogue turns in which all slot values are correctly predicted. The slot accuracy is defined as the average accuracy of all slots. As shown in Table 3, all models achieve much higher performance on MultiWOZ 2.4. SimpleTOD 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. SAVN also shows a low performance increase, as it has already utilized value normalization to tackle label variants in MultiWOZ 2.1. We then report the joint goal accuracy of SUMBT and TRADE on different versions of the dataset in Table 4, in which MultiWOZ 2.3-cof means MultiWOZ 2.3 with coreference applied. As can be seen, both methods perform better on MultiWOZ 2.4 than on all previous versions. We include the domain-specific accuracy of SOM-DST and STAR in Table 5, which shows that except SOM-DST in the *taxi* domain, both methods demonstrate higher performance in each domain of MultiWOZ 2.4.

4 Human Evaluation

We also perform a human evaluation on the quality of the refined annotations. We randomly sampled 50 dialogues from the test set and recruited 5 computer science students to compare our refinement against the annotations in MultiWOZ 2.1. Specifically, the raters were asked to assign a score to each turn of the sampled dialogues based on the following criteria: 1) **-2**: A score of -2 means that both the refined annotation and original annotation are not completely correct; 2) **-1**: A score of -1 means that the original annotation is correct while the refined annotation is problematic; 3) **0**: A score of 0 means that both the refined annotation and original annotation are correct, that is, no changes have been made to the original annotation; 4) **1**: A score of 1 means that the refined annotation is correct while the original annotation is invalid.

We obtain an average score of 0.1653, meaning that our refined annotations are more accurate. We further employ Fleiss' kappa (Fleiss, 1971) to measure the level of agreement among different raters. We obtain $\kappa = 0.9226$, which indicates an almost perfect agreement across the five raters.

We illustrate the score distributions of different raters in Figure 3. From this figure, we can intuitively observe that there is a high level of agree-

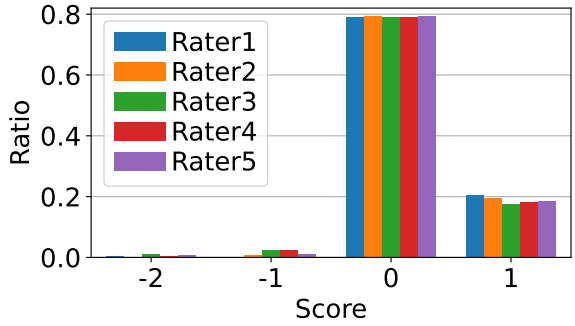

Figure 3: The score distribution of different raters.

ment among the five raters. Figure 3 also shows that in most cases, the refined annotation and the original annotation are both correct, meaning that there is no need to make any changes to the original annotation. This is desirable, as our refinement is based on MultiWOZ 2.1 which has already fixed lots of annotation errors. Around 20% of annotations in MultiWOZ 2.4 are deemed to be more accurate than MultiWOZ 2.1, while only about 1% of annotations in MultiWOZ 2.1 are evaluated as better. This verifies again that our refinement has higher quality.

We further inspected the annotations in MultiWOZ 2.1 that are assessed to be more appropriate. We found that these annotations are mainly related to the slot *hotel-type*. This slot has four candidate values {"hotel", "guest house", "none", "dontcare"}, which are relatively confusing because the term "hotel" is also one candidate value. In practice, when a user says "I am looking for a hotel with 4 stars", the user may actually mean that "I am looking for a place to stay with 4 stars". However, by convention, the term "hotel" is used more often, even though the user does not mean that the hotel type must be "hotel". In our refinement procedure, we chose to annotate this slot based on the whole dialogue session to understand the true user intention (i.e., *hotel type=hotel*?) while the raters tended to take into account only the dialogue history. This ambiguous slot tells us that it is crucial to develop appropriate slots and candidate values that will not cause any confusions to the annotators.

5 Caveats and Lessons Learned

Although we have tried our best to correct as many annotations in the validation set and test set as possible, it is unlikely that we have fixed all the annotation errors. In fact, there are several challenges we faced during the refinement process that are particularly difficult to overcome. Firstly, as discussed earlier, the candidate values of some slots are confusing, which makes it really challenging to choose the most appropriate value. Secondly, in some scenarios, the user intention can have different interpretations. For example, the user utterance "the hotel does not need to have internet though" can mean that the user does not need internet at all (*hotel-internet=no*) or the user does not care about if the internet is provided (*hotel-internet=dontcare*). Thirdly, some slots may have multiple values. Sometimes these values should even be ordered according to users' preferences. When there are too many values (more than two), it is also questionable if the corresponding slot should be annotated. Suppose that the system recommended 10 museums to the user and the user asked "Does any of them have zero entrance fee?", should the slot *attraction-name* be annotated?

Further, the dialogue state can be regarded as a structured representation of the complex user intentions. Due to the complexity of the language itself, some information will be inevitably lost when transforming unstructured user utterances into structured state representations. In this regard, dialogue state annotating is in essence a challenging task.

Given these challenges, it is necessary to define unambiguous slots and unconfusable candidate values to facilitate state annotating. It is also important to provide annotators with full instructions for each slot so that they can make consistent annotations.

6 Conclusion

We introduce MultiWOZ 2.4, an updated version of MultiWOZ 2.1, by rectifying (almost) 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 eight state-of-the-art dialogue state tracking models on MultiWOZ 2.4 to facilitate future research. All the benchmark models have demonstrated much better performance on MultiWOZ 2.4 than on MultiWOZ 2.1.

MultiWOZ 2.4 can also be applied to train better overall dialogue systems, e.g., by utilizing data augmentation techniques to generate high-quality training data based on the clean validation set.

Acknowledgments

This project was funded by the EPSRC Fellowship titled "Task Based Information Retrieval" and grant reference number EP/P024289/1.

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.

Fethiye Irmak Doğan, Ilaria Torre, and Iolanda Leite. 2022. Asking follow-up clarifications to resolve ambiguities in human-robot conversation. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, pages 461–469.

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.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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

Ting Han, Ximing Liu, Ryuichi Takanabu, Yixin Lian, Chongxuan Huang, Dazhen Wan, Wei Peng, and Minlie Huang. 2021. 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.

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*.

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.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

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.

A Additional Statistics on the Refined Annotations

In Table 6, we report the value vocabulary size (i.e., the number of candidate values) of each slot in MultiWOZ 2.1 & 2.4, respectively. We also report their value change ratios. As can be observed, 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 6 also indicates that the `name`-related slots have the highest value change ratio. Since these slots usually have "longer" values, the annotators are more likely to make incomplete and inconsistent annotations.

Slot	2.1	2.4	Val(%)	Test(%)
attraction-area	7	8	1.97	1.93
attraction-name	106	92	5.34	5.16
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	7.81	5.90
rest.-pricerange	5	7	1.51	2.05
taxi-arriveby	62	61	0.41	0.56
taxi-departure	177	172	0.92	0.86
taxi-destination	185	181	1.14	0.75
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 6: 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.

B Per-Slot (Slot-Specific) Accuracy

In Section 3, we have presented the joint goal accuracy and average slot accuracy of eight state-of-the-art dialogue state tracking models. The results have demonstrated that much better performance can be achieved on our refined annotations in terms of the two metrics. 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 7, 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 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.

C Case Study

Except for the quantitative analyses provided in the benchmark evaluation and human evaluation, we also conduct a qualitative analysis to understand more intuitively why and how the refined annotations boost the performance of evaluation. To this end, we showcase several dialogues from the test set in Table 8, 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 and therefore are valid. From Table 8, we also observe that the predictions of both SOM-DST and STAR are the same as the annotations of MultiWOZ 2.4 in the first four dialogues. In the last dialogue, the prediction of STAR is consistent with the annotation of MultiWOZ 2.4, whereas the predicted slot value of SOM-DST is different from the annotations of both MultiWOZ 2.1 and MultiWOZ 2.4. These examples show that the performance of existing dialogue state tracking models is underestimated because of the invalid annotations in MultiWOZ 2.1. While MultiWOZ 2.4 can better manifest the true model performance owing to the refined annotations that align well with the dialogue context.

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 7: Per-slot (slot-specific) accuracy (%) of SUMBT on different versions of the MultiWOZ dataset. The results on MultiWOZ 2.1-2.3 and MultiWOZ 2.3-cof are from (Han et al., 2021). It is shown that most slots demonstrate stronger performance on MultiWOZ 2.4 than on all the other versions.

D 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 the benchmark models can 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 have not fully resolved the incorrect and inconsistent annotations. Therefore, although there have been three refined versions, our refinement is still necessary and meaningful. In addition, the refined validation set and test set can be combined with the training set of MultiWOZ 2.3. Since MultiWOZ 2.3 has the cleanest training set by far, this combination has the potential to result in even higher performance of existing methods.

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

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 8: 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 dialogue context. Given that the annotations of MultiWOZ 2.4 are consistent with the dialogue context as well, there is no doubt that higher performance can be achieved when performing evaluation on MultiWOZ 2.4. Note that only the problematic slots are presented.

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 view of this, 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., 2020; Song et al., 2022) 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 that these techniques can also be applied to the study of dialogue systems and thus accelerate the development of conversational AI.

E Potential Impacts

We believe that our refined dataset MultiWOZ 2.4 would have substantial impacts in academia. First of all, the cleaned validation set and test set can help us evaluate the performance of dialogue state tracking models more properly and fairly, which is undoubtedly 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. The advantage of MultiWOZ 2.4 is that it is a multi-label dataset with real noise in the training set. In the machine learning community, it has been recognized as a future research direction to study noisy label learning for multi-label classification (Song et al., 2022).