Oh My Mistake!: Toward Realistic Dialogue State Tracking including Turnback Utterances

Takyoung Kim[†], **Yukyung Lee**[†], **Hoonsang Yoon**[†],
Pilsung Kang[†], **Junseong Bang**[‡], **Misuk Kim**[§]
Korea University, Seoul 02841, Republic of Korea[†]
Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea[‡]
Sejong University, Seoul 05006, Republic of Korea[§]
{takyoung_kim, yukyung_lee, hoonsang_yoon, pilsung_kang}@korea.ac.kr
misuk.kim@sejong.ac.kr

Abstract

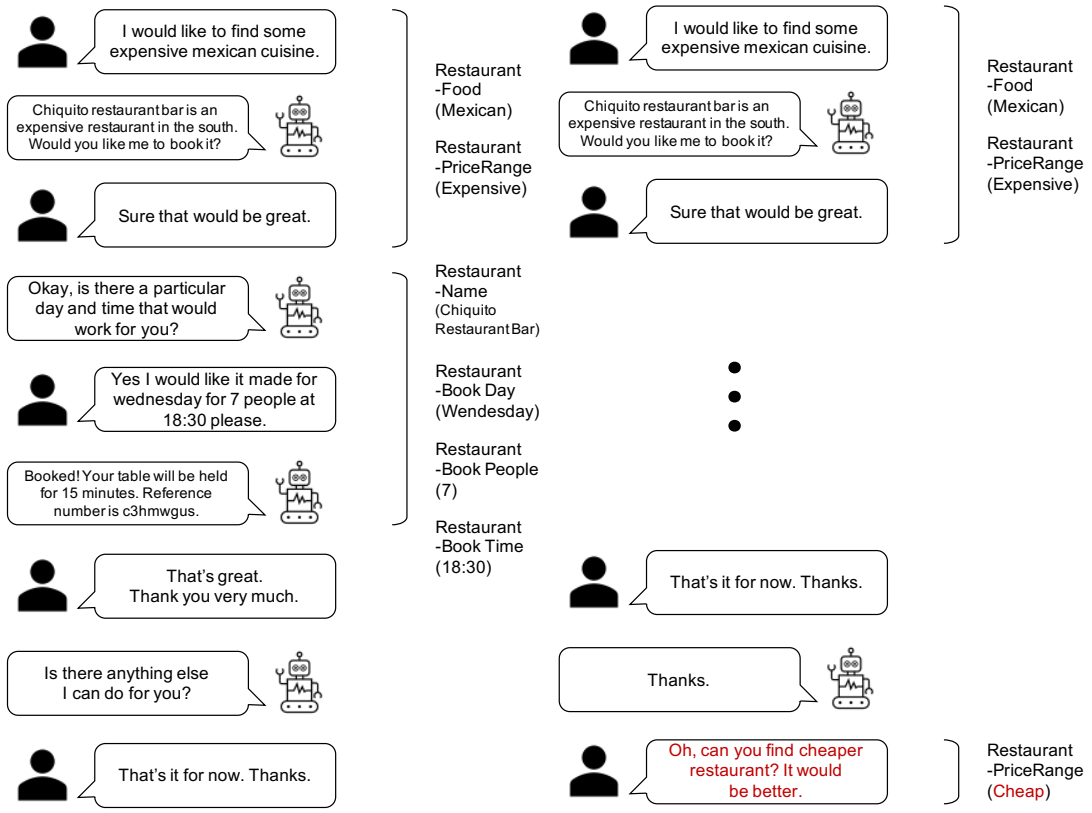
The primary purpose of dialogue state tracking (DST), a critical component of an end-to-end conversational system, is to build a model that responds well to real-world situations. Although we often change our minds from time to time during ordinary conversations, current benchmark datasets do not adequately reflect such occurrences and instead consist of over-simplified conversations, in which no one changes their mind during a conversation. As the main question inspiring the present study, "Are current benchmark datasets sufficiently diverse to handle casual conversations in which one changes their mind after a certain topic is over?" We found that the answer is "No" because DST models cannot refer to previous user preferences when template-based turnback utterances are injected into the dataset. Even in the the simplest mind-changing (turnback) scenario, the performance of DST models significantly degenerated. However, we found that this performance degeneration can be recovered when the turnback scenarios are explicitly designed in the training set, implying that the problem is not with the DST models but rather with the construction of the benchmark dataset.

1 Introduction

The dialogue state tracking (DST) module is a part of a task-oriented dialogue system, the main role of which is to extract essential information of user preferences from various conversational situations. Based on the given information from the previous module, the DST module finds appropriate slot-value pairs to understand the current conversational situations, and these pairs are then delivered to the next module to continue the conversation. Hence, building an accurate DST model is a key success factor of the overall task-oriented dialogue system not only because it can convince users that the system perfectly understands what they are talking about, but also because appropriate responses can be generated based on the result of the DST model. As in other natural language processing (NLP) tasks, two main components are mandatory to build a good DST model: (1) well-structured machine learning models and (2) sufficiently large datasets that contain various real-world conversational situations with fewer biases for training the model. Since the introduction of Transformer and BERT (Vaswani et al., 2017; Devlin et al., 2018), various breakthrough model structures have been designed for DST, such as SUMBT and SOM-DST (Lee et al., 2019; Kim et al., 2020), and have shown an excellent performance. With respect to DST-specific datasets, by contrast, some benchmark datasets, such as WOZ (Wen et al., 2017) and MultiWOZ (Budzianowski et al., 2018), have been introduced; however, their sizes and coverage are not yet satisfactory owing to the relatively high labeling cost. For example, the MultiWOZ only consists of approximately 10,000 dialogues from some different domains, which is significantly smaller than other NLP datasets such as SQuAD or IMDB (Rajpurkar et al., 2016; Maas et al., 2011).

Whereas the MultiWOZ has been used as a standard benchmark dataset for DST, there has been an increasing number of recent studies reporting the concerns regarding the inherent limitations of this dataset. First, newer versions of MultiWOZ have been proposed to address certain issues such as annotation errors, typos, standardization, annotation consistency, and other factors (Eric et al., 2019; Zang et al., 2020; Han et al., 2020; Ye et al., 2021). In addition, Qian et al. (2021) pointed out an entity bias issue, i.e., only a small number of values in the ontology account for the majority of labels. For example, a large number of *'train-destination'* slots take the value *'cambridge'* in the MultiWOZ (Qian et al., 2021). In addition, with CoCo (Li et al., 2020), an overestimation of the held-out accuracy was pointed out by showing that the training and evaluation sets of the MultiWOZ

Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD), pages 1 - 12



(a) Benchmark (b) In reality

Figure 1: Dialogue flow example of MultiWOZ 2.1 (MUL1514.json).

have a similar distribution, and controllable counterfactual goals were proposed that do not change the original dialogue flow but generate a new dialogue with different responses.

Although previous studies have raised inherent problems in the MultiWOZ, most have tended to focus on correcting the annotation inconsistency or entity biases, which enforces the dialogue in the dataset to be more idealistic. However, in real-world conversations, the dialogue flow between two speakers is not always as fluent as those in the MultiWOZ, e.g., one can occasionally change one's mind during a conversation. For example, Figure 1a shows a sample dialogue in the MultiWOZ. No slot that appears once appears again in the subsequent dialogue turns. As the main hypothesis motivating this study, real conversations do not always continue as shown in Figure 1a, but often continue as shown in Figure 1b. Individuals change their mind during a conversation, and thus some slot-value pairs (same slot but different values) repeatedly appear in an entire dialogue. This hypothesis has led us to raise the main question of this paper: *"Can the current benchmark dataset handle a situation in which users change their mind after a certain amount of turn?"* Our assumption is that the turnback situation of a user will hamper the robust evaluation of DST models because such models do not have a chance to learn the situation in which the values of specific slots are changed during the conversation. To experimentally verify our assumption, we investigate how DST models handle additional turnback dialogues by injecting template-based utterances under different scenarios on the MultiWOZ.

It is common for users to change their decisions in various ways in the real world, and thus we define four turnback situations as follows:

- **SINGLE TURNBACK** : This is the simplest form in which the user changes the decision of a single slot only once.
- **RETURN TURNBACK** : This is the reverse of a decision twice but returning to the original value of a single slot.
- **DUAL-VALUE TURNBACK** : The decision for a single slot is changed twice and thus the corresponding values are also changed twice.
- **DUAL-SLOT TURNBACK** : The decision for

two slots are sequentially changed. The corresponding values are changed only once.

The remaining states are more complicated variants of the simplest versions by modifying the number of repetitions or slots. There are some ways to generate turnback utterances such as manually annotating dialogues or generating with the help of language models (Raffel et al., 2020). In this study, we injected turnback utterances at the end of the existing dialogue using pre-defined templates for two reasons. First, locating turnback utterances at the end of the dialogue is a better way to verify the ability handling long-range contexts for the model. Second, template-based-generated utterances explicitly mention the information of `domain`, `slot`, and `value` in a raw text, which can play a role as the minimal form of turnback scenarios. We found even these simple and explicit forms of turnback utterances are sufficient to disclose the problem.

In this paper, we evaluate the performance of turnback situations with TRADE, SUMBT, and Transformer-DST (Wu et al., 2019; Lee et al., 2019; Zeng and Nie, 2021). The results show that existing models cannot detect changing user preferences when injecting turnback utterances in the test set; the same trends are also shown in all variants of turnback scenarios. We further determined that including turnback utterances appropriately during the training phase can make a model robust because the model performance rebounds. To summarize, the main contributions of this paper can be summarized as follows:

- We define the problem that the current benchmark cannot handle, i.e., the change in decision of the user after a certain topic is over, which must be considered when constructing an realistic conversational system.
- We quantitatively and qualitatively evaluate three representative DST models to verify the effect of the turnback situation by injecting template-based utterances into the existing dataset.
- We explore the effect of various turnback proportions in both the training and testing datasets: When turnback utterances appear in the test set, models trained with the data including turnback utterances become more robust.

2 Related Work

2.1 Limitation of Benchmark Dataset

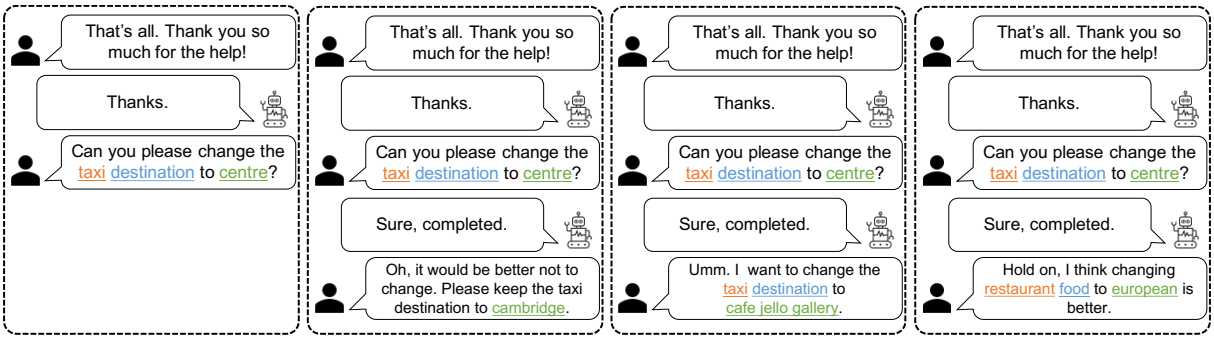
MultiWOZ (Budzianowski et al., 2018) is one of the most popular multi-domain task-oriented dialogue datasets. Although a new task-oriented dialogue dataset, such as SGD (Rastogi et al., 2020), has been recently proposed, most previous studies still evaluate the performance based on MultiWOZ (Kim et al., 2022). However, it has been revealed that the MultiWOZ has inherent errors and biases, and several studies have been proposed to resolve the reported issues.

Annotation error Even the recent versions of MultiWOZ still have incorrect labels and inconsistent annotations (Eric et al., 2019; Zang et al., 2020; Han et al., 2020; Ye et al., 2021). These noises are the primary reason why it is challenging to accurately evaluate the model performance. Fortunately, the benchmark is continuously updated by progressively correcting any annotation errors found.

Biased slots The slots in MultiWOZ are biased. The slots in the training and test sets overlap by more than 90%, and the co-occurrence between slots in the test set is also unequally distributed. DST models are vulnerable to unseen slots because biased slots do not consider rare but realistic slot combinations. To relieve this assumption, CoCo (Li et al., 2020) generates counterfactual dialogues to allow the existing dataset to cover realistic conversation scenarios.

Biased entities Entities in the MultiWOZ are also significantly biased. The test dataset has most of the entities that appear in the training dataset, and existing models are vulnerable to unseen entities (e.g., *"cambridge"* appearing in 50% of the destination cities in the *train* domain) (Qian et al., 2021). Thus, the new test dataset consisting of unseen entities is proposed, which also results in a decrease in performance (Qian et al., 2021).

Change my mind During a real conversation, people often change their minds. For example, when making a reservation for a restaurant, one might change the number of visitors, arrival time, or menu. When catching a taxi, the rider might ask the driver to go to their office first, and suddenly decide to go home to take a rest instead. Someone might want to sleep more, so they might delay their departure time. There are many other examples in which speakers change their mind or decision



(a) SINGLE TURNBACK (b) RETURN TURNBACK (c) DUAL-VALUE TURNBACK (d) DUAL-SLOT TURNBACK

Figure 2: An example of proposed turnback situations. Text in orange denotes a domain, blue denotes a slot, and green denotes a value.

[Train]
1: Umm. I think it's better to change {domain} {slot} to {value}.
2: Can you change {domain} {slot} to {value}? I forgot it.
3: Oh, I need to change {domain} {slot} to {value}. Please fix it.

[Validation]
1: Oh, I took a mistake. Change {domain} {slot} to {value} please.
2: It would be better to change {domain} {slot} to {value}. Can you make it?
3: I forgot about it, I want to change {domain} {slot} to {value}.

[Test]
1: I think {value} is better. I want to change {domain} {slot} to {value}.
2: Wait, it might be better to change {domain} {slot} to {value}.
3: Hold on, I've been thinking about it and
I think changing {domain} {slot} to {value} will be better.

Figure 3: Template utterances of each phase (train, validation, and test).

during a conversation. Unfortunately, the current well-known DST benchmark dataset does not seem to take these scenarios into serious consideration. All conversations continue naturally, and no one reverses what they have said. Some approaches reflect changing decisions of the user but only cover changes in the same dialog topic (Bordes et al., 2017; Mosig et al., 2020). Our contention regarding the conditions of a good DST benchmark dataset is that the conversations in the dataset should reflect more realistic situations, e.g., frequent turnback utterances, which are a main component of ordinary conversations in the real world.

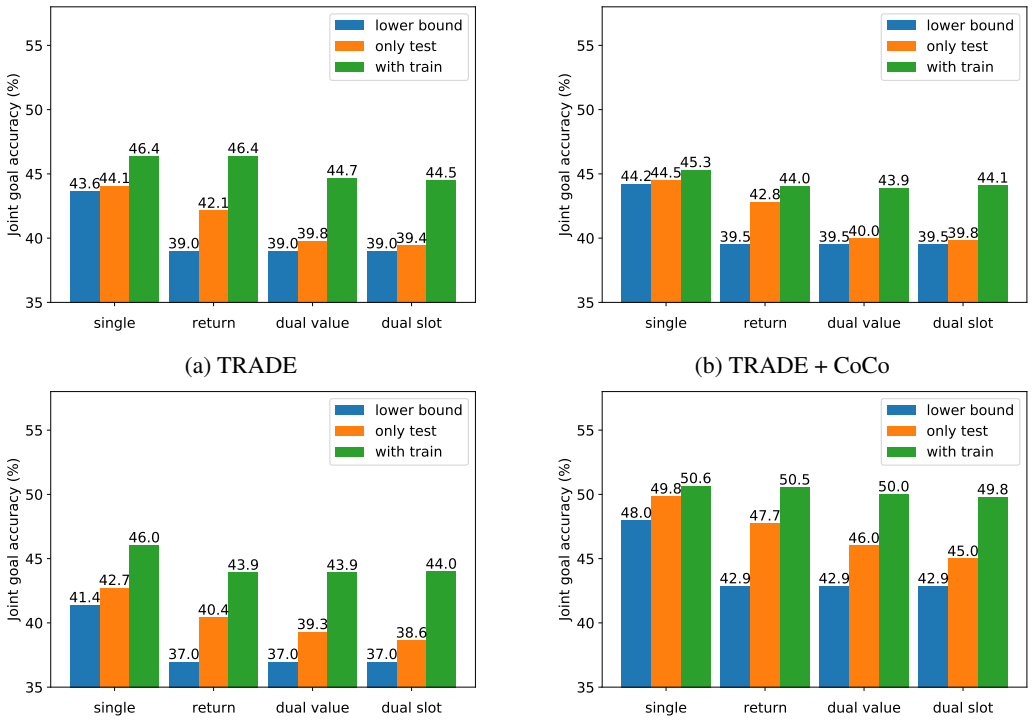
This paper is partially related to Jakobovits et al. (2022), which points out the current task-oriented dialogue benchmark only considers short-term context rather than long history. Our turnback scenarios are the representative phenomena that show the lack of *conversationality* of the benchmark dataset, defined in Jakobovits et al. (2022).

3 Method

To test whether the model trained with the current DST dataset can track the change in value of the turnback situation, we assume four turnback scenarios and inject these turnback utterances at the end of every dialogue, as represented as Figure 2. In other words, each data containing dialogue of t turns can be formulated as $X_t = \{(U_1^{sys}, U_1^{usr}), \ldots, (U_t^{sys}, U_t^{usr})\}$, and we then append an extra template-generated turn with one of the aforementioned turnback situations at the end of the existing data, resulting in $X_k = \{(U_1^{sys}, U_1^{usr}), \ldots, (U_k^{sys}, U_k^{usr})\}$, where $k = t{+}1$ for a single turnback situation or $k = t{+}2$ for multiple situations. Figure 3 shows examples of a turnback used in each dataset. Note that we used different templates for different datasets to avoid an overlap across the datasets. Whenever applying a template-based utterance generation, the arbitrary template of each phase is selected at each turn of dialogue.

As the main purpose of this paper is to investigate whether the model can follow the user's mind-changing utterances, we designed the simplest form of turnback utterances: injecting them to the last turn and generating utterances using templates. The former is to assume the mind-changing within the longest history in a single dialogue, and the latter is to show that models cannot track changing values when even the most informative turnback utterances are explicitly provided. Accordingly, we defined four variants of turnback situations as follows:

SINGLE TURNBACK Users change the value of a particular slot only once, as shown in Figure 2a. Basically, a single turnback utterance is constructed using the last turn of the dialogue because it contains accumulated belief states that appeared throughout the dialogue. Figure A1 shows the process of generating a single turnback utterance and skipping the process when there is no belief stated



(a) TRADE (b) TRADE + CoCo

(c) SUMBT (d) Transformer-DST

Figure 4: Performance gap based on the existence of turnback in the training data. Lower bound indicates the performance of not correctly predicting additional turnback turns at all.

during the dialogue.

RETURN TURNBACK Users change the value of a particular slot but return to the original value again, as shown in Figure 2b. This means that the final belief state after injecting a return turnback utterance is the same as the belief state of the original dataset. In this case, the first turnback utterance can be generated like a single turnback process, and the second turnback utterance is then generated identically by simply replacing the changed value with the original value.

DUAL-VALUE TURNBACK Users sequentially change the value of a particular slot twice, as shown in Figure 2c. Dual value turnback utterances can be generated in the same way as return turnback utterances, but can be generalized to a triple or quadruple value turnback if there are more than two available values in the slot on the ontology.

DUAL-SLOT TURNBACK Users first change the value of a particular slot and then also change the value of a different slot, as represented in Figure 2d. This can be generated simply by applying a single turnback twice; however, there must be more than two total belief states to apply this scenario.

4 Experiments

4.1 Experimental setup

We verified our hypothesis using the MultiWOZ 2.1 (Eric et al., 2019), the most commonly used DST dataset in previous studies. As a performance metric, the joint goal accuracy was employed. The joint goal accuracy is a standard criterion used to check if the model tracks the triplet of `(domain, slot, value)` precisely. When tracked correctly, the joint goal accuracy is marked as 1, and is otherwise 0. The numbers of training, validation, and test sets are 8420, 1000, and 999, respectively. The open-source code for the TRADE model was from CoCo repository[1], while the code for SUMBT[2] and Transformer-DST[3] was from the original author, respectively. For the TRADE model, we also considered the model trained jointly with CoCo-augmented dataset (Li et al., 2020). All the experiments explained later were conducted using a machine with the NVIDIA GeForce RTX 3090 GPU.

[1]https://github.com/salesforce/coco-dst
[2]https://github.com/SKTBrain/SUMBT
[3]https://github.com/zengyan-97/Transformer-DST

4.2 Main results

Figure 4 shows the main results. As the extra turnback utterances are appended to the original dataset, we reported the performance lower bound where the model does not predict the additional states of turnback utterances at all (blue color). In other words, the joint goal accuracy of every turn of turnback scenarios is zero in the lower bound setting. The performance of the original model with turnback-included test set is reported with the orange color. Compared to the lower bound, the model trained with the original set correctly predicts only a few altered dialogue states. In the case of multiple turnbacks (i.e., RETURN TURNBACK , DUAL-VALUE TURNBACK , and DUAL-SLOT TURNBACK), the models with RETURN TURNBACK resulted in relatively better performance than the others. This is not because the model predict the state values in the turnback utterance correctly, but because RETURN TURNBACK has the same value with the original state value. Note that these turnback utterances generated using templates are the *easiest* form of the situation, explicitly providing the entire information of `domain`, `slot`, and `value`.

4.3 Including turnback dialogues in the training set

Because the main hypothesis was sufficiently supported by the first experiment, we further investigated whether including turnback situations in the training dataset can prevent the model from not being able to trace the changing values. We inserted turnback utterances at the end of all training train and validation data, and different template utterances were randomly used for the training and validation phases, as illustrated in Figure 3.

The green-colored bar in Figure 4 shows the joint goal accuracy for each turnback scenario before and after the turnback utterance are included in the training and validation datasets with a performance lower bound of newly added turnback turns. The performance always improves irrespective of the turnback scenarios and DST models. Also note that the performance recovery is more significant for more complicated turnback scenarios. Injecting turnback utterances increases the joint goal accuracy by 1.83%p on average for a single turnback, whereas the average improvement is 4.90%p for the dual slot turnback.

In addition to achieving a quantitative rebound in performance, we also conducted a qualitative comparison of the model predictions before and after the turnback injection in the training and validation datasets. Table 1 shows an example of dual slot turnback dialogue, and the predicted states of the Transformer-DST model are as shown in Table 2. The prediction results of the remaining three turnback situations are also provided in Tables A1, A2, and A3. The first row of Table 2 is the last turn of the original dialogue, and we can see that both the original and dual-trained model predict the belief states correctly. In the second and third rows of the same table, when the values of two slots are sequentially changed, the original model can catch only one changing value (*'finches bed and breakfast'*). Not being able to follow all changes is frequently detected with the original model in other test dialogues. By contrast, the model trained with the turnback utterances can correctly predict the entire belief state, as shown in the last row and the last column of Table 2.

Based on the results shown in Figure 4 and Table 2, we can conclude that the performance degeneration of the DST models is not because the DST model structures are incorrect but because they do not have a chance to train such turnback utterances with the current benchmark DST dataset, which means that the MultiWOZ dataset does not have a sufficient coverage yet for dialogues in the real-world.

4.4 Difference in performance according to turnback proportion

We also conducted an ablation study on how the turnback utterance proportions in the training and test dataset affect the DST performance. We evaluate five different proportions of turnback-injected training and test datasets (i.e., 0%, 30%, 50%, 70%, and 100%) with corresponding turnback-test situations, resulting in a total of 25 combinations of training-test turnback proportions. We named each turnback-mixed dataset *phase-N%*. For example, Train-30% denotes the dataset in which 30% of the turnback utterances are applied to the existing dialogues, and the remaining 70% of the original dialogues are unmodified. The performances of Transformer-DST are shown in Table 3. The performance of the other models are provided in Tables A4, A5, and A6. The last column of the table is the difference between the best-proportion model performance and the original performance.

Turn #	Dialogue History
1	System: " " User: "I need a taxi. I'll be departing from la raza."
2	System: "I can help you with that. When do you need to leave?" User: "I would like to leave after 11:45 please."
3	System: "Where will you be going?" User: "I'll be going to restaurant 17."
4	System: "I have booked for you a black volkswagen, the contact number is 07552762364. Is there anything else I can help you with?" User: "No, that's it. Thank you!"
5	System: "Completed." User: "Wait , it might be better to change **taxi leave at** to **15:00**."
6	System: "Sure. Anything else?" User: "Hold on , I've been thinking about it and I think changing **taxi destination** to **finches bed and breakfast** will be better."

Table 1: Sample dialogue of test set with additional DUAL-SLOT TURNBACK situation (SNG01367.json).

Gold state (label)	**Predicted state (original model)**	**Predicted state (DUAL-SLOT-trained model)**
"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"
"taxi-departure-la raza", "taxi-leaveat-**15:00**", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-**15:00**", "taxi-destination-restaurant 17"
"taxi-departure-la raza", "taxi-leaveat-**15:00**", "taxi-destination-**finches bed and breakfast**"	"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination-**finches bed and breakfast**"	"taxi-departure-la raza", "taxi-leaveat-**15:00**", "taxi-destination-**finches bed and breakfast**"

Table 2: The model prediction on DUAL-SLOT TURNBACK situation at turn 4, 5, and 6 (SNG01367.json).

	SINGLE TURNBACK					
	Train-0%	Train-30%	Train-50%	Train-70%	Train-100%	Difference
Test-0%	**54.47**	54.40	54.32	54.44	52.80	-0.03%p
Test-30%	53.04	53.81	53.84	**54.00**	52.22	0.96%p
Test-50%	52.06	53.44	53.36	**53.46**	51.88	1.40%p
Test-70%	50.90	**52.81**	52.78	52.73	51.12	1.91%p
Test-100%	49.84	51.98	52.23	**52.32**	50.65	2.48%p

* Bold denotes the best, and underline denotes the second-best performance.

Table 3: Joint goal accuracy (%) of Transformer-DST with different SINGLE TURNBACK proportions.

Based on Table 3, we can draw the following observations. First, adding moderate turnback utterances does not significantly affect the performance on Test-0%, which is the original test dataset. The joint goal accuracies of Train-30%, Train-50%, and Train-70% are very close to that of Train-0%. Second, high proportions of turnback utterances in the training set help recover the performance in most cases. With regard to turnback ratio in the training dataset, above 70% of the turnback utterance show the best performance in Table 3, A4, and A6. In the case of Table A5, we expect that conterfactual slot combinations provided in CoCo-augmented dataset can assist the model's robust prediction.

5 Conclusion

A DST model should focus on properly reacting to unpredictable scenarios from a human speaker. From this perspective, using realistic benchmark datasets for the model is crucial. To validate recent DST models trained on the commonly used DST benchmark dataset, we first designed a template-based (but enough to verify the hypothesis) data injection method to create a turnback situation and modified the test dataset by appending one of four trunback scenarios to the end of the dialogue. Our experiment showed that the current model trained using the existing benchmark cannot track the changing values well when users change their decisions. We also conducted additional experiment to investigate whether the model performance can be recovered if the turnback utterances are properly included in the training dataset. Experimental results showed that the joint goal accuracy was improved for all turnback scenarios when the models were trained on the dataset with turnback utterances. The ablation study shows that moderately including the turnback utterances can manage a broader range of turnback proportions. Our experimental results emphasize that constructing a right benchmark dataset is as important as developing an advanced model structure in NLP tasks.

Despite the meaningful results, we argue that the turnback utterance is just one of many situations that can happen in a real-world conversation. If more diverse realistic dialogue scenarios are reflected in the DST benchmark dataset, the bias of models trained on it can be significantly reduced.

Acknowledgment

This research was supported and funded by the Korean National Police Agency. [Pol-Bot Development for Conversational Police Knowledge Services / PR09-01-000-20]

References

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *Proceedings of the 5th International Conference on Learning Representations*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, page 4171–4186.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tür. 2019. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. *CoRR*, abs/1907.01669.

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

Alice Shoshana Jakobovits, Francesco Piccinno, and Yasemin Altun. 2022. What did you say? task-oriented dialog datasets are not conversational!?

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.

Takyoung Kim, Hoonsang Yoon, Yukyung Lee, Pilsung Kang, and Misuk Kim. 2022. Mismatch between multi-turn dialogue and its evaluation metric in dialogue state tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 297–309, Dublin, Ireland. 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. 2020. Coco: Controllable counterfactuals for evaluating dialogue state trackers. In *Proceedings of the 8th International Conference on Learning Representations*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of*

the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Johannes E. M. Mosig, Shikib Mehri, and Thomas Kober. 2020. STAR: A Schema-Guided Dialog Dataset for Transfer Learning. *arXiv e-prints*.

Kun Qian, Ahmad Beirami, Zhouhan Lin, Ankita De, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. 2021. Annotation inconsistency and entity bias in multiwoz. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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, and Emine Yilmaz. 2021. Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *arXiv preprint arXiv:2104.00773*.

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, ACL 2020*, pages 109–117.

Yan Zeng and Jian-Yun Nie. 2021. Jointly optimizing state operation prediction and value generation for dialogue state tracking.

A Appendix

System Utterance: “Is there anything to help?”
User Utterance: “No, that’s all. Thanks.”
Turn Index: 7
Belief State:

{
‘restaurant-food’: ‘european’,
‘taxi-destination’: ‘cambridge’,
‘hotel-name’: ‘lensfield hotel’
}

(a) Get the last turn’s dialog.

System Utterance: “Is there anything to help?”
User Utterance: “No, that’s all. Thanks.”
Turn Index: 8
Belief State:

{
‘restaurant-food’: ‘european’,
‘taxi-destination’: ‘cambridge’,
‘hotel-name’: ‘lensfield hotel’
}

(b) Duplicate dialog and randomly select belief state.

System Utterance: “Is there anything to help?”
User Utterance: “No, that’s all. Thanks.”
Turn Index: 8
Belief State:

{
‘restaurant-food’: ‘european’,
‘taxi-destination’**: ‘stansted airport’**,
‘hotel-name’: ‘lensfield hotel’
}

(c) Replace the value of selected belief state with a different value on ontology.

System Utterance: “Thanks.”
User Utterance: “Wait, it might be better to change taxi destination to stansted airport.”
Turn Index: 8
Belief State:
{
‘restaurant-food’: ‘european’,
‘taxi-destination’: ‘stansted airport’,
‘hotel-name’: ‘lensfield hotel’
}

(d) Change system utterance and apply a template to user utterance.

Figure A1: Process of SINGLE TURNBACK dialogue generation.

Gold state (label)	Predicted state (original model)	Predicted state (SINGLE-trained model)
"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination restaurant 17"
"taxi-departure-, **london liverpool street**", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"	"taxi-departure-**london liverpool street**", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"

Table A1: Model prediction on SINGLE TURNBACK situation at turns 4 and 5 (SNG01367.json).

Gold state (label)	Predicted state (original model)	Predicted state (RETURN-trained model)
"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"
"taxi-departure-, **the copper kettle**", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"	"taxi-departure-**the copper kettle**", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"
"taxi-departure-**la raza**", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"	"taxi-departure-**la raza**", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"	"taxi-departure-**la raza**", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"

Table A2: Model prediction on RETURN TURNBACK situation at turns 4, 5, and 6 (SNG01367.json).

Gold state (label)	Predicted state (original model)	Predicted state (DUAL-VALUE-trained model)
"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-11:45", "taxi-destination-restaurant 17"
"taxi-departure-la raza", "taxi-leaveat-**10:15**", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-**10:15**", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-**10:15**", "taxi-destination-restaurant 17"
"taxi-departure-la raza", "taxi-leaveat-**12:00**", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-10:15", "taxi-destination-restaurant 17"	"taxi-departure-la raza", "taxi-leaveat-**12:00**", "taxi-destination-restaurant 17"

Table A3: Model prediction on DUAL-VALUE TURNBACK situation at turn 4, 5, and 6 (SNG01367.json).

	SINGLE TURNBACK					
	Train-0%	Train-30%	Train-50%	Train-70%	Train-100%	Difference
Test-0%	**49.55**	48.47	48.25	48.11	<u>48.81</u>	-0.74%p
Test-30%	**47.82**	47.41	47.16	47.16	**47.82**	0.00 %p
Test-50%	46.52	46.62	46.41	<u>46.67</u>	**47.24**	0.72%p
Test-70%	45.31	<u>45.92</u>	45.63	45.85	**46.50**	1.19%p
Test-100%	44.05	45.12	45.13	<u>45.29</u>	**46.36**	2.31%p

* Bold denotes the best, and underline denotes the second-best performance.

Table A4: Joint goal accuracy (%) of TRADE with different SINGLE TURNBACK proportions.

	SINGLE TURNBACK					
	Train-0%	Train-30%	Train-50%	Train-70%	Train-100%	Difference
Test-0%	**50.21**	48.40	<u>49.80</u>	47.73	48.05	-0.41%p
Test-30%	<u>48.36</u>	47.30	**48.74**	46.81	47.22	0.38%p
Test-50%	<u>47.13</u>	46.57	**48.16**	46.07	46.62	1.03%p
Test-70%	<u>46.02</u>	45.57	**47.42**	45.38	45.89	1.40%p
Test-100%	44.49	44.75	**46.73**	44.75	<u>45.30</u>	2.24%p

* Bold denotes the best, and underline denotes the second-best performance.

Table A5: Joint goal accuracy (%) of TRADE + CoCo with different SINGLE TURNBACK proportions.

	SINGLE TURNBACK					
	Train-0%	Train-30%	Train-50%	Train-70%	Train-100%	Difference
Test-0%	46.99	46.24	46.32	**47.16**	<u>47.10</u>	0.17%p
Test-30%	45.59	46.57	46.17	<u>47.18</u>	**47.38**	1.79%p
Test-50%	44.80	46.29	45.70	<u>46.70</u>	**47.22**	2.42%p
Test-70%	43.73	45.54	45.13	<u>46.11</u>	**46.39**	2.66%p
Test-100%	42.72	45.01	44.70	<u>45.62</u>	**46.04**	3.32%p

* Bold denotes the best, and underline denotes the second-best performance.

Table A6: Joint goal accuracy (%) of SUMBT with different SINGLE TURNBACK proportions.