Counterfactual Dialogue Mixing as Data Augmentation for Task-Oriented Dialogue Systems

Sebastian Steindl, Ulrich Schäfer, Bernd Ludwig
Ostbayerische Technische Hochschule Amberg-Weiden, University Regensburg
{s.steindl,u.schaefer}@oth-aw.de, bernd.ludwig@ur.de

Abstract

High-quality training data for Task-Oriented Dialogue (TOD) systems is costly to come by if no corpora are available. One method to extend available data is data augmentation. Yet, the research into and adaptation of data augmentation techniques for TOD systems is limited in comparison with other data modalities. We propose a novel, causally-flavored data augmentation technique called Counterfactual Dialogue Mixing (CDM) that generates realistic synthetic dialogs via counterfactuals to increase the amount of training data. We demonstrate the method on a benchmark dataset and show that a model trained to classify the counterfactuals from the original data fails to do so, which strengthens the claim of creating realistic synthetic dialogs. To evaluate the effectiveness of CDM, we train a current architecture on a benchmark dataset and compare the performance with and without CDM. By doing so, we achieve state-of-the-art on some metrics. We further investigate the external generalizability and a lower resource setting. To evaluate the models, we adopted an interactive evaluation scheme.

Keywords: Task-Oriented Dialog, Data Augmentation, Dialogue Mixing

1. Introduction

Chatbots or *dialogue systems* have undergone extensive research as a tool for human-computer interaction. Depending on the targeted use case, dialogue systems can be roughly separated into two types: *chat-oriented* and *Task Oriented* (Yan et al., 2022). The latter aim to serve as a personal assistant and, through the usage of external services, fulfill tasks, e.g., book a restaurant. Such a TOD system can be interpreted as an user interface to these (possibly multiple) external services. As such, it has to work in three dimensions: *Natural Language Understanding*, *Policy Planning* and *Dialogue Generation* (He et al., 2022). While models for all three dimensions could be constructed and trained separately, end-to-end models incorporating all three aspects are gaining increasingly widespread acceptance (e.g., Peng et al., 2021; Lin et al., 2020; He et al., 2022). This paradigm shift can be attributed to the general prevalence of (Large) Language Models based on the transformer architecture (Vaswani et al., 2017) and the publication of large datasets, e.g., the MultiWOZ-Dataset (Budzianowski et al., 2018), that enable the usage of Deep Learning methods.

In this work, we investigate a data augmentation technique to increase the amount of training data and thus improve model performance for applications in which only a small amount of data is available. The method can be seen as a synthetic oversampling algorithm. We mount the method in the research field of Causality to motivate the generation of realistic, synthetic dialogs.

To study the effect of the proposed data augmentation method, we train and compare two models of the MTTOD (Lee, 2021) architecture. We do this both on the full dataset and a lower resource setting. To evaluate the models, we adopt the interactive evaluation framework proposed by Cheng et al. (2022). Based on the low resource setting, we qualitatively analyzed errors that occurred during evaluation. The usage of CDM boosted the metrics by up to 5 points in the normal setting and 15 points in the low resource setting.

Our main contributions are:

1. We introduce CDM, a new methodology to systematically generate synthetic conversational data in order to mitigate the effort of constructing representative corpora.

2. We compare a model trained with this extended data to the baseline data in an interactive evaluation setting on the MultiWOZ dataset.

3. We simulate a lower resource setting and evaluate CDM in this scenario.

4. We show that the model generalizes better to a different dataset from the same domain.

2. Motivation

The gold standard to generate realistic training multi-turn TOD data is the Wizard-of-Oz (WOZ) technique (Kelley, 1984), in which two humans interact with each other, each thinking that they are talking to a machine. Depending on the domains or the goals of the dialogs, this can be done in a crowdsourcing setting without the need for experts, as was the case in the generation of the MultiWOZ dataset (Budzianowski et al., 2018). Still, the need to generate problem-specific data with limited possibilities of using open-source data renders the compilation of datasets and their labeling a time-consuming and expensive task. To the best of our knowledge, none of the current state-of-the-art works on end-to-end dialogue systems for the MultiWOZ dataset use data augmentation. Instead, they, for example, add auxiliary tasks (Lee, 2021), insert subtask-specific prompts (Su et al., 2022), or perform semi-supervised pretraining on a large dataset (He et al., 2022).

We therefore argue in favor of data augmentation to make the most of the collected data, placing our work in the lineage of data-centric AI. Thus, the goal is to generate dialogs that, even though they did not happen, are still just as realistic as those that were actually collected by WOZ. To this end, we adopt the idea of counterfactuals for data augmentation.

3. Related Work

Data Augmentation. In Computer Vision literature and practice, data augmentation is widely used, and multiple methods exist, e.g., geometric transformations, color space transformations and mixing images (Shorten and Khoshgoftaar, 2019). The data augmentation both acts as a regularization method and increases the amount of training data (Lewy and Mańdziuk, 2023).

In the field of Natural Language Processing (NLP), augmentation methods are classified as either paraphrasing-based, noising-based or sampling-based and include, e.g., backtranslation, word-level swapping and rule-based sampling (Li et al., 2022).

Task-Oriented Dialogue systems and Data Augmentation. While classic approaches opted for a modular system to solve the TOD subtasks (Natural Language Understanding, Policy Planning and Dialogue Generation) (Young et al., 2013), most current systems rely on one pretrained language model (LM) to handle all subtasks in an end-to-end fashion (e.g., Lin et al., 2020; Peng et al., 2021; He et al., 2022). Different data augmentation techniques have been used for TOD systems. Xu et al. (2021) increase their training data by leveraging external datasets and formulate this as a type of data augmentation. Gritta et al. (2021) propose to construct graphs that represent the dialogue states as nodes and the transitions between them as edges. Based on the graph, they generate new data by following the most frequent outgoing edge as observed in the original data. Kulhánek et al. (2021) use backtranslation with 10 languages as a data augmentation technique. While they found this to improve their model's performance, they did not consequently outperform other models that did not use backtranslation. So while data augmentation has been investigated for TOD systems in general, to the best of our knowledge, none of the current or recent SOTA models on the MultiWOZ benchmark dataset have adopted it. Moreover, to the best of our knowledge, the mixing of dialogues controlled by the domains of subtasks has not been proposed in literature before.

Interactive Evaluation. Cheng et al. (2022) propose an interactive evaluation scheme for TOD systems. They argue that due to a policy mismatch during traditional evaluation, the tested models might appear weaker than they actually are. This policy mismatch arises because the utterances that are obtained from the dialogue system are evaluated against the annotated data without regard for the dialogue history. Thus, without taking correct but different policies into account. For example, the dialogue system might try to fill the slots in a different order than in the annotated data. They train a user simulator that, given the goals, generates the user utterances and therefore can simulate an interaction between the user and the dialogue system.

Causality and NLP. Interdisciplinary research on Causality and Machine Learning has recently gained increased attention. We use the term Causality as a collective term for research trying to incorporate causal inference and causal thinking into classical statistical work, but focus on the framework proposed by Pearl (2009). A key concept in this framework is the counterfactual. These are hypothetical facts, i.e., situations that would have arisen if some circumstances changed in a specific way.

Regarding the subfield NLP, the integration of Causality and Machine Learning has been comparatively limited (Feder et al., 2022). Feder et al. (2022) identified two main directions. On one hand, causal inference is practiced with text (e.g., Wood-Doughty et al., 2018; Veitch et al., 2020). On the other hand, causal concepts are used to improve NLP models, e.g., focusing on robustness (Wang and Culotta, 2021) or explanations (Feder et al., 2021). With counterfactuals as hypothetical facts, we hope that a causally flavored method will lead to realistic counterfactuals, which in turn improve the model's performance.

4. Method

The term counterfactual refers to a hypothetical situation that would have arisen if (at least) one element of the original situation were consciously changed while all other elements stayed the same. A standard example is the question *"Would my headache have gone away had I not taken medicine?"* Counterfactuals have recently been in the focus of the Explainable AI literature (Guidotti, 2022) as well as in the field of Causality (Pearl, 2009). In the following, we present how a counterfactual is generally generated in the Causality framework by means of an example from Pearl and Mackenzie (2018):

Say we want to investigate the relationship between Salary (S), Education (ED) and Experience (EX). We define ED and EX as having a causal influence on S. *Everything else* that might have influenced S is not measured, i.e., is exogenous, and is represented by the variable U_S. With this, we define the structural equation to calculate the salary as

$$\text{S} = 65000 + 2500 * \text{EX} + 5000 * \text{ED} + U_S. \quad (1)$$

It is called structural because we defined EX and ED as causal parents of S. Using the observations of one individual's S, ED and EX, we can solve for his U_S. U_S now incorporates every factor other than experience and education that led to the concrete salary. We can then perform so-called do-interventions, i.e., enforcing certain values for the endogenous variables. Say for one individual we observe S $= 72000$, EX $= 2$ and ED $= 0$. We calculate $U_S = 2000$.

Now we can ask the question *"What if she/he had higher education?"*, i.e., $do(\text{ED} = 1)$. We calculate the counterfactual

$$\begin{aligned}\tilde{\text{S}} &= 65000 + 2500 * 2 + 5000 * 1 + 2000 \\ &= 77000.\end{aligned} \quad (2)$$

The collection of all structural equations is called the Structural Causal Model (SCM) (Pearl and Mackenzie, 2018). However, in real-world applications, we usually do not have a fully specified SCM. Without equations like (1), generating realistic counterfactuals is notoriously hard, especially for textual data (Feder et al., 2021).

With CDM we take up the concept of counterfactuals and transfer the idea to a causally-flavored data augmentation technique in which the counterfactuals serve as additional training data.

Since we do not have a fully specified SCM, we cannot generate counterfactuals as shown above but have to approach the problem practically while keeping the causality theory in mind. Instead, we implicitly assume the simple SCM

$$\begin{aligned}X &:= f_X(U_X) \\ Y &:= f_Y(X) + U_Y\end{aligned}$$

with X being the user utterances, Y the system responses. Thus, $f_X(U_X)$ is the mechanism by which the user generated the text based on the latent, exogenous variable U_X. We therefore regard the user's state of mind, which in the context of the multi-turn TOD is best approximated by his current goal within the dialog, as the exogenous variable U. As demonstrated by the example above, one crucial step in generating counterfactuals in the framework proposed by (Pearl, 2009) is to fix the exogenous variables U. To perform an equivalent step in CDM, we have to fix U_X, which we referenced to the user's goal, and account for this by only selecting text patches with matching topics. Thus, the idea of performing an intervention on a variable while keeping *everything else* the same is realized by having matching topics when replacing the utterance.

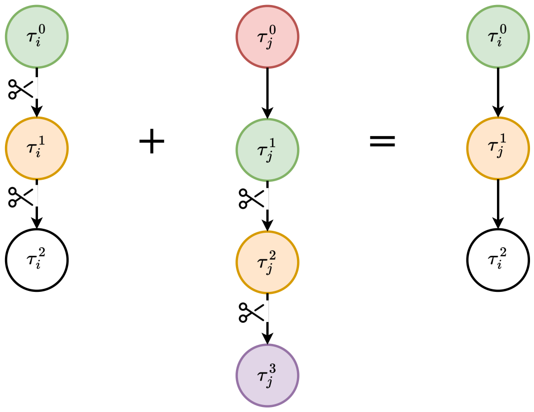


Figure 1: Visualization of the dialogue mixing procedure from the view of conversational graphs. Each node represents one topic and contains multiple utterances. Best viewed in color.

In the MultiWOZ dataset, each dialogue contains multiple domains. We make use of this annotated information to generate synthetic dialogs by mixing two real dialogs from the dataset. That is, we take two dialogs that share a topic change and copy the corresponding utterances (and their annotations) from the second into the first dialog. We regard these synthetic dialogs as counterfactuals, in which the goals of the user stayed the same but the way she/he expresses her/himself, i.e., the utterances, changed.

The method can also be seen from the perspective of conversational graphs. In the work of Gritta et al. (2021), the authors chose to model dialogue states as nodes and the transitions between them as edges in a graph. Our method is the equivalent of constructing a graph with the topics (or subtasks) τ^m as nodes, with their transitions being represented as edges. And the mixing itself would be represented as swapping the node that this transition leads to from D_i into D_j. Fig. 1 visualizes the CDM approach from the perspective of conversational graphs.

4.1. Formalized CDM Approach

Let $D_i = u_i^0, s_i^1, ..., u_i^{n-1}, s_i^n$ be a dialogue in the training data of length n made up of user utterances u_i and system responses s_i, both called *turns*.

Since the MultiWOZ dataset spans multiple domains, each dialogue will usually have multiple domains or topics τ_i (e.g., hotel, taxi). Let $t(D_i)$ be the m element sequence of topics, which we notate as $[\tau_i^0, ..., \tau_i^m] = t(D_i)$. These topics are the equivalent of patches for image mixing. Each τ is confined to a certain sequence of user-system utterance pairs. Topics are allowed to arise multiple times within a conversation.

Analogously, let $D_j = u_j^0, s_j^1, ..., u_j^{l-1}, s_j^l$ be a second dialogue with $[\tau_j^0, ..., \tau_j^k] \in D_j$. The annotation of the dataset contains information about the topic each turn belongs to. Therefore, we know for τ_i the range of utterances in the dialogue D_i in which it is being talked about.

We argue that it is feasible to generate a plausible counterfactual by a form of data mixing, i.e., combining existing data. Thus, for D_i we first randomly choose one topic change (τ_i^c, τ_i^{c+1}). Then we randomly choose one D_j from all D_j, $j \neq i$ that contain the same topic change. Thus, our selection criterion is $(\tau_i^c, \tau_i^{c+1}) = (\tau_j^v, \tau_j^{v+1})$. From the point of view of dialogue graphs, this is analogous to having equal transitions. The new counterfactual dialogue $\widetilde{D}_i$ will be constructed by replacing every turn in τ_i^{c+1} with the turns from τ_j^{v+1}. We see this as an intervention-like action, in which the value for the intervention, i.e., the turns in τ_j^{v+1}, is randomly drawn from the training data. The structural character is taken into consideration by fixing the user's goal.

Let $|\tau|$ be the number of turns contained within a topic. $\widetilde{D}_i$ will then have the length $n - |\tau_i^{c+1}| + |\tau_j^{v+1}|$. The process is visualized in Fig. 2.

While the general "thanks" and "bye" turns can be regarded as topics, we exclude them from the candidates for the mixing operation since these phrases are usually at the end of the dialogue and offer extremely limited variance, thus creating counterfactuals from which little can be learned.

Note that both D_i and D_j must not be part of the develop or test data split to avoid any leakage into the training data.

4.2. Possible Problems due to Mixing

At first glance, one might think that the mixing of dialogs might lead to problems, since the consistency could be impaired. We can divide this concern into two main points: inconsistency regarding i) the dialogue flow and ii) the entities. The integrity of the dialogue flow will, in general, be kept intact. This is the result of a turn comprising only one topic. Moreover, a topic often ends with the dialogue system providing further assistance with phrases along the lines of "is there anything else I can do for you?" and/or the user starts a new topic with, e.g., "I also need to". Speaking in the graph analogy, the node selection process makes sure that the transitions remain sensible after the mixing. If we were to mix single turns within a topic, this would be more difficult to ensure.

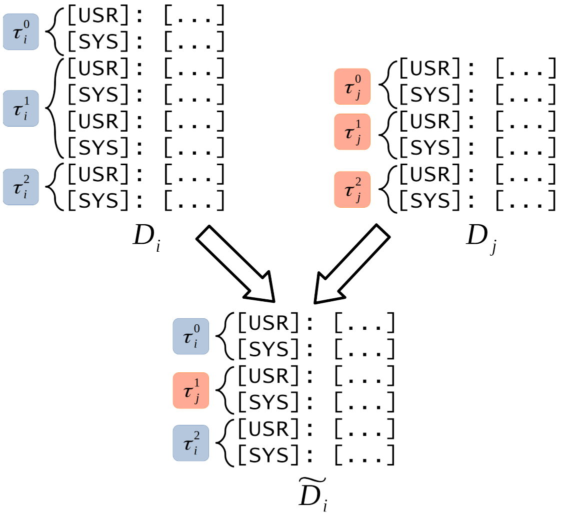
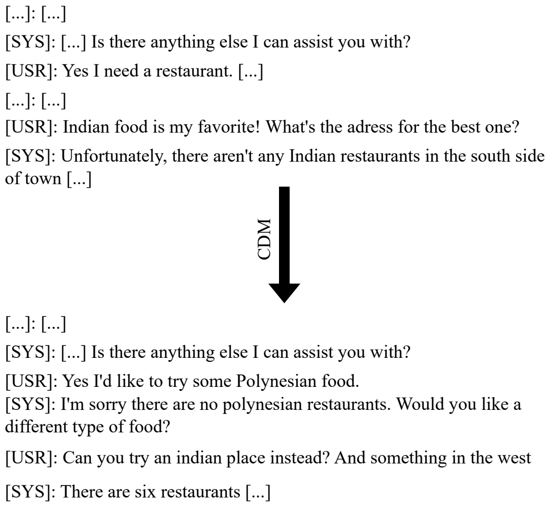


Figure 2: Visualization of the dialogue mixing procedure. Best viewed in color.



[...]: [...]

[SYS]: [...] Is there anything else I can assist you with?

[USR]: Yes I need a restaurant. [...]

[...]: [...]

[USR]: Indian food is my favorite! What's the adress for the best one?

[SYS]: Unfortunately, there aren't any Indian restaurants in the south side of town [...]

[...]: [...]

[SYS]: [...] Is there anything else I can assist you with?

[USR]: Yes I'd like to try some Polynesian food.

[SYS]: I'm sorry there are no polynesian restaurants. Would you like a different type of food?

[USR]: Can you try an indian place instead? And something in the west

[SYS]: There are six restaurants [...]

Figure 3: Example dialogue without CDM (top) and the dialogue generated through CDM (bottom).

Due to the mixing, the entities within a counterfactual conversation will, in general, not be consistent. For example, say the user is looking for an attraction in the northern part of the city. The dialogue system provides the information and offers further service. The user then asks for a restaurant nearby. Say in the counterfactual dialog, the restaurant topic was replaced. Now we have to expect that the mixed-in restaurant will not be close to the attraction from the earlier part of the dialog. An example of this is given in Fig. 3. It is apparent that due to the mixed-in restaurant topic, the location moved from the south to the west side of town. As a side effect, the original dialogue mentions that there are no indian restaurants nearby, while the counterfactual states that there are six of them.

This inconsistency depicted by the example is unproblematic since they are only of concern in the lexicalized dialogs, that contain information on entities. During the training of the dialogue system, analogous to (Cheng et al., 2022), we use the *delexicalized* data that only contains placeholders instead of concrete values (cf. Fig. 4). This would be a common approach for real-world applications as well, since the entities to replace the placeholders would be provided by the external services and need to use real time, e.g., train information.

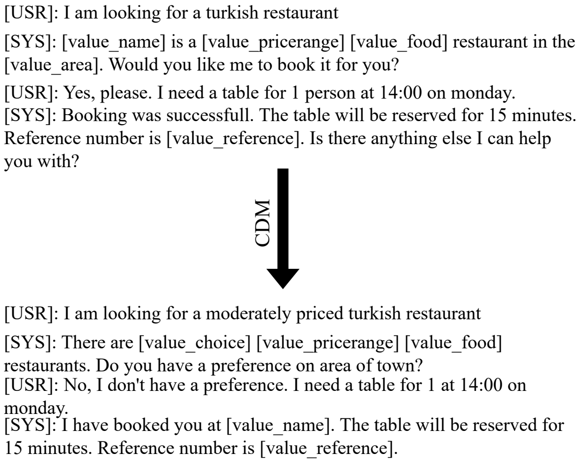
5. Experiments

To evaluate the benefit of the CDM data augmentation, we performed multiple experiments. If not specified otherwise, we use the MultiWOZ data (Budzianowski et al., 2018) with its official train-test split, to be able to compare the results to those reported in (Cheng et al., 2022). Firstly, we investigate if a LM can discern between original dialogues and those created by CDM to determine if the dialogs are realistic. Secondly, we train two MTTOD (Lee, 2021) models in different settings, once on the base dataset and once on the dataset augmented with CDM to evaluate the effect the extended training data has on the test set performance. In another experiment, we simulate a lower resource setting, to investigate the effect of CDM in a scenario that can be seen as more realistic for different applications where fewer data is available. Lastly, we evaluate both models that were trained on the MultiWOZ data on the restaurant domain of the SGD (Rastogi et al., 2020) test data to assess the generalization capability.

We performed one round of CDM, that is, one iteration over all dialogs in the training data, and tried to construct a counterfactual. Due to the sampling process of CDM, where we randomly draw a patch to be replaced, it is possible that for some dialogs D_i there is no matching D_j to generate the counterfactual $\tilde{D}_i$. This procedure increased the training data by roughly 76%.

5.1. Classification of counterfactual dialogs

The first experiment evaluates if a language model is capable of discerning between the original dialogs and the counterfactuals generated by CDM. To this end, we train a DistilBERT (Sanh et al., 2019) model as a binary classifier that takes as input all the dialogue turns at once, evaluating each dialogue as a whole. We perform a train-val-test split with 70%, 10% and 20% of the data, respectively. We train the model for up to 50 epochs or until convergence, stopping when the validation accuracy does not improve for five epochs. In this scenario, the model did not improve after the first epoch. The model reached an accuracy of 55.4% on the test set. This is not better than chance, since due to the sampling method, there is a slight imbalance between the two classes. The real dialogs make up for 56.7% of the whole dataset.


[USR]: I am looking for a turkish restaurant

[SYS]: [value_name] is a [value_pricerange] [value_food] restaurant in the [value_area]. Would you like me to book it for you?

[USR]: Yes, please. I need a table for 1 person at 14:00 on monday.

[SYS]: Booking was successfull. The table will be reserved for 15 minutes. Reference number is [value_reference]. Is there anything else I can help you with?

[USR]: I am looking for a moderately priced turkish restaurant

[SYS]: There are [value_choice] [value_pricerange] [value_food] restaurants. Do you have a preference on area of town?

[USR]: No, I don't have a preference. I need a table for 1 at 14:00 on monday.

[SYS]: I have booked you at [value_name]. The table will be reserved for 15 minutes. Reference number is [value_reference].

Figure 4: Example of a delexicalized dialogue without CDM (top) and the delexicalized dialogue generated through CDM (bottom).

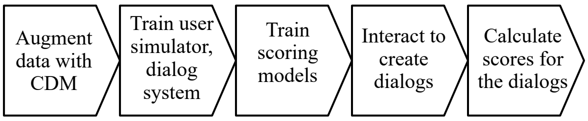

Augment data with CDM
Train user simulator, dialog system
Train scoring models
Interact to create dialogs
Calculate scores for the dialogs

Figure 5: Overview of the training and evaluation process with its different components and steps.

Seeing as the model is not better than chance at the binary classification task, we argue that the generated counterfactuals are realistic enough to aid in model training by increasing the amount of available data.

5.2. Evaluation on MultiWOZ

We use the official code published by Cheng et al. (2022) to train a T5-based (Raffel et al., 2020) dialogue system in multiple settings. In a first step, a user simulator model is trained that, based on the predefined goal state of the dialog, generates user utterances. This sequence-to-sequence model is trained by identifying, based on the dialogue history, which goals have been achieved already and which are not yet finished. Once all goals in the predefined state are achieved, the dialogue is terminated.

To further improve the simulated dialogue interaction, Cheng et al. (2022) incorporate Reinforcement Learning where the generation of a token is interpreted as an agent action. Thus, we also adopt this additional training step. Depending on the configuration, different scores are used as the reward. We optimize the user simulator jointly for the generation of the user action and user utterance, and the dialogue system is jointly optimized for the generation of the belief state, action and utterance. Since Cheng et al. (2022) reported possible instability due to unfortunate seeds, note that we use their default seed of "1998".

The sequence of actions to train the models, generate dialogs and evaluate them is depicted in Figure 5. The components consist of the user simulator and the dialogue system that will perform the interactions, and the sentence and session score models that will score them.

The evaluation of the models follows the interactive setting (Cheng et al., 2022) to avoid a potential policy mismatch distorting the evaluation results. The metrics used are the standard inform and success scores. Instead of the BLEU score, which cannot be used in the interactive evaluation, we report the sentence and session scores as proposed by Cheng et al. (2022).

The sentence score measures the quality of the language generation for a single sentence and is defined as

$$Sent = -\sum_{i=1}^{L} \frac{1}{L} \log p(y_i | y_{<i}, \theta), \qquad (3)$$

where y_i is the i-th token and $y_{<i}$ are the previous tokens in a sequence of length L generated by a fine-tuned GPT-2 (Radford et al., 2019).

The session score measures the coherence of the whole conversation. To this end, a BERT-base model is trained as a binary classifier by randomly sampling system responses to create negative utterance pairs. The session score is then calculated as the average softmaxed confidence over all utterance pairs, both starting with a user utterance and with a system utterance.

Metric	Base	Base+CDM
Sentence Score	1.44	**1.43**
Session Score	0.89	**0.92**

Table 1: Evaluation of the sentence score and session score Model. For the sentence score lower is better and for session score higher is better.

The evaluation of the sentence and session score model are listed in Table 1. The models trained with CDM outperform the standard models in both cases by a small margin.

The performance of the dialogue systems as measured by inform, success, sentence and session scores is shown in Table 2. Analogously to Cheng et al. (2022), we evaluate different settings of reinforcement learning. wRL-Succ only uses the success as a reward, RL-Sent uses sucess and sentence score and RL-Sess uses success and session score. The model trained with CDM systematically outperforms the model trained on the standard data on three out of the four metrics while only producing worse sentence scores. Moreover, to the best of our knowledge, the inform and success of 99.1 achieved by the RL-Succ model improve the state of the art.

Fig. 4 shows examples for dialogs created by the model without and with CDM, respectively. Notice that both models were evaluated on the same dialogue from the test set (SNG01608) but differ in the user utterances due to the interactive evaluation scheme.

5.3. Lower Resource Evaluation

Furthermore, we investigate a lower resource (LR) setting by randomly sampling 20% of the training dialogs and using them to either train directly or to perform one round of CDM before training. We argue that this is a more realistic scenario for real-world applications, where the collected datasets might not be as large as MultiWOZ.

The results in Table 3 show that in this scenario, increasing the amount of training data via CDM significantly increases the model's performance, especially in its ability to complete the task. Interestingly, while the experiments on the base dataset showed that CDM led to systematically better session scores but worse sentence scores, this behavior is flipped in the LR scenario.

5.4. Qualitative Analysis of Errors

Since the models, even without CDM, achieve a high level of task completion in the base scenario, we focus on the LR setting to analyze the errors qualitatively and better understand where the models fail. We found that while the architecture with a user simulator and interactive evaluation alleviates the problem of underestimating performance due to policy mismatches, it also introduces a new possible source of error.

That is, the user simulator might fail to produce user utterances that are a sufficient input for the dialogue system. The user simulator tries to generate utterances that correspond to the dialog's predefined goals. However, if the user simulator's performance is not sufficient, this will generate a mismatch of its own: The conversation might read as if all goals were achieved based on the utterances, despite not fulfilling all predefined goals. Analogously to the term "policy mismatch" we name this "goal mismatch". This might arise, e.g., if the user simulator does not request all needed information and the dialogue system consequently does not provide all of it. Thus, during evaluation, errors

	Inform		Success		Sentence		Session	
Model	Base	Base+CDM	Base	Base+CDM	Base	Base+CDM	Base	Base+CDM
RL-Succ	95.9	**99.1**	93.9	**99.1**	**0.799**	0.812	0.876	**0.957**
RL-Sent	94.6	**98.6**	89.4	**97.6**	**0.746**	0.834	0.953	**0.959**
RL-Sess	95.6	**96.7**	90.3	**96.6**	**0.73**	0.799	0.957	**0.962**

Table 2: Evaluation of the MTTOD model trained either on the base dataset, or the extended dataset with CDM.

Metric	LR	LR+CDM
Inform	73.5	**88.1**
Success	69.5	**77.9**
Sentence Score	1.01	**0.90**
Session Score	**0.92**	0.82

Table 3: Result of a RL-Succ model on the test data, trained in the lower resource setting.

Metric	Base	Base+CDM
Inform	**17.93**	**17.93**
Success	**17.93**	**17.93**
Sentence	1.19	**1.16**
Session	0.70	**0.84**

Table 4: Result of the external evaluation with the RL-Succ model.

that are due to the user simulator will impair the metrics that should measure the performance of the dialogue system.

We found that the interaction of the user simulator and the RL-Succ model trained on the LR data without CDM produced multiple dialogs that are deemed errors due to the lack of sufficiency of user utterances. For example, the user simulator produced the same question "What is the address?" for five consecutive turns, even though the dialogue system correctly answered each time. Moreover, we found multiple instances where the user simulator was not specific enough, i.e., it did not fill all slots the predefined goal state dictates.

The application of CDM reduced these errors. For instance, the model without CDM produced the utterance "I am also looking for a place to go" when the predefined goal state demanded information on an attraction of the type of entertainment. The dialogue system correctly interprets the utterance as a request for an attraction of whatever type, and not of the attraction type. On the other hand, the model trained with CDM is more specific and directly demands an entertainment attraction when asking for the same test data dialog: "I am also looking for an entertainment attraction."

We conclude from this error analysis that a significant part of the increased performance is due to the improvements to the user simulator that benefited from the additional training data.

5.5. External Evaluation

One of the goals of data augmentation in general is to improve the generalization of the model by increasing the size of the training dataset (Lewy and Mańdziuk, 2023). Therefore, we perform an external evaluation on dialogues that are in the same domain but from a different dataset. Multiple task-oriented dialogue datasets have been proposed in the literature, e.g., CamRest676 (Wen et al., 2016), SGD (Rastogi et al., 2020), and MultiWOZ (Budzianowski et al., 2018). However, it remains hard to use multiple datasets combined, since they differ in, e.g., their domains, labels and usage of databases. Harmonizing these datasets has been studied in the literature, but concessions have to be made (Hudeček et al., 2022).

To perform the external evaluation, we use the dialogs in the restaurant domain of the schema-guided SGD test data, based on the DIASER (Hudeček et al., 2022) unified presentation, in which a database usage that was not part of the original dataset has been emulated. We transform the data to fit the preprocessing of the previously used system and add the entries to the MultiWOZ database. Without further finetuning, we evaluate the RL-Succ model on this test data.

The results in Table 4 show that while both models fail in generalizing on achieving the task (low inform and success), the model trained with CDM produces more coherent language (better sentence and session score). That is, its language generalization abilities were improved.

6. Conclusion

We propose CDM, a new method for textual data augmentation that increases the amount of training data by strategically mixing two dialogues. The comparison of a model trained with CDM to one without it showed promising results. This is true as well when using the whole dataset, as in a LR setting, where the improvement was even clearer. One of the models trained on the CDM enhanced data improves the state of the art on the inform and success metrics for the MultiWOZ dataset. During the external evaluation, the model with CDM showed better language generation capabilities than that without. Still, the generalization of task performance remains unsatisfying for both models.

Cheng et al. (2022) suggest that MultiWOZ is a solved dataset since their interactive evaluation reveals high inform and success values. However, data augmentation with CDM to increase the training data still lead to small improvements when using all available data and significant improvements in the lower resource setting. Moreover, the experiments revealed that while the interactive evaluation scheme alleviates policy mismatches that distort the evaluation results, it also introduces the risk of a goal mismatch. This can be due to an insufficient user simulator and will lead to the same problem the interactive evaluation set out to solve: concealing the true performance of the dialogue system.

While Cheng et al. (2022) argue that we need more complex datasets, we think that a more promising direction would be to start with a unified data representation (e.g., Hudeček et al., 2022) of multiple data sources and try to improve external generalization. CDM can be studied in this setting in future work by using external data as the basis for the mixed in dialogs. Moreover, an accurate but automatic evaluation of the dialogue systems remains a challenging task, as shown by the newly identified possible goal mismatch.

Ethics Statement

Current research into TOD systems is decisively enabled by the availability of large datasets. However, especially for low resource languages, this assumption will not be met in general when trying to develop real world applications. The proposed data augmentation method can reduce the amount of required training data, making TOD systems more achievable with lower resources.

Moreover, a common way to collect data is via paid crowdsourcing. If, during this process, one does not ensure that the workers get paid at least the minimum wage for their on-demand task solving, the data collection has ethical problems. Thus, reducing the necessity for this controversial process through data augmentation can be seen as a positive aspect.

Nevertheless, as is true for any technological advancement, we cannot keep bad actors from using it. Possible malicious use cases include using the data augmentation method to create chatbots that are part of fraud schemes or spread fake news.

7. Bibliographical References

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. 2022. On Pearl's Hierarchy and the Foundations of Causal Inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, first edition, pages 507–556. Association for Computing Machinery, New York, NY, USA.

Qinyuan Cheng, Linyang Li, Guofeng Quan, Feng Gao, Xiaofeng Mou, and Xipeng Qiu. 2022. Is MultiWOZ a solved task? An interactive TOD evaluation framework with user simulator. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1248–1259, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Milan Gritta, Gerasimos Lampouras, and Ignacio Iacobacci. 2021. Conversation graph: Data augmentation, training, and evaluation for non-deterministic dialogue management. *Transactions of the Association for Computational Linguistics*, 9:36–52.

Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery*.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, and Luo Si. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10749–10757.

J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems*, 2(1):26–41.

Jonáš Kulhánek, Vojtěch Hudeček, Tomáš Nekvinda, and Ondřej Dušek. 2021. AuGPT: Auxiliary tasks and data augmentation for end-to-end dialogue with pre-trained language models. In *Proceedings of the 3rd Workshop on*

Natural Language Processing for Conversational AI, pages 198–210, Online. Association for Computational Linguistics.

Yohan Lee. 2021. Improving end-to-end task-oriented dialog system with a simple auxiliary task. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1296–1303, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dominik Lewy and Jacek Mańdziuk. 2023. An overview of mixing augmentation methods and augmentation strategies. *Artificial Intelligence Review*, 56(3):2111–2169.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405, Online. Association for Computational Linguistics.

Judea Pearl. 2009. *Causality*, second edition. Cambridge University Press.

Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*, 1 edition. Basic Books, New York.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Victor Veitch, Dhanya Sridhar, and David Blei. 2020. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14024–14031.

Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, page 4586.

Yan Xu, Etsuko Ishii, Genta Indra Winata, Zhaojiang Lin, Andrea Madotto, Zihan Liu, Peng Xu, and Pascale Fung. 2021. CAiRE in DialDoc21: Data augmentation for information seeking dialogue system. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 46–51, Online. Association for Computational Linguistics.

Rui Yan, Juntao Li, and Zhou Yu. 2022. Deep Learning for Dialogue Systems: Chit-Chat and Beyond. *Foundations and Trends® in Information Retrieval*, 15(5):417–589.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. POMDP-Based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

8. Language Resource References

Budzianowski, Paweł and Wen, Tsung-Hsien and Tseng, Bo-Hsiang and Casanueva, Iñigo and Ultes, Stefan and Ramadan, Osman and Gašić, Milica. 2018. *MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling*. Association for Computational Linguistics. PID https://github.com/budzianowski/multiwoz.

Hudeček, Vojtěch and Schaub, Leon-paul and Stancl, Daniel and Paroubek, Patrick and Dušek, Ondřej. 2022. *A Unifying View On Task-oriented Dialogue Annotation*. European Language Resources Association. PID https://github.com/ufal/diaser.

Rastogi, Abhinav and Zang, Xiaoxue and Sunkara, Srinivas and Gupta, Raghav and Khaitan, Pranav. 2020. *Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset*. AAAI. PID https://github.com/google-research-datasets/dstc8-schema-guided-dialogue.

Wen, Tsung-Hsien and Gašić, Milica and Mrkšić, Nikola and Rojas-Barahona, Lina M. and Su, Pei-Hao and Ultes, Stefan and Vandyke, David and Young, Steve. 2016. *Conditional Generation and Snapshot Learning in Neural Dialogue Systems*. Association for Computational Linguistics. PID https://www.repository.cam.ac.uk/handle/1810/260970.