Counterfactual Matters: Intrinsic Probing For Dialogue State Tracking

Yi Huang, Junlan Feng*, Xiaoting Wu, Xiaoyu Du
JIUTIAN Team, China Mobile Research
{huangyi, fengjunlan, wuxiaoting, duxiaoyu}@chinamobile.com

Abstract

A Dialogue State Tracker (DST) is a core component of modular task-oriented dialogue systems. Tremendous research progress has been made in past ten years to improve performance of DSTs especially on benchmark datasets. However, their generalization to novel and realistic scenarios beyond the held-out conversations is limited. In this paper, we design experimental studies to answer: 1) How does the distribution of dialogue data affect the performance of DSTs? 2) What are effective ways to probe counterfactual matter for DSTs? Our findings are: the performance variance of generative DSTs is not only due to the model structure itself, but can be attributed to the distribution of cross-domain values. Evaluating iconic generative DST models on MultiWOZ dataset with counterfactuals results in a significant performance drop of up to 34.64% (from 50.91% to 16.27%) in absolute joint goal accuracy. It is believed that our experimental results can guide the future work to better understand the intrinsic core of DST and rethink the suitable way for specific tasks given the application property.

1 Introduction

A dialogue state tracker (DST) is a pillar of today's task-oriented dialogue systems, which maintains user's intentional goals through the course of a dialogue.

In recent years, the creation of large-scale datasets, such as MultiWOZ (Budzianowski et al., 2018), has fueled the advance of DST models, pushing the accuracy of DST from 15.8%, baseline from (Budzianowski et al., 2018) to above 50% (Lee et al., 2019; Eric et al., 2020; Chen et al., 2020; Goel et al., 2019; Gao et al., 2019; Wu et al., 2019; Zhang et al., 2019; Huang et al., 2020). The common belief is that the more abundant the labeled data, the higher the likelihood of learning diverse phenomena, which in turn leads to models that generalize well. In practice, however, generalization remains as a huge challenge (Yogatama et al., 2019; Linzen, 2020).

Motivated by this phenomenon, we aim to address and provide insights into the following question: how well do DST models generalize to the novel but realistic scenarios that are not captured well enough by the held-out evaluation set? Answering this question may take us a step closer to bridging the gap between dataset collection and broader task objectives (Li et al., 2020; Heck et al., 2020).

Most prior work (Iyyer et al., 2018; Jin et al., 2020) focus on adversarial example generation for robustness evaluation. They rely on perturbations made directly on test examples in the held-out set and assume direct access to evaluated models' gradients or outputs, which often leads to unnatural examples or hurt target models deliberately. Our studies in this paper are not this line of research.

Recently, the generation-model based approaches for DST instead of a close-set classification approach have attracted more attention. Wu proposed a TRAnsferable Dialogue statE generator (TRADE) (Wu et al., 2019) that generates dialogue states from utterances using a copy mechanism, facilitating knowledge transfer between domains. The prominent difference from previous one-domain DST models is that TRADE is based on a generation approach instead of a close-set classification approach. Huang proposed a Meta-Reinforced Multi-Domain State Generator (MERET) (Huang et al., 2020) which introduces an end-to-end generative framework with pre-trained language model and copy-mechanism, using RL-based generator to encourage higher semantic relevance in greater exploration space for DST. MERET holds the similar underlying architecture with TRADE. Quan released Modeling Long Context for Task-Oriented Dialogue State Generation

*Corresponding author.

(LCDSG) (Quan and Xiong, 2020), which is a multi-task learning model with a simple yet effective utterance tagging technique and a bidirectional language model as an auxiliary task for task-oriented dialogue state generation. LCDSG follows the similar overall framework with TRADE, too.

We conduct our studies in this paper on top of these models TRADE, MERET and LCDSG: first, we propose a simple and efficient counterfactual-maker policy as a principled approach to generate novel scenarios; then, we take a closer look at data sets, and conduct a deep qualitative analysis on data distribution and model structure impact for the DSTs. The main contributions of this paper are two-fold:

- This paper provides deep analysis of counterfactual probing to mainstream generative DST models.
- This paper empirically examines the performance degradation of generative DSTs at different granularities.

2 Proposed Approach: SVS

Let us define $D = \{(U_1, R_1), ..., (U_T, R_T)\}$ as the set of user utterance and system response pairs in T turns of a dialogue, and $B = \{B_1, ..., B_T\}$ as the dialogue state for each turn. Dialogue state is represented as slot-value pairs, denoted as $B_t = \{(S_1, V_1), ..., (S_J, V_J)\}$ where S_j and V_j ($1 \leq j \leq J$) denote the j-th slot name and slot value at this turn.

We propose a simple counterfactual-maker approach, Slot Value Substitution (SVS). It is used to generate counterfactual dialogue D' and corresponding dialogue state B'. Parameter m is used to represent the ratio of SVS. For each dialogue, m percent slot values in B_T are selected to be substituted. Specifically, for each slot value V_j, if the value does not appear in dialogue history, we keep it as it is. For values that can be substituted, new values are sampled from ontology, a predefined value set for each domain-slot. Then dialogue history is updated by these new values and counterfactual dialogue D' is generated. For state of each previous turn, B_t ($1 \leq t \leq T$) is updated and denoted as B'_t. We get $B' = \{(S_1, V'_1), ..., (S_J, V'_J)\}$ for D' after the update and we do post-processing human validation on the counterfactuals generated by SVS to ensure quality. An example of SVS process is shown in Figure 1.

3 Experiments and analysis

In this section, we first describe our observations and concerns from the experiments and then investigate the reason behind.

To evaluate the DSTs' performance on counterfactual dialogue data, we train DST models following their publicly released implementations on the standard train/dev/test split of MultiWOZ[1] from scratch. Joint goal accuracy is used to be the evaluation metric. It measures the accuracy of model prediction at each dialogue turn, and the output is considered correct if and only if all the predicted values exactly match the ground truth values.

3.1 Behavioral probe : Characterization

We compare joint goal accuracy of TRADE, MERET and LCDSG on counterfactuals generated by SVS with different m. Experimental results are listed in Table 1. Surprisingly with m increasing, joint goal accuracy of each model significantly drops, up to 34.64% when m=100. This behavior makes us very curious about the causes behind. We probe the reasons from the perspective of data distribution, characterizing data instances.

Figure 2 shows an overview of error rate of fifteen domain-slots on three test sets. The error rate of most slots increases as m increases, and *train-departure* increases the most, from 2.55% to 20.69%.

To further understand the deep-dive reasons, we conduct qualitative analysis on the data generated by SVS in the next. Figure 3 shows the slot value distribution of *train-departure*, with left vertical axis referring to the proportion of data sets. Through the visual display, we can see the variance between training and test sets with different m clearly: for example, the slot value *cambridge* gets a proportion drop of 0.431 on test set (0.468 vs. 0.037). The data distribution between training and test sets matches well when m=0, while significantly differs when m=100. More generally, the increase of m exacerbates this difference. Extreme situation goes for those unseen slot values appearing in the new test set when m=100. This illustrates the qualitative distribution in-depth, that is, out-of-distribution (OOD) resulting in performance drop of DSTs responding to above.

We also calculate overall F1 on *train-departure* in Figure 3, with right vertical axis representing

[1] https://www.repository.cam.ac.uk/handle/1810/280608

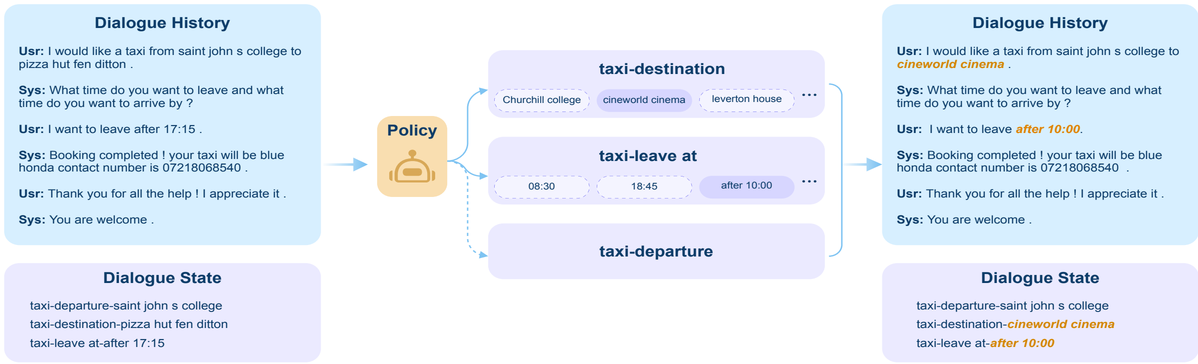

Figure 1: The process of SVS. The left is a dialogue example and the dialogue-level belief state. Value candidates for every checked slot will be substituted by a policy. The right is the substituted dialogue and belief states with new slot values.

DST	Joint goal accuracy							
Model	m=0	m=10	m=20	m=30	m=50	m=80	m=90	m=100
TRADE	0.4913	0.4585	0.3826	0.3290	0.2595	0.1907	0.1768	0.1553
MERET	0.5091	0.4769	0.3977	0.3442	0.2686	0.2039	0.1855	0.1627
LCDSG	0.5103	0.4777	0.4084	0.3557	0.2819	0.2112	0.2032	0.1759

Table 1: Joint goal accuracy of MultiWOZ held-out set with different proportion of slot value substitution. As the proportion of SVS increases, the accuracy drops significantly.

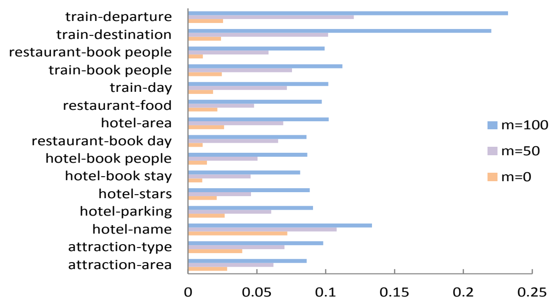

Figure 2: Error rate of multi-domain slots on three test sets. As the ratio of SVS increases, the error rate of *train-departure* and *train-destination* increase by 18.14% and 17.23%.

the results of F1. It shows that for every single slot value, F1 is strongly influenced by the data distribution consistency between training and test sets. The F1 is relatively high when the distribution is consistent. It decreases when m increases. Take *cambridge* for example, 0.956 vs. 0.321 for m=0 and m=100, respectively.

3.2 Structural probe: Attention

In the following, we take TRADE as a representative here, considering the similar structure beings. TRADE consists of two parts in general: a classifier and a copy-mechanism. Copy-mechanism utilizes the generative composition to realize copy action. Detailed experimental results in Table 2 show that the classifier maintains a high accuracy under different conditions. Hence, in the following we focus on analyzing the impact to the second part, the generative composition.

First, we calculate the accuracy of generative composition in counterfactuals generated by SVS with different m. Table 2 is the accuracy of generative composition which shows that as the ratio of SVS increases, the accuracy of generative composition gradually decreases, indicating that the network structure of generative composition is not robust when the ratio of SVS increases.

We reason that counterfactual probing leads to two fundamental changes: the final output distribution and the underlying attention change in different test set. Technically, the final output distribution is:

$$p_{jk}^{f} = p_{jk}^{gen} \times P_{jk}^{v} + (1 - p_{jk}^{gen}) \times P_{jk}^{h} \qquad (1)$$

At decoding step k for the j-th (domain, slot) pair, p_{jk}^{f} is final output distribution. P_{jk}^{h} is the probability of the dialogue and P_{jk}^{v} is the probability of the vocabulary, both of them are impacted by attention. The scalar p_{jk}^{gen} is trainable to combine this two distributions. Figure 4 shows p_{jk}^{gen} rises

Compositions	Accuracy of different m							
	m=0	m=10	m=20	m=30	m=50	m=80	m=90	m=100
Generative composition	0.8949	0.8737	0.8313	0.7932	0.7263	0.6265	0.6015	0.5679
Classifier composition	0.9761	0.9758	0.9746	0.9737	0.9722	0.9700	0.9692	0.9681

Table 2: The accuracy of the generative composition and classifier composition under different ratio of SVS. With m increasing, the accuracy drops from 89.94% to 56.79% in generative composition.

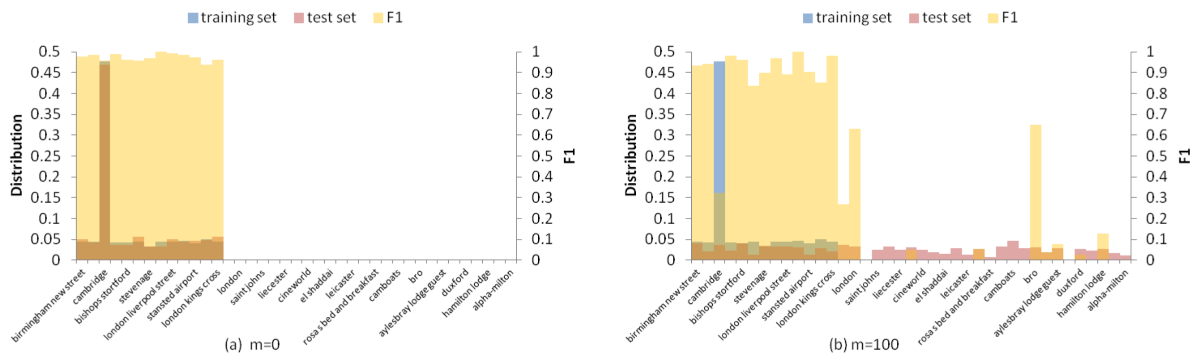

Figure 3: Qualitative analysis for the slot value distribution of *train-departure*. Best viewed in color.

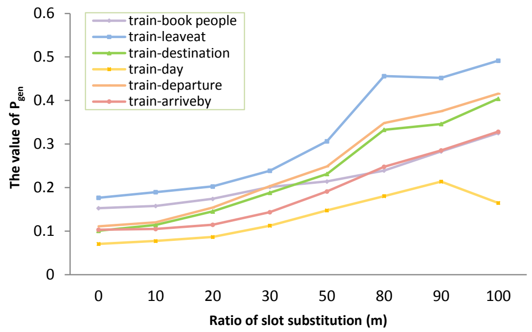

Figure 4: The trend of p_{jk}^{gen} in the *train* domain on different ratio of SVS.

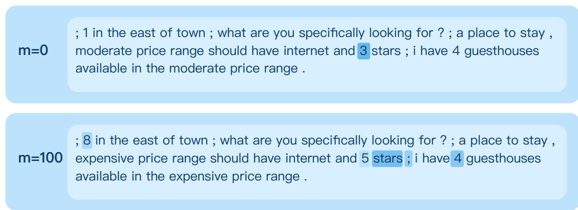

Figure 5: In the dialogue PMUL2513.json, the distribution of attention changes in *hotel-stars* when m differs.

with the increase of SVS ratio in the *train* domain.

It can be seen when the ratio of SVS increases, the model tends to generate from the vocabulary rather than the dialogue. Intuitively, this is one of the reasons for the decline of joint goal accuracy. More unseen values mean larger problem space, existing DST model is insensitive to the unseen test data, which oughts to make the model more inclined to choose the slot value in the dialogue to improve the situation going forward.

To further evaluate the attention matter, we decompose the attention tensor apart from the model structure. Figures 5 shows the distribution of attention in the dialogue when TRADE generates slot values in *hotel-stars*, where darker blue shades indicate larger attention weights. It reveals that the increase of SVS ratio will affect the attention, and then affect the results of generative composition.

4 Conclusions

This paper analyzes reasons leading to performance degradation of generative DSTs across controllable counterfactuals. We propose a simple and efficient counterfactual-maker policy as a principled approach to generate novel scenarios beyond the held-out conversations. We find that performance degradation of DSTs comes from the OOD of counterfactuals and generative composition. These findings are confirmed through experiments on behavior and structure probing, with similar trends. This is of practical interest for applications of DST models, with respect to unlock a true potential of generalization capability.

Acknowledgements

This work is supported by the National Key R&D Program of China under Grant No. 2020AAA0108600.

References

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5016–5026.

Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7521–7528.

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.

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tür. 2019. Dialog state tracking: A neural reading comprehension approach. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pages 264–273.

Rahul Goel, Shachi Paul, and Dilek Hakkani-Tür. 2019. Hyst: A hybrid approach for flexible and accurate dialogue state tracking. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1458–1462.

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, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, pages 35–44.

Yi Huang, Junlan Feng, Min Hu, Xiaoting Wu, Xiaoyu Du, and Shuo Ma. 2020. Meta-reinforced multi-domain state generator for dialogue systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7109–7118.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1*, pages 1875–1885.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025.

Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. SUMBT: slot-utterance matching for universal and scalable belief tracking. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5478–5483.

Shiyang Li, Semih Yavuz, Kazuma Hashimoto, Jia Li, Tong Niu, Nazneen Fatema Rajani, Xifeng Yan, Yingbo Zhou, and Caiming Xiong. 2020. Coco: Controllable counterfactuals for evaluating dialogue state trackers. *CoRR*, abs/2010.12850.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5210–5217.

Jun Quan and Deyi Xiong. 2020. Modeling long context for task-oriented dialogue state generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7119–7124. 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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 808–819.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomás Kociský, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei

Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and evaluating general linguistic intelligence. *CoRR*, abs/1901.11373.

Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wan, Philip S. Yu, Richard Socher, and Caiming Xiong. 2019. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. *CoRR*, abs/1910.03544.