# Less is More: Mitigate Spurious Correlations for Open-Domain Dialogue Response Generation Models by Causal Discovery

**Tao Feng** and **Lizhen Qu*** and **Gholamreza Haffari**

Faculty of Information Technology
Monash University, Australia
{tao.feng, lizhen.qu, gholamreza.haffari}@monash.edu

## Abstract

In this paper, we conduct the first study on spurious correlations for open-domain response generation models based on a corpus CGDialog curated by ourselves. The current models indeed suffer from spurious correlations and have a tendency to generate irrelevant and generic responses. Inspired by causal discovery algorithms, we propose a novel model-agnostic method for training and inference using a conditional independence classifier. The classifier is trained by a constrained self-training method, coined ConSTrain, to overcome data sparsity. The experimental results based on both human and automatic evaluation show that our method significantly outperforms the competitive baselines in terms of relevance, informativeness, and fluency.

## 1 Introduction

Open-domain response generation models have achieved impressive empirical success due to the recent advances in large-scale pre-trained transformers (Caldarini et al., 2022). However, although those models can generate fluent responses, it is still difficult for them to deeply understand conversation histories, and produce *coherent* and semantically *diverse* responses, especially when the conversation histories are *long* (Sankar et al., 2019; Qiu et al., 2019). We conjecture that one of the key reasons is *spuriously correlated* utterances in histories, which do not directly result in responses. Although the vulnerability to *spurious correlations* is a well-known problem in deep learning models (Wang et al., 2021), to the best of our knowledge, there is no study on this topic from a causal perspective for response generation models.

To investigate spurious correlations in dialogues, we are concerned with identifying non-spurious ones, which are the *direct causes* of responses. In this work, a direct cause of a response refers to a text or an utterance in a conversation history that directly results in the response. Table 1 shows an example dialogue between a help-seeker and a peer-supporter randomly picked from the Emotion Support Conversation corpus (ESConv) (Liu et al., 2021). The utterance $\boldsymbol{u}_3$ serves as the direct cause of the response $\boldsymbol{u}_6$, because it is the only utterance mentioning online learning. Otherwise, if we remove it from the history or significantly alter its meaning, the response $\boldsymbol{u}_6$ becomes groundless. In contrast, if we remove an utterance non-causally related to a human response, such as $\boldsymbol{u}_1$ or $\boldsymbol{u}_5$ related to $\boldsymbol{u}_6$, the direct causes still provide sufficient and necessary information to the responses.

Causal discovery algorithms provide a theoretically grounded way to learn causal relations between random variables from observational data (Nogueira et al., 2021). Although they can be applied to identify which utterances in conversation histories are direct causes of responses in theory, the research on such methods for natural language processing problems is still in its infancy.

In this work, we conduct the first study on spurious correlations for response generation models from a causal perspective. We empirically show that non-cause utterances, including spurious correlated ones, have *significantly more influence* on response generation models than the direct cause utterances human would rely on.

Inspired by causal discovery algorithms, we propose a *model-agnostic* training and inference method for mitigating spurious correlations in long conversations. The method aims to automatically identify key utterances in histories, which serve as direct causes for response generation. Herein we convert the cause identification problem into a problem of conditional independence (CI) tests. The CI tests are realized by building

*Corresponding author.



Transactions of the Association for Computational Linguistics, vol. 11, pp. 511–530, 2023. https://doi.org/10.1162/tacl_a_00561
Action Editor: Hua Wu. Submission batch: 11/2022; Revision batch: 2/2023; Published 5/2023.

| | | | |
|---|---|---|---|
| History | Supporter: | Hello | $\boldsymbol{u}_0$ |
| | Help seeker: | Hi, how are you? | $\boldsymbol{u}_1$ |
| | Supporter: | Doing good.. How are you? | $\boldsymbol{u}_2$ |
| | Help seeker: | **I'm feeling really anxious these days. I'm finding the COVID online learning experience to be too much for me at this time.** I want to stop school, but I don't think I can afford to. I need to get done with school. | $\boldsymbol{u}_3$ |
| | Supporter: | I understand your frustration. All of us are challenged due to COVID. | $\boldsymbol{u}_4$ |
| | Help seeker: | School was always hard. Now it's gotten harder. I think a lot of people are stressed. | $\boldsymbol{u}_5$ |
| Human | Supporter: | **How long are you doing the online school?** | $\boldsymbol{u}_6$ |
| BLENDERBOT | Supporter: | You are welcome. I wish you all the best in your future endeavors. Take care. | $\boldsymbol{u}_7$ |

Table 1: An emotion support dialogue annotated with direct causes of human response (in **bold**).

a classifier to infer whether an utterance in the history statistically depends on the response conditioned on its preceding utterance. As there is no training data for such a classifier, we start with manually annotating causal relations on a small portion of public open-domain dialogues. To overcome the scarcity of the training data, we propose a Constrained Self-Training method, coined CONSTRAIN, which is able to identify causal relations with high precision and recall. This classifier is applied to filter out utterances in histories, which are not direct causes of responses, before training response generation models. Furthermore, the classifier serves as a scoring function to select the most relevant response from all generated candidates.

To sum up, our contributions are as follows:

- We conduct the first empirical study on spurious correlations for dialogue response generation models. To investigate this problem in depth, we curate a corpus CGDIALOG by annotating causal relations on dialogues.
- We reduce the direct cause identification problem to a problem of CI tests and propose a constrained self-training method, coined CONSTRAIN, to train the corresponding classifier.
- We propose to train response generation models by taking only direct causes as inputs and perform inference using the CI classifier.
- The extensive human evaluation results show that the response generation models, such as BLENDERBOT, using our method outperform the baselines in terms of relevance, informativeness, and fluency.[1]

[1]Our dataset, models, and code can be found at https://github.com/WilliamsToTo/CGDIALOG.

## 2 Causal Discovery Background

Given a set of random variables, causal discovery from observational data is concerned with discovering causal relations between the random variables. A set of causal relations constitutes a causal graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where a node $v \in \mathcal{V}$ denotes a random variable and a directed edge $v_i \rightarrow v_j \in \mathcal{E}$ indicates that $v_i$ is a *direct cause* of $v_j$ (Neal, 2020). A change in $v_i$ results in a change in $v_j$, but an intervention in $v_j$ does not necessarily lead to a change in $v_i$.

Our work is motivated by constraint-based causal discovery approaches (Nogueira et al., 2021), which iteratively apply independence and CI tests to infer causal structures. Those approaches make the *faithfulness* assumption that independencies in a distribution imply the structure of the corresponding causal graph. The most commonly used algorithm in this family is the PC algorithm (Spirtes et al., 2000). It starts with adding undirected edges between two nodes if both of them are dependent by not passing independence tests. Then it remove an edge between two nodes if they are identified as conditionally independent after running CI tests. The algorithm would continue with larger conditioning sets until the skeleton of the graph is identified. Finally, it orients the edges when possible by using heuristics and identifying the specific structure, $v_i \rightarrow v_k \leftarrow v_j$, referred to as immorality, as illustrated in Figure 2b (Neal, 2020).

In this work, we do not need to recover the complete causal structure between utterances in dialogues. Instead, we only focus on identifying direct causes of responses, namely, the parents of the response nodes in a causal graph. A causal graph satisfies the *Causal Markov Condition*, which states that *each variable is independent of all its non-descendants, given its parents in the causal graph*. Hence the value of a response variable is only determined by its parents

(Pearl and Verma, 1991; Pearl, 2009). Under the faithfulness assumption, if a response variable $v_j$ is dependent on $v_i$ conditioning on arbitrary any other nodes, and we know the influence direction is from $v_i$ to $v_j$, then we conclude that $v_i \rightarrow v_j$.

## 3 Spurious Correlations in Dialogues

The slogan ''Spurious correlation is no proof of causation'' is well known in statistics (Simon, 1954). A correlation between a response and an utterance in a conversational history is spurious if it does not directly result in the response.

Spurious correlations are an inherent problem of statistical machine learning (ML) models. Wang et al. (2021) point out that ML models relying on core features may well achieve similar training errors on the same training data as those relying on spurious features. However, the models relying on spurious correlations lead to high test errors because spurious correlations are inconsistent across datasets. Overparameterization further exacerbates spurious correlations by memorizing examples containing spurious features (Sagawa et al., 2020). Unfortunately, almost all the SotA open-domain dialogue models are based on large-scale transformers, which are overparameterized with respect to small dialogue training datasets in target domains.

To study the impact of spurious correlations for dialogue models, we leverage two public dialogue corpora (ESConv and MSC) to construct a small evaluation corpus for Causal Graphs in dialogues, coined CGDialog, and evaluate two SotA dialogue models, BlenderBot (Roller et al., 2021) and DialogGPT (Zhang et al., 2020), on that corpus in terms of spurious correlations.

### 3.1 Annotation of Causal Graphs

We randomly sampled 80 dialogues from ESConv (Liu et al., 2021) and MSC (Xu et al., 2022) each, then employed four graduate computer science students and four well-trained crowd-workers to annotate direct causes of responses. All annotators were instructed to have a good understanding about what are direct causes of responses and used Amazon Mechanical Turk (AMT) for annotation. We trained them by letting them first annotate on a dry-run dataset, and provided feedback if there was a misunderstanding. After training, annotators were asked to read the provided responses and their conversation histories, then highlight which utterances or clauses serve as direct causes of the responses. We include clause level annotations because sometimes only one clause in a long utterance is the direct cause of a response. For quality check, a human expert having a good grasp of this task reviewed all annotations and corrected mistakes. CGDialog-ESConv is splitted into a training set, a validation set, and a test set, containing 272, 211, and 211 context-response pairs, respectively, while CGDialog-MSC contains 300, 250, and 250 context-response pairs, respectively.

| Number of items | ESConv | MSC | Total |
|---|---|---|---|
| Dialogues | 80 | 80 | 160 |
| History-response pairs | 694 | 800 | 922 |
| Utterances | 2301 | 3807 | 6108 |
| Direct causes utterance | 1347 | 1525 | 2872 |
| Average token length | 24.01 | 22.22 | 23.05 |
| of direct causes | ($\sigma = 16.61$) | ($\sigma = 13.79$) | ($\sigma = 15.20$) |
| The proportion of direct causes | 0.86 | 0.72 | 0.79 |
| in original utterances | ($\sigma = 0.22$) | ($\sigma = 0.27$) | ($\sigma = 0.26$) |

Table 2: Statistics of the CGDialog.

We measured the inter-annotator agreement between the expert and an annotator at both the utterance level and the clause level. At the utterance level, we computed Cohen's Kappa and obtained 0.8149. At the clause level, because marked text boundaries may vary between annotators, we compute the averaged F1 score for all possible pairs of annotators, as detailed in Rajpurkar et al. (2016) and Poria et al. (2021). We obtained a F1 score of 0.8449, which indicates a high-level of inter-annotator agreement.

We show the corpus statistics in Table 2 and Figure 1. Most of the preceding utterances of responses are annotated as direct causes, which are over $80\%$ and $95\%$ on ESConv and MSC, respectively. The proximity of utterances to responses matters: The closer utterances are to the responses, the higher the chance to be direct causes.

### 3.2 Analysis of Spurious Correlations

We conduct experiments to investigate the impact of spurious correlations on two SotA response generation models: BlenderBot and DialogGPT. Both models are fine-tuned on the training sets of ESConv and MSC by taking full conversation histories as inputs. Inspired by Sankar et al. (2019), we perturb conversation histories by removing either direct causes or non-causes from histories. We hope that the outputs of a robust model should have little changes if only spuriously correlated utterances are removed. The removal is conducted in two ways: i) replacing each removed token with

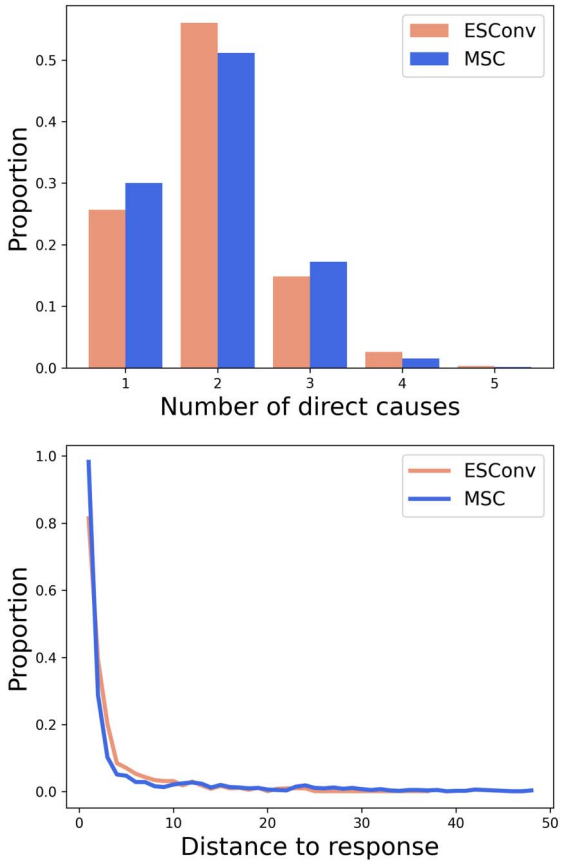


Figure 1: **Top:** The ratio between the number of the history-response pairs with a particular number of direct causes and all history-response pairs. **Bottom:** Proximity between direct causes and responses, measured by the percentage of such pairs in all history-response pairs. Most direct causes are very close to responses.

the pad token <pad>; and ii) directly dropping the removed tokens. We apply such perturbations to the test set of CGDIALOG and compare their results with the ones without any perturbations.

If a response model captures the same genuine correlations between key utterances in histories and responses as humans, the perplexities of human responses estimated by the model should change only slightly if non-cause utterances are excluded from conversation histories. However, as shown in Table 3, the increase of perplexities caused by dropping or replacing non-cause utterances is significantly sharper than that resulted by the removal of cause utterances.

To further investigate the effects of perturbing conversation histories, we apply the same decoding method of both models to the histories after perturbations. We compare the responses generated before and after perturbations in terms of BLEU. Lower BLEU indicates larger changes of generated outputs. As we can see, dropping or replacing direct causes leads to notably smaller changes of outputs than applying the same operations to non-cause utterances.

To eliminate the concern that the above observations are caused by the number of perturbed utterances, we remove or replace the same number of non-cause utterances as that of direct causes each time. More specifically, as the number of direct causes is always smaller than that of non-causes, we apply the perturbations to $k$ utterances randomly chosen from non-cause utterances if the number of direct causes is $k$, and compute the corresponding perplexities and BLEU. To mitigate the influence of randomness, we repeat each experiment for five times and compute statistical significance based on two-sample t-test (Dror et al. 2020). As one can see from Table 3, both generative models are sensitive to the removal of utterances that are weakly associated with human responses. The perturbations on the equal number of non-cause utterances lead to larger changes of the model outputs than those on causes, as indicated by BLEU. For DIALOGGPT, the increase of perplexities by perturbing non-causes is still significantly higher than that by perturbing causes. Therefore, both models do not really learn on the utterances that humans use as causes to articulate responses, but rely heavily on non-cause utterances.

## 4 Causal Discovery Motivated Training and Inference

As shown by our empirical study, spurious correlations are detrimental to the SotA dialogue models. To remedy this, we propose to automatically identify the utterances in conversation histories, which serve as direct causes to responses, and only use them as history representations during both training and inference. Based on the theoretical analysis in Section 2, this identification problem is reduced to running CI tests between responses and utterances in their history. Herein, we propose a constrained self-training procedure to build a classifier for classifier-based CI tests (Lopez-Paz and Oquab 2017; Sen et al., 2017, 2018; Bellot and Schaar 2019).

Formally, given a conversation history $\mathbf{C}_t = \{\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{t-1}\}$ at time $t$, a dialogue model aims to produce a word sequence $\boldsymbol{r}_t$ as the response based on $\mathbf{C}_t$. Both $\boldsymbol{u}_i$ and $\boldsymbol{r}_t$ are regarded as collections

| Datasets | Models | No Perturbations | Replace non-causes with <pad> | Replace non-causes with <pad> randomly | Replace causes with <pad> | Drop non-causes | Drop non-causes randomly | Drop causes |
|---|---|---|---|---|---|---|---|---|
| PPL↓ | | | | | | | | |
| ESConv | Blenderbot | 12.16 | $25.00^{*}$ | 12.10 | 12.81 | $22.65^{*}$ | 13.13 | 12.35 |
| | DialoGPT | 400.15 | $588.16^{*}$ | $569.60^{\dagger}$ | 514.09 | $474.42^{*}$ | $469.51^{\dagger}$ | 452.91 |
| MSC | Blenderbot | 48.29 | $57.52^{*}$ | 47.52 | 49.65 | $58.53^{*}$ | 49.69 | 48.82 |
| | DialoGPT | 404.08 | $875.15^{*}$ | $703.61^{\dagger}$ | 613.95 | $590.28^{*}$ | $575.12^{\dagger}$ | 480.95 |
| Average BLEU↑ | | | | | | | | |
| ESConv | Blenderbot | – | $0.11^{*}$ | $0.56^{\dagger}$ | 0.82 | $0.15^{*}$ | $0.48^{\dagger}$ | 0.86 |
| | DialoGPT | – | $0.08^{*}$ | $0.48^{\dagger}$ | 0.56 | $0.11^{*}$ | $0.35^{\dagger}$ | 0.81 |
| MSC | Blenderbot | – | $0.14^{*}$ | $0.47^{\dagger}$ | 0.94 | $0.09^{*}$ | $0.39^{\dagger}$ | 0.95 |
| | DialoGPT | – | $0.28^{*}$ | $0.49^{\dagger}$ | 0.81 | $0.37^{*}$ | $0.48^{\dagger}$ | 0.82 |

Table 3: Performance comparison with respect to conversation history perturbations. PPL indicates perplexity of human responses. Average BLEU scores are computed as the mean over the four orders of the n-grams. Because responses generated in ''No Perturbations'' setting are treated as references, average BLEU scores are empty in the ''No Perturbations'' column. $*$ indicates a significant difference between ''Replace (or Drop) causes'' and ''Replace (or Drop) non-causes'', while † represents a significant difference between ''Replace (or Drop) causes'' and ''Replace (or Drop) non-causes randomly''. The significant difference is computed by two sample t-test with $p \leqslant 0.05$.

of random variable, where each variable in the collection denotes if a single word is present or not. Because the same event can be expressed in various linguistic forms, we assume there is a projection function $g(\boldsymbol{u})$, which maps an utterance to a latent random variable vector $\boldsymbol{z} \in \mathcal{Z}$ denoting the meaning of the corresponding event.

A causal graph in the semantic space is a directed acyclic graph $\mathcal{G} = \{\boldsymbol{V}, \boldsymbol{E}\}$, where a node represents a latent random variable vector $\boldsymbol{z}_i$ and an edge is denoted by a causal relation between a pair of nodes. We do not define causal graphs in the word space because i) it is the meanings of utterances that are causally correlated and ii) the same words in different contexts may be involved in different causal relations. Identifying direct causes of responses can thus be regarded as recognizing causal relations between those latent random variables. To simplify notation, we denote the output of $g(\boldsymbol{u}_i)$ by $\boldsymbol{z}_i$, unless stated otherwise.

### 4.1 From Cause Identification to the Conditional Independence Tests

If a latent semantic vector $\boldsymbol{z}_i$ of an utterance is a direct cause of the meaning of a response $\boldsymbol{z}_j$, then $\boldsymbol{z}_i \not\perp\!\!\!\perp \boldsymbol{z}_j | \mathcal{Z}_{t,-i}$, where $\mathcal{Z}_{t,-i}$ denotes any subset of latent random variables derived from the history $\mathbf{C}_t$ excluding $\boldsymbol{z}_i$. In other words, $\boldsymbol{z}_i$ provides additional useful information for $\boldsymbol{z}_j$ given any other utterances in a history. However, it is computationally expensive to consider all possible subsets of a conversation history for running CI tests for a single utterance.

To address the computational challenge, we observe that a response often only depends on the preceding utterance and at most two utterances in total. As evident in Figure 1, $81\%$ of the responses in CGDialog have one or two direct causes and $90\%$ of the preceding utterances serve as direct causes of the following responses. Therefore, we can sharply reduce the computational overhead by making the following assumptions.

**Assumption 1.** *For each response $\boldsymbol{r}_t$, $g(\boldsymbol{u}_{t-1}) \rightarrow g(\boldsymbol{r}_t)$ always holds.*

**Assumption 2.** *There are at most two direct causes for the latent random variable vector of a response.*

**Assumption 3.** *If there is an edge between $g(\boldsymbol{u}_i)$ and $g(\boldsymbol{u}_j)$ in a causal graph and $i < j$, then $g(\boldsymbol{u}_i) \rightarrow g(\boldsymbol{u}_j)$.*

The last assumption articulates the fact that what people said in the past influences what people will say in the future. If the temporal order in a conversation is known, there is no need to apply statistical methods to infer the orientation.

Under the above assumptions, for a given response $\boldsymbol{r}_t$, there are only four possible neighborhood structures, as illustrated in Figure 2. We have $\boldsymbol{z}_t \not\perp\!\!\!\perp \boldsymbol{z}_j | \boldsymbol{z}_{t-1}$ for Figure 2a and Figure 2b, but $\boldsymbol{z}_t$ is conditionally independent of $\boldsymbol{z}_j$ in the remaining cases. Herein, we make the faithfulness assumption that CIs imply graph structures. Under our assumptions, it is sufficient to determine if an utterance $\boldsymbol{u}_j$ with $j < t$ is a cause of $\boldsymbol{t}_t$ by checking whether $\boldsymbol{z}_t \not\perp\!\!\!\perp \boldsymbol{z}_j | \boldsymbol{z}_{t-1}$. Hence, we only need to run $t-2$ CI tests for a response $\boldsymbol{r}_t$. Note

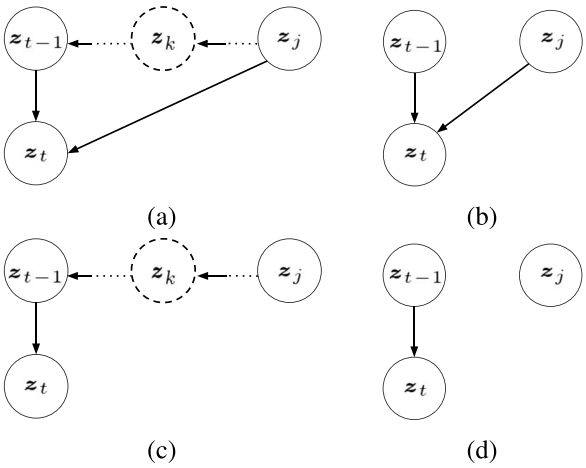


Figure 2: In a, the response variable has two direct causes that may be connected through $z_k$ ($k > j$) or directly connected, while the response variable in b has two disconnected cause variables. In c and d there is only one direct cause $z_{t-1}$ linking to $z_t$.

that it is *important* to run CI tests instead of dependence tests to find a direct cause of a response. As illustrated in Figure 2c, although $z_j$ is not a direct cause of $z_t$, both of them are still dependent through $z_k$ and $z_{t-1}$ according to dependence tests. If we run a CI test conditioned on $z_{t-1}$, the path through $z_k$ is blocked so that the test result reveals $z_t \perp\!\!\!\perp z_j | z_{t-1}$. More details of identifying independence structures in a graphical model can be found in Neal (2020) and Pearl (2009).

### 4.2 Conditional Independence Tests

To perform CI tests over a set of latent random variables $z$ on observational data, we need to i) project utterances to the latent space, and ii) choose a scalable test method which can work with texts. However, the first step is already challenging because the latent random variables are unknown and we even do not know the number of them for an arbitrary dialogue corpus.

To tackle both challenges, we opt for the classifier-based CI test. As $z_t \perp\!\!\!\perp z_j | z_{t-1}$ implies $p(z_t, z_j | z_{t-1}) = p(z_t | z_{t-1}) p(z_j | z_{t-1})$, this family of tests builds a classifier to determine if a sample of data is drawn from $p(z_t | z_{t-1}) p(z_j | z_{t-1})$ or $p(z_t | z_j, z_{t-1}) p(z_j | z_{t-1})$. To train the classifier, we label a tuple $(z_t, z_{t-1}, z_j)$ with $l = 1$ if it is drawn from $p(z_t | z_j, z_{t-1}) p(z_j | z_{t-1})$, otherwise $l = 0$. Then the classifier aims to capture the conditional distribution $p(l | z_t, z_{t-1}, z_j)$.

The recent advances of deep learning show that hidden representations of deep neural networks can well capture meanings of input texts (Yang et al., 2020). Hence, it is straightforward to consider a deep encoder as a function $g(u)$ from an utterance $u$ to a hidden representation $z$. Specifically, we employ a pre-trained RoBERTa (Liu et al., 2019) as the encoder to map a tuple $(r_t, u_{t-1}, u_j)$ to a sequence of hidden representations $(z_t, z_{t-1}, z_j)$, where adjacent utterances are separated by the special token </s>. Taking the representations $(z_t, z_{t-1}, z_j)$ as input, the CI classifier consists of a mean-pooling layer, a linear layer, and a sigmoid layer for characterizing $p(l | z_t, z_{t-1}, z_j)$.

Inspired by Sun et al. (2019), we first train the pre-trained RoBERTa with the masked language model objective on the publicly available Reddit dataset (Baumgartner et al., 2020) to adapt it to dialogues. After training 10 epochs with the learning rate $5 \times 10^{-5}$, we fine-tune the model with our self-training procedure detailed below.

**Incremental Self-training with Constraints.** It is straightforward to collect a small training dataset $\mathbb{D}_L$ from the training set of CGDialog by considering $(u_j, u_{t-1}, r_t)$ annotated with $g(u_j) \rightarrow g(r_t)$ as positive examples and the remaining as negative examples. However, the size of $\mathbb{D}_L$ is small by having only 922 examples in total.

To address the scarcity of $\mathbb{D}_L$, we adapt the self-training procedure introduced in Zou et al. (2019) to train the CI classifier. It starts with training an initial classifier $f_0$ on $\mathbb{D}_L$ in a supervised manner. Then we apply this classifier to unlabeled utterance tuples. The tuples predicted with labels 1 are added to the training set as positive examples if they satisfy the *threshold* and *context* constraints:

i) The probability $p(l = 1 | u_j, u_{t-1}, r_t)$ exceeds a predefined threshold 0.9;

ii) $u_j$ is either $u_{t-2}$ or $u_{t-3}$ with respect to a response $r_t$.

For each response $r_t$, negative examples are collected by randomly sampling $u_j$ from the utterances that are not selected as positive examples. We keep the number of positive examples the same as the number of negative examples in each batch. The extended training set is used to fine-tune the classifier. The process is repeated until the classifier achieves the highest performance on the validation set of CGDialog. More

**Algorithm 1** Incremental Self-training

**Require:**

Labeled training and validation set: $\mathbb{D}_L^{tr}$, $\mathbb{D}_L^{va}$

Unlabeled dataset: $\mathbb{D}_U$

Pseudo-labeled data selection constraint: $C$

Classifier with pre-trained RoBERTa: $f_\theta$

**Ensure:**

$i \leftarrow 0$

$\mathbb{D}^i \leftarrow \mathbb{D}_L^{tr}$

$f_i \leftarrow$ fine-tuning $f_\theta$ on $\mathbb{D}^i$

**while** $f_i$ does not have the best performance on $\mathbb{D}_L^{va}$ **do**

    Apply $f_i$ on unlabeled dataset $\mathbb{D}_U$

    Construct pseudo-labeled dataset $\mathbb{D}_{PL}^i$ with constraint $C$

    $\mathbb{D}^{i+1} \leftarrow \mathbb{D}^i \cup \mathbb{D}_{PL}^i$

    $f_{i+1} \leftarrow$ fine-tuning $f_i$ on $\mathbb{D}^{i+1}$

    $i \leftarrow i + 1$

**end while**

details can be found in Algorithm 1. Note that the main difference to the original self-training algorithm is that we add a positive example to the training set only if $\boldsymbol{u}_j$ is either $\boldsymbol{u}_{t-2}$ or $\boldsymbol{u}_{t-3}$. The constraint is proven to be empirically useful in our experiments.

### 4.3 Training and Inference for Generative Response Models

To overcome spurious correlations, we propose to feed only direct causes of responses to dialogue models during training and inference, where direct causes are selected by the CI classifier. This approach is model-agnostic because it only ''cleans'' the inputs of a response model regardless which neural architecture is used.

The training set of mainstream open-domain dialogue models consists of conversation history and response pairs $\{\mathbf{C}_t, \boldsymbol{r}_t\}_{t=1}^n$. Before training, we preprocess the training set by keeping only direct causes in each conversation history. As $\boldsymbol{u}_{t-1}$ is always one of the direct causes according to Assumption 1, we find another cause by using the CI classifier. In particular, for each conversation history $\mathbf{C}_t$, we perform max inference on all tuples $(\boldsymbol{u}_j, \boldsymbol{u}_{t-1}, \boldsymbol{r}_t)$ using the classifier, where $j \in [0, t-2]$. We select the $\boldsymbol{u}_j$ that has the highest probability $p(l = 1 \mid \boldsymbol{u}_j, \boldsymbol{u}_{t-1}, \boldsymbol{r}_t)$ as another direct cause. Dialogue models are subsequently trained on the preprocessed training set.

The input selection for inference is conducted in a similar manner. In particular, we feed each possible $(\boldsymbol{u}_j, \boldsymbol{u}_{t-1})$ with $j \in [0, t-2]$ to the trained dialogue model to generate a response by beam search. Then we apply the CI classifier to identify the tuple $(\boldsymbol{u}_j, \boldsymbol{u}_{t-1}, \boldsymbol{r}_t)$ with the highest $p(l = 1 \mid \boldsymbol{u}_j, \boldsymbol{u}_{t-1}, \boldsymbol{r}_t)$. To allow selecting responses based on $p(\boldsymbol{r}_t|\boldsymbol{u}_j, \boldsymbol{u}_{t-1})$ or $p(\boldsymbol{r}_t|\boldsymbol{u}_{t-1})$, we choose the response conditioned on $(\boldsymbol{u}_j, \boldsymbol{u}_{t-1})$ if the highest $p(l = 1 \mid \boldsymbol{u}_j, \boldsymbol{u}_{t-1}, \boldsymbol{r}_t)$ exceeds the threshold 0.5, tuned on a validation set, otherwise we take the response conditioned on $\boldsymbol{u}_{t-1}$.

## 5 Experiments

### 5.1 Datasets

We experiment on the following two open-domain dialogue corpora that have long conversation histories. The longer a conversation history is, the more likely utterances in the history are spuriously correlated with responses. In contrast, most open-domain dialogue corpora contain short conversations, in which there are dramatically less spuriously correlated utterances. For example, DailyDialog (Li et al., 2017), WizardOfWikipedia (Dinan et al., 2019), and EmpatheticDialogues (Rashkin et al., 2019) have 7.9 utterances, 9 utterances, and 4.31 utterances per conversation, respectively.

**Emotion Support Conversation (ESConv).** ESConv (Liu et al., 2021) contains conversations between mental health help seekers and supporters, with 29.8 utterances per dialogue on average. In each dialogue, help seekers talk about their problems, such as unemployment, losing a family member, or being infected with COVID. Dialogue response models play the role of supporters to provide supportive responses to help seekers. Each utterance from supporters is annotated with a strategy such as providing suggestions, paraphrasing, or questioning, which are not considered in our models. It is splitted into training, validation, and test sets with the ratios of 80%, 10%, and 10%, respectively.

**Multi-Session Chat (MSC).** MSC (Xu et al., 2022) contains human-human chit-chats over five sessions, each of which contains up to 14 utterances. The average number of utterances per dialogue is 53.3. In each session, two interlocutors conduct a conversation based on given personas. Each persona describes personal information with

multiple sentences. We experiment on its official splits of training, validation, and test sets.

### 5.2 Baseline Models

We compare our method ConSTrain and its variations, based on BlenderBot, with the following generative models:

**BlenderBot.** This transformer-based encoder-decoder model achieves superior performance over the prior models in terms of engagingness and humanness (Roller et al., 2021). We fine-tune the pre-trained model with varying settings of conversational histories. As such, a conversational history contains either: 1) only the preceding utterance $\boldsymbol{u}_{t-1}$, 2) the preceding two utterances $(\boldsymbol{u}_{t-2}, \boldsymbol{u}_{t-1})$ when available, 3) the preceding three utterances $(\boldsymbol{u}_{t-3}, \boldsymbol{u}_{t-2}, \boldsymbol{u}_{t-1})$ when available, 4) the complete conversational history $(\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{t-1})$, or 5) the preceding utterance $\boldsymbol{u}_{t-1}$ and a randomly selected utterance $\boldsymbol{u}_j$ between 0 and $t-2$. All hyperparameters remain the same in different settings.

**DialoFlow.** Li et al. (2021) propose a dialogue system that models dynamic information flow across utterances. The model generates a response based on a distributed representation predicted based on past information flow.

**Retrieval-guided Model.** We implement the retrieval-guided response generation model proposed in Zhong et al. (2022) without using user ids, because they are not available in both corpora. Herein, we first map the tokens in the preceding utterance $\boldsymbol{u}_{t-1}$ and the tokens in the previous history $\{\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{t-2}\}$ into a set of BERT embeddings, respectively. Then we compute a similarity matrix between the two sets of embeddings in terms of dot product. As there is a similarity vector for each token in the previous history, we score each of them by using the highest similarity score in the corresponding vector. We pick the top 30 scored ones as the final set of retrieved tokens. The input to their response generation model is the concatenation of $\boldsymbol{u}_{t-1}$ and the corresponding retrieved tokens.

**ESConv Baseline.** Liu et al. (2021) provide two response models on ESConv. The first one directly fine-tunes the BlenderBot model on ESConv without using annotations of negotiation strategies. Another one fine-tunes BlenderBot by taking as input both negotiation strategies and conversation histories. Both models consider the preceding five utterances as conversation history.

**TransferTransfo.** As MSC can be viewed as an extension of PersonaChat dataset (Zhang et al., 2018), we consider TransferTransfo (Wolf et al., 2019), which reports the SotA performance on PersonaChat. We fine-tune this model on the training set of MSC for a fair comparison.

**Retriever-generator.** Xu et al. (2022) propose a model consisting of a retriever and a generator. The retriever selects relevant utterances from a history, while the generator produces responses conditioned on the utterances selected by the retriever.

Among the above models, BlenderBot, DialoFlow, and retrieval-guided model are evaluated on both corpora. TransferTransfo is evaluated only on MSC because the same model shows inferior performance than the one proposed in Liu et al. (2021) on ESConv. Furthermore, the baseline (Liu et al., 2021) is only evaluated on ESConv because it requires annotations of strategies.

### 5.3 Implementation Details

All the models are implemented with PyTorch (Paszke et al., 2019) and the Transformers library (Wolf et al., 2020). We use the same BlenderBot model[2] in all relevant experiments. All models are trained with Adam (Kingma and Ba, 2015) optimizer with hyperparameters tuned on the validation sets. As a result, we run Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is $2 \times 10^{-5}$ for the CI classifier and $5 \times 10^{-5}$ for the response model. We use a linear learning rate scheduler that dynamically decreases learning rate after a warm-up period. CI classifiers were trained for 10 epochs with the batch size 16 on one NVIDIA RTX 16G V100 GPU; the response models were trained with 5 epochs and a batch size of 8. The beam search width is set to 5 during decoding.

### 5.4 Metrics

**Human Evaluation.** In practice, we had the same observations as in other reports (Belz and Kow 2010; Callison-Burch et al., 2007;

[2] https://huggingface.co/facebook/blenderbot-400M-distill.

Kiritchenko and Mohammad, 2017), that asking crowd-workers to directly score responses on a scale usually receives low-quality evaluation. Thus, following the evaluation design proposed in other work (Novikova et al., 2018; Bojar et al., 2016; Zheng et al., 2021; Zhou et al., 2018; Liu et al., 2021), we opt for pairwise comparison between responses from different sources. In each comparison experiment, we compared our model with a baseline or human responses on a set of 100 conversations randomly sampled from our test set. Given a conversation history, we presented crowd-workers with a pair of responses, one of which is generated by our model and the other is either from humans or a baseline. Five well-trained crowd-workers from AMT are asked to choose the better one in terms of four metrics: **Empathy** (Which response shows better understanding of the partner's feelings?), **Fluency** (Which response has better fluency and readability?), **Relevance** (Which response is more relevant and coherent to the context?), and **Informativeness** (Which response provides more information when both are relevant?). For quality control, we selected only crowd-workers who have an approval rating greater than $90\%$ and a minimum of 10,000 approved tasks. Inter-rater agreement using Krippendorff's $\alpha$ was 0.41. In addition, we presented both good and bad example responses for each metric to educate crowd-workers.

The results of all comparison experiments are summarized by using ranking-based Best-Worst Scaling, a method shown to be more reliable than rating-based Likert scaling in prior studies (Kiritchenko and Mohammad, 2017; Puduppully and Lapata, 2021; Steen and Markert, 2021; Tang et al., 2022; Louviere et al., 2015). For each pair of models in comparison, the score of a model is calculated as the number of times rated best minus the number of times rated worst (Amplayo and Lapata, 2021; Puduppully and Lapata, 2021). Thus, for such a pair of models, their scores have the same absolute value but opposite signs. For example, in a comparison experiment between System A and System B, the score of System A is 13, then that of System B is -13. Thus, we only need to know the score of one system, then obtain the score of the other system automatically. To summarize those results, we put the scores of baselines and human responses in one table, which are compared with our model. As our model is always used as a reference, we set the scores of our model to be zero in that table. Therefore, a negative score in the table means the corresponding system performs worse than our model, while a positive score indicates a better performance of the corresponding system.

**Automatic Evaluation** Although automatic metrics are still not reliable for response evaluation (Liu et al., 2016), to facilitate comparisons with prior works, we consider the four automatic metrics for evaluating the quality of responses: BLEU (Papineni et al., 2002), BERTScore (Zhang* et al., 2020), MAUVE (Pillutla et al., 2021), and METEOR (Banerjee and Lavie, 2005). In addition, we evaluate the diversity of model outputs in terms of Distinct-1/2 (Li et al., 2016).

### 5.5 Experimental Results

**Response Generation.** We compare BlenderBot using our method (ConSTrain) with multiple strongest baselines for response generation. Table 4 summarizes the human evaluation results based on the Best-Worst Scaling. Our response model outperforms all baselines in terms of all the metrics on both ESConv and MSC, as indicated by their negative scores. Most of the results are statistically significant. The automatic evaluation results with MAUVE in Table 5, one of the best automatic metrics for NLG tasks, also demonstrates the strengths of our method over the baselines. This meets our expectation that responses generated based on direct causes perform better than responses generated on histories including spuriously correlated utterances.

Surprisingly, the BlenderBot using our method outperforms human responses on ESConv in terms of fluency and informativeness. A close look at the results reveals that i) some of the responses generated by our model are longer than the corresponding human responses because they cover more specific details in contexts, and ii) a significant amount of responses in ESConv contain grammatical errors while the model generated ones rarely make grammatical errors. Unfortunately, our model does not reach human-level performance on MSC in terms of informativeness and relevance, in which the majority of the multi-session conversations span more than 40 turns.

The two model variations in Liu et al. (2021) are the reported strongest baselines on ESConv, while

| Models | Empathy↑ | Fluency↑ | Informativeness↑ | Relevance↑ |
|---|---|---|---|---|
| | | ESConv | | |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-1})$ | −22* | −48* | −15* | −4 |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-2:t-1})$ | −83* | −46* | −12 | −26* |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-3:t-1})$ | −28* | −39* | −31* | −31* |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{0:t-1})$ | −54* | −36* | −16* | −38* |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_j, \boldsymbol{u}_{t-1})$ | −69* | −61* | −25* | −51* |
| DialoFlow | −38* | −54* | −6 | −28* |
| (Liu et al., 2021) w/o strategy | −64* | −45* | −6 | −9 |
| (Liu et al., 2021) with strategy | −52* | −36* | −13* | −19* |
| Retrieval-guided | −3 | −14* | −12* | −18* |
| ConSTrain (Ours) | 0 | **0** | **0** | 0 |
| Human | **12** | −30* | −16* | **3** |
| | | MSC | | |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-1})$ | – | −31* | −25* | −7 |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-2:t-1})$ | – | −54* | −24* | −35* |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-3:t-1})$ | – | −12 | −8 | −4 |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{0:t-1})$ | – | −80* | −30* | −80* |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_j, \boldsymbol{u}_{t-1})$ | – | −82* | −71* | −66* |
| DialoFlow | – | −54* | −35* | −51* |
| TransferTransfo | – | −49* | −44* | −48* |
| Retriever-generator | – | −64* | −10 | −14 |
| Retrieval-guided | – | −12 | −29* | −32* |
| ConSTrain (Ours) | – | 0 | 0 | 0 |
| Human | – | **3** | **19*** | **19*** |

Table 4: Results of human evaluation using best-worst scaling (higher is better). The results in **Bold** are better than all the competitors. Systems significantly different from our method are marked with an asterisk * (using a one-way ANOVA with post hoc Tukey HSD tests; $p \leqslant 0.05$).

the retriever-generator model is the strongest one on MSC in literature. Both the retriever-generator and the retrieval-guided model apply retrieval techniques to identify the most relevant texts in context. The retrieval-guided model starts with employing the tokens in the preceding utterance $\boldsymbol{u}_{t-1}$ as queries to retrieve the most relevant tokens in the context $\{\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{t-2}\}$, followed by concatenating them with the ones in $\boldsymbol{u}_{t-1}$ as model inputs. In contrast, retriever-generator identifies relevant utterances in histories. Despite that, all of them still fall short of our method according to human and automatic evaluations. Those results indicate that retrieval techniques are still limited for identifying key utterances from conversation histories.

We compare different ways of selecting utterances from conversation histories as the inputs of the same neural architecture. Table 4 and Table 5 include the corresponding results of BlenderBot on both corpora. Taking the full conversation histories as input, which is widely used in practice, turns out to be a poor choice on both corpora. The responses generated in this setting are often too general, such as ''I'm sorry to hear that.'', without touching specific details in contexts. As a comparison, using the preceding utterances is evident as a good heuristic on ESConv, while the best heuristic on MSC is to use the preceding three utterances. The worse case is $P(\boldsymbol{r}_t|\boldsymbol{u}_j, \boldsymbol{u}_{t-1})$, which randomly selects an utterance between the first utterance and $\boldsymbol{u}_{t-2}$ to combine with $\boldsymbol{u}_{t-1}$. The corresponding ratio of spurious correlations is one of the highest among all settings. Those results again demonstrate the harm of spuriously correlated utterances for generative models.

To demonstrate that our method is model-agnostic, we apply our method to DialogGPT[3]

[3]https://huggingface.co/microsoft/DialoGPT-medium.

| Models | BLEU↑ | BERTScore↑ | MAUVE↑ | METEOR↑ | D-1↑ | D-2↑ |
|---|---|---|---|---|---|---|
| | | | ESConv | | | |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-1})$ | 0.09 | 0.19 | 0.24 | 0.12 | 0.26 | 0.72 |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-2:t-1})$ | 0.09 | 0.19 | 0.32 | 0.12 | 0.27 | 0.73 |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-3:t-1})$ | 0.08 | 0.18 | 0.24 | 0.13 | 0.27 | 0.73 |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{0:t-1})$ | 0.08 | 0.15 | 0.09 | 0.11 | 0.27 | 0.73 |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_j, \boldsymbol{u}_{t-1})$ | 0.07 | 0.14 | 0.29 | 0.11 | 0.24 | 0.70 |
| DialoFlow | 0.05 | 0.14 | 0.19 | 0.07 | 0.23 | 0.72 |
| (Liu et al., 2021) w/o strategy | 0.09 | 0.18 | 0.31 | 0.12 | 0.24 | 0.70 |
| (Liu et al., 2021) with strategy | 0.07 | 0.18 | 0.21 | 0.13 | 0.27 | 0.73 |
| Retrieval-guided | 0.07 | 0.17 | 0.27 | 0.12 | 0.26 | 0.72 |
| ConSTrain (Ours) | 0.08 | 0.18 | 0.33 | 0.13 | 0.26 | 0.73 |
| | | | MSC | | | |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-1})$ | 0.09 | 0.20 | 0.28 | 0.11 | 0.28 | 0.74 |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-2:t-1})$ | 0.09 | 0.20 | 0.30 | 0.10 | 0.29 | 0.76 |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-3:t-1})$ | 0.08 | 0.18 | 0.23 | 0.11 | 0.29 | 0.76 |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_{0:t-1})$ | 0.06 | 0.13 | 0.02 | 0.08 | 0.26 | 0.75 |
| BlenderBot - $P(\boldsymbol{r}_t\|\boldsymbol{u}_j, \boldsymbol{u}_{t-1})$ | 0.07 | 0.16 | 0.07 | 0.09 | 0.27 | 0.74 |
| DialoFlow | 0.05 | 0.14 | 0.16 | 0.08 | 0.33 | 0.74 |
| TransferTransfo | 0.07 | 0.13 | 0.10 | 0.05 | 0.50 | 0.89 |
| Retriever-generator | 0.09 | 0.20 | 0.25 | 0.10 | 0.29 | 0.75 |
| Retrieval-guided | 0.08 | 0.18 | 0.20 | 0.11 | 0.26 | 0.74 |
| ConSTrain (Ours) | 0.09 | 0.20 | 0.31 | 0.13 | 0.29 | 0.76 |

Table 5: Automatic evaluation results contain BLEU, BERTScore (F1), MAUVE, METEOR, and Distinct (D1 and D2). Distinct score is calculated on 1-gram and 2-gram on corpus level.

| Models | Empa↑ | Fluen↑ | Info↑ | Rele↑ |
|---|---|---|---|---|
| | | ESConv | | |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-1})$ | −3 | −11 | −14* | −21* |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-2:t-1})$ | −12 | −17* | −25* | −28* |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-3:t-1})$ | −11 | −5 | −25* | −18* |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{0:t-1})$ | −26* | −32* | −22* | −20* |
| ConSTrain | **0** | **0** | **0** | **0** |
| | | MSC | | |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-1})$ | – | −9 | −7 | −11 |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-2:t-1})$ | – | −5 | −10 | −15* |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-3:t-1})$ | – | −16* | −28* | −18* |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{0:t-1})$ | – | −13* | −23* | −17* |
| ConSTrain | – | **0** | **0** | **0** |

Table 6: Model-agnostic experiment results where all models use DialogGPT as backbone. * indicates significant difference with ConSTrain - $\boldsymbol{u}_{MaxCI,t-1}$.

instead of BlenderBot, and evaluate the models on both ESConv and MSC with varying input settings. As one can see from Table 6, our method outperforms the other DialogGPT models with different input settings in terms of all metrics. As DialogGPT uses only a transformer-based decoder, we show that our training and inference methods improve the performance of both decoder-only and encoder-decoder neural architectures.

**Ablation Study of Response Generation.** We conduct ablation studies to demonstrate that *conditional dependence* is crucial for selecting direct causes during training and inference. The corresponding results are summarized in Table 7.

Training generative models with the utterances selected by our method improves model performance significantly. Without our method, empathy, informativeness and relevance drop for all BlenderBot variations on ESConv. Only the fluency increases slightly when using the preceding two utterances as input during training. It is worth noting that training models with the utterances selected by our CI classifier improves the diversity of response candidates consistently. From Table 8 we can see the diversity of response candidates produced by different response models. The model trained with our method generates more diverse response candidates than the other ones in terms of all metrics. We conjecture that training with direct causes can let model parameters focus on associating key differences among inputs with responses, thus becoming more sensitive to input variations.

Using BlenderBot trained with our method (ConSTrain), we compare our inference method,

| | ESConv | | | | MSC | | | |
|---|---|---|---|---|---|---|---|---|
| **Models** | **Empa↑** | **Fluen↑** | **Info↑** | **Rele↑** | **Empa↑** | **Fluen↑** | **Info↑** | **Rele↑** |
| ConSTrain (Ours) | 0 | 0 | 0 | 0 | – | 0 | 0 | 0 |
| ConSTrain - $\boldsymbol{u}_{t-2,t-1}$ | −21* | 2 | −6 | −7 | – | −2 | 3 | −4 |
| ConSTrain - $\boldsymbol{u}_{MaxDep,t-1}$ | −9 | −13* | −14* | −10 | – | −17* | 4 | −18* |
| ConSTrain - $\boldsymbol{u}_{Random,t-1}$ | −22* | −18* | −19* | −23* | – | −28* | −14 | −20* |
| ConSTrain - $\boldsymbol{u}_{Entropy,t-1}$ | −26* | −28* | −8 | −23* | – | −21* | −17* | −21* |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-2},\boldsymbol{u}_{t-1})$ - $\boldsymbol{u}_{MaxCI,t-1}$ | −17* | **11** | −19* | −5 | – | **10** | −21* | −5 |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{0:t-1})$ - $\boldsymbol{u}_{MaxCI,t-1}$ | −23* | −25* | −10 | −21* | – | −5 | **8** | −18* |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{0:t-1})$ - $\boldsymbol{u}_{t-2,t-1}$ | −22* | −11 | −9 | −13* | – | −9 | −16* | −15* |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{random},\boldsymbol{u}_{t-1})$ - $\boldsymbol{u}_{MaxCI,t-1}$ | −43* | −36* | −30* | −43* | – | −12* | −15* | −25* |
| ConSTrain - Beam | **3** | −2 | **1** | **3** | – | 5 | −5 | **6** |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-1})$ - Beam | −20* | −35* | −33* | −6 | – | −39* | −30* | −9 |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{0:t-1})$ - Beam | −40* | −28* | −23* | −37* | – | −52* | −49* | −14* |

Table 7: The comparisons between inference methods. All models are fine-tuned on BlenderBot. * indicates a significant difference with our model. ''Beam'' indicates regularized beam search that employs a width of 10 with 3-grams blocking and a minimum length of 20.

| Models | **Self-BLEU ↓** | **D-1 ↑** | **D-2 ↑** |
|---|---|---|---|
| ESConv | | | |
| ConSTrain | **0.42** | **0.27** | **0.74** |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-2},\boldsymbol{u}_{t-1})$ | 0.69 | 0.27 | 0.70 |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{0:t-1})$ | 0.71 | 0.24 | 0.62 |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{random},\boldsymbol{u}_{t-1})$ | 0.91 | 0.19 | 0.59 |
| MSC | | | |
| ConSTrain | **0.69** | **0.32** | **0.78** |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-2},\boldsymbol{u}_{t-1})$ | 0.78 | 0.30 | 0.75 |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{0:t-1})$ | 0.80 | 0.27 | 0.74 |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{random},\boldsymbol{u}_{t-1})$ | 0.93 | 0.20 | 0.53 |

Table 8: Response candidates diversity. All models are fine-tuned on BlenderBot.

coined $\boldsymbol{u}_{MaxCI,t-1}$, with alternative methods: i) randomly selecting $\boldsymbol{u}_j$ between 0 and $t-2$ and combining it with $\boldsymbol{u}_{t-1}$, coined $\boldsymbol{u}_{Random,t-1}$; ii) taking both $\boldsymbol{u}_{t-2}$ and $\boldsymbol{u}_{t-1}$ as input, coined $\boldsymbol{u}_{t-2,t-1}$; iii) applying the entropy-based method proposed in Csáky et al. (2019) to remove generic response candidates and select optimal response, coined $\boldsymbol{u}_{Entropy,t-1}$; and iv) replacing the CI classifier with a dependence classifier for inference, coined $\boldsymbol{u}_{MaxDep,t-1}$. The dependence classifier is trained by setting $(\boldsymbol{u}_{t-1},\boldsymbol{r}_t)$ as positive samples, $(\boldsymbol{u}_j,\boldsymbol{r}_t)$ as negative samples, where $\boldsymbol{u}_j$ far from responses is randomly sampled from dialogue histories. During inference, we generate response candidates in the same way as our method $\boldsymbol{u}_{MaxCI,t-1}$, but select the candidate that has the highest dependence probability $P_{depend}(l=1|\boldsymbol{u}_j,\boldsymbol{r}_t^j)$ as the final output.

The results in Table 7 show that our inference method outperforms alternative inference methods, when the models are trained with our method. Replacing the CI classifier with the dependence classifier ($\boldsymbol{u}_{MaxDep,t-1}$) leads to substantial performance drops in terms of all metrics. It is also noteworthy that generating responses using the preceding two utterances ($\boldsymbol{u}_{t-2,t-1}$) is a fairly effective heuristic, which only falls short of our method in terms of empathy. This can be explained by the statistics that $40\%$ of direct causes on ESConv are the preceding two utterances, while the corresponding percentage on MSC is $29\%$. Selecting key utterances randomly or using entropy to pair with $\boldsymbol{u}_{t-1}$ is worse than that simple heuristic.

In addition, we compare our method with regularized beam search (Roller et al., 2021) in three settings: i) replacing the unregularlized beam search with the regularized one using our method, ii) using only preceding utterances as input, and iii) using full conversation histories as input. In all settings, the beam search employs a width of 10 with 3-grams blocking and a minimum length of 20. Regularized beam search with full conversation histories ($P(\boldsymbol{r}_t|\boldsymbol{u}_{0:t-1})$-Beam) or only preceding utterances ($P(\boldsymbol{r}_t|\boldsymbol{u}_{t-1})$-Beam) achieve dramatically lower performance than our inference method. If the beam search is used together with the CI classifier (ConSTrain-Beam), the model performance increases slightly but the differences are not statistically significant.

**Qualitative Studies.** To further investigate the differences between the CI classifier and the dependence classifier, we apply the model to

| Human | How long are you doing the online school? (−2, 0, 0, 1) | | |
|---|---|---|---|
| Input | Generated Response | Conditional dependence | Dependence |
| | ConSTrain (Ours) | | |
| $(\boldsymbol{u}_0, \boldsymbol{u}_5)$ | I am sorry to hear that. What is causing you to be so stressed right now? | 0.0004 | 0.8314 |
| $(\boldsymbol{u}_1, \boldsymbol{u}_5)$ | Do you have any friends or family that you can talk to about it? | 0.0003 | 0.0115 |
| $(\boldsymbol{u}_2, \boldsymbol{u}_5)$ | I am doing well. I am sorry to hear that school is stressful for you right now. | 0.0004 | 0.9996 |
| $(\boldsymbol{u}_3, \boldsymbol{u}_5)$ | **I can understand how stressful this is for you. Have you thought about talking to a school counselor?** | 0.7287 | 0.9287 |
| $(\boldsymbol{u}_4, \boldsymbol{u}_5)$ | Yes, I can understand that. Do you have any friends or family that you can talk to about it? | 0.0112 | 0.9981 |
| | $P(\boldsymbol{r}_t\|\boldsymbol{u}_{0:t-1})$ - $\boldsymbol{u}_{MaxCI,t-1}$ | | |
| $(\boldsymbol{u}_0, \boldsymbol{u}_5)$ | I am sorry to hear that. I can understand how stressful it is to be in school right now. | 0.0003 | 0.0028 |
| $(\boldsymbol{u}_1, \boldsymbol{u}_5)$ | I'm sorry to hear that. What is stressing you out the most about school right now? | 0.0003 | 0.999 |
| $(\boldsymbol{u}_2, \boldsymbol{u}_5)$ | I can understand how stressful school can be. What is stressing you out the most about it? | 0.0003 | 0.9472 |
| $(\boldsymbol{u}_3, \boldsymbol{u}_5)$ | **I can understand how stressful online learning can be, especially online learning.** (1, 0, −1, −1) | 0.872 | 0.7682 |
| $(\boldsymbol{u}_4, \boldsymbol{u}_5)$ | I can understand that. It seems like everyone is stressed about the vaccine. | 0.2121 | 0.9996 |
| | Other baselines | | |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-1})$ | It sounds like you're stressed about the way things are going in school. Is that correct? (−3, 0, −3, 1) | | |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-2:t-1})$ | Have you tried to talk to someone about it? (−2, 0, −3, −2) | | |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-3:t-1})$ | I can understand how stressful this is for you. Do you have anyone you can talk to about this? (−1, 0, 1, −1) | | |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{0:t-1})$ | I can understand how stressful it is to be in school right now. ( −2, 1, −3, 0) | | |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_j, \boldsymbol{u}_{t-1})$ | I am sorry to hear that. What are you studying? (−2, −1, −2, −2) | | |
| DialoFlow | I'm sorry to hear that. I hope you feel better soon. (−1, 0, −2, −2) | | |
| (Liu et al., 2021) | I'm sorry to hear that. What is stressing you out the most about school right now? (−1, 0, −1, −2) | | |
| w/o strategy | | | |
| (Liu et al., 2021) | I can understand how stressful school can be. What is stressing you out the most about it? (−1, 0, −2, 1) | | |
| with strategy | | | |
| Retrieval-guided | I can understand how stressful this is for you. I am sure you are doing your best. (0, 1, −1,−2) | | |

Table 9: Response candidates generated by ConSTrain and one baseline $P(\boldsymbol{r}_t|\boldsymbol{u}_{0:t-1})$ based on the conversation history in Table 1. We use $\boldsymbol{u}_{MaxCI,t-1}$ to select final responses, which are in **bold**. Behind responses generated by baselines, we append pair-wise comparison results annotated by five workers between baselines and our model, (Empathy, fluency, informativeness, relevance). In a pair-wise comparison, if baseline is better, it gets a +1 score; if baseline is worse, it gets a -1 score; if baseline is the same with our model, both get 0 score. The sum of the five workers' evaluations is the score shown in this Table.

generate all candidate responses and score the candidates with the probabilities yielded by the dependence and the CI classifiers. Using the example conversation in Table 1, we show all generated candidate responses and the corresponding scores in Table 9. With $\boldsymbol{u}_3$, the direct cause used by humans, the corresponding response achieves the highest conditional dependence probability but not the highest dependence probability. Perplexity is also not reliable. Moreover, the distributions of the conditional dependence scores are more skewed towards the true direct causes than those of dependence scores. Hence, the conditional dependence, which measures the conditional mutual information obtained from a selected utterance beyond that from the preceding utterance, is more informative and robust than mutual information between responses and single utterances in contexts.

Furthermore, we apply our method to BlenderBot on example dialogues and show qualitative differences to the baselines. Table 9 shows the responses generated by our method and the baselines using the running example in Table 1. The responses generated by our method give a specific suggestion to ''talk to a school counselor'' or refer to the most specific detail of ''online learning'', while the remaining ones talk about school or irrelevant contents. In addition, we provide the Best-Worse Scaling scores of five crowd-workers, who compare the baseline outputs with those of our method. Most crowd-workers consider our model output is better than that of the baselines in terms of informativeness and relevance.

For error analysis, we find that model cannot always generate natural and relevant responses by relying on the same direct causes as humans. As shown in Table 10, although there are overlapped direct causes between humans and our model, the response generated by our model is reasonable and relevant by capturing context specific entities ''son'' and ''boyfriend'', while the other models fail to do so. In those cases, even if our model uses different direct causes than humans for response generation, most of them are reasonable and fluent. To further investigate to what degree our model utilizes the same direct causes as humans, we apply our model to the test set of CGDialog and collect the direct causes used during inference. The percentage of using exactly same causes, partially overlapped causes and totally different causes amount to 26.47%,

| | | |
|---|---|---|
| History | Supporter | Hi! How are you doing today? |
| | Seeker | I am struggling with how to turn in this situation.<br>My son is unreasonable but I am trying to help him get through school.<br>My boyfriend finds the situation intolerable. |
| | Supporter | That sounds really hard. Does your son fight with you a lot? |
| | Seeker | Yes, since he got back from NYC he is disrespectful. That is what is upsetting my boyfriend so much. |
| | Supporter | What was he in NYC for? Have you tried getting counseling together? |
| | Seeker | **He went to school there and left school to get acting jobs. He did get his sag and equity cards.**<br>**He was also working as a caterer. Everything dried up due to Covid.**<br>**In June, I convinced him to get into therapy but it is slow and I see jealousy with directed at my boyfriend.** |
| | Supporter | Has he gone to therapy with you? Maybe you can try doing some fun activties together.<br>I know my brother had a lot of tension with my mom and finding something they could bond over really helped. |
| | Seeker | **I have tried biking and hiking as well as other activities. He is openly hostile.**<br>**He does not want anything to do with my boyfriend. He does not want me to go to therapy with him.** |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-1})$ | Supporter | I can understand why he is hostile to you. Have you tried talking to him about it? (−2, 0, −2, 1) |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-2:t-1})$ | Supporter | I am so sorry to hear that. I have been in a similar situation and I know how hard it is. (0, 0, −3, −2) |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{t-3:t-1})$ | Supporter | Have you tried talking to him about how you feel? Sometimes it helps to talk things through. (−2, 0, −1, 0) |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_{0:t-1})$ | Supporter | I am so sorry to hear that. Is there anything else I can help you with today? (−2, 0, −2, −3) |
| $P(\boldsymbol{r}_t\|\boldsymbol{u}_j, \boldsymbol{u}_{t-1})$ | Supporter | I can understand why that would upset you. Have you talked to him about how you feel? (1, 0, −2, 0) |
| DialoFlow (Liu et al., 2021) | Supporter | I understand how you feel.. It's hard to get away from it. (−1, −1, −2, −1) |
| w/o strategy (Liu et al., 2021) | Supporter | I can understand why that would upset you. Have you talked to him about how you feel? (1, 0, −1, −1) |
| with strategy | Supporter | I am so sorry to hear that. I have been in a similar situation before and I know how hard that can be. (0, 0, −2, −2) |
| Retrieval-guided | Supporter | I can understand how stressful this must be for you. Is there anyone you can talk to about this? (−1, 0, −2, −1) |
| ConStrain (Ours) | Supporter | It sounds like you are trying your best to help your son and your boyfriend at the same time. |
| Human | Supporter | Does he give a reason why? (−2, 0, −1, 2) |

Table 10: An example where causes of the human response and the generated response partially overlap. The causes of human response are in bold. The causes of the response generated by our model are highlighted. Behind responses generated by baselines, we append pair-wise comparison results annotated by five workers between baselines and our model, (Empathy, fluency, informativeness, relevance).

62.13%, and 11.40%, respectively. Overall, comparing with the baselines, the model with our method produces more specific, relevant, and natural responses than the baselines regardless if it uses the same direct causes as humans or not.

**CI Classification Results.** We evaluate our method ConStrain to identify direct causes of responses in the test sets of CGDialog, and compare them with two simple but strong baselines: ''Always $\boldsymbol{u}_{t-1}$'' and ''Always $\boldsymbol{u}_{t-2}, \boldsymbol{u}_{t-1}$''. The former always considers $\boldsymbol{u}_{t-1}$ of responses as direct causes, while the latter considers the preceding two utterances as direct causes. In the test sets, we keep the manually annotated cause-response pairs as positive examples, while combining all non-cause utterances with $\boldsymbol{u}_{t-1}$ and $\boldsymbol{r}_t$ as negative samples. As a result, the number of negative samples is much larger than the number of positive examples. Due to such an imbalance, we adopt precision, recall, and F1 as the evaluation metrics.

| Models | Precision | Recall | F1 |
|---|---|---|---|
| CGDialog - ESConv | | | |
| Always $\boldsymbol{u}_{t-1}$ | **0.80** | 0.41 | 0.54 |
| Always $\boldsymbol{u}_{t-2}, \boldsymbol{u}_{t-1}$ | 0.60 | 0.61 | 0.61 |
| INIT | 0.63 | 0.41 | 0.49 |
| FC | 0.43 | 0.54 | 0.47 |
| IST | 0.67 | 0.33 | 0.44 |
| ConStrain | 0.70 | **0.71** | **0.70** |
| CGDialog - MSC | | | |
| Always $\boldsymbol{u}_{t-1}$ | **0.98** | 0.51 | 0.67 |
| Always $\boldsymbol{u}_{t-2}, \boldsymbol{u}_{t-1}$ | 0.64 | 0.66 | 0.65 |
| INIT | 0.70 | 0.60 | 0.65 |
| FC | 0.49 | 0.59 | 0.54 |
| IST | 0.73 | 0.54 | 0.62 |
| ConStrain | 0.73 | **0.72** | **0.73** |

Table 11: The results of direct cause identification on the test sets of CGDialog.

Table 11 reports the results of cause identification. ConStrain reaches the highest recall and F1 on this task. ''Always $\boldsymbol{u}_{t-1}$'' reaches the highest precision because preceding utterances have the highest probability to be direct causes, as we discussed in Section 3.1. We also created a balanced test set by randomly sampling non-cause utterances and combining them with $\boldsymbol{u}_{t-1}$ and

$\boldsymbol{r}_t$ as negative examples. The accuracy of ConSTrain is 0.83 on CGDialog - ESConv, and 0.86 on CGDialog - MSC, much higher than random guess.

Furthermore, we evaluate the effectiveness of incremental self-training with constraints on the test sets of CGDialog by comparing it with three options: i) training only the initial classifier on the labeled training set $\mathbb{D}_L$ of CGDialog (INIT), ii) fine-tuning the initial classifier on the full unlabeled training set with the context constraint (FC), and iii) incremental self-training without the context constraint on the full unlabeled training set (IST). As shown in Table 11, ConSTrain outperforms the three options in terms of recall by a wide margin, hence achieves the highest F1 scores on both datasets. Applying the context constraint during self-training filters out mislabeled data far from responses, dropping it leads to the largest reduction of recall and F1. The threshold constraint is still effective by boosting both the precision and the recall of direct cause identification.

## 6 Related Work

**Dialogue Datasets** Recently, state-of-the-art open-domain dialogue agents have utilized DailyDialog (Li et al., 2017), PersonaChat (Zhang et al., 2018), EmpatheticDialogues (Rashkin et al., 2019), and Wizard of Wikipedia (Dinan et al., 2019). Dialogues in these datasets usually have 3-15 turns. Dialogue agents trained on these dataset don't have the ability to deal with dialogue with very long history. This weakness encourages researchers to crowdsource long conversations, such as Emotion Support Conversation (Liu et al., 2021) and Multi-Session Chat (Xu et al., 2022). The number of utterances per dialogue in two datasets is 30 and 53, respectively.

**Dialogue Models** Recently, seq2seq dialogue models, such as DialoGPT, Blenderbot, and PLATO (Zhang et al., 2020; Roller et al., 2021; Bao et al., 2020), showed significant improvement in generating fluent and relevant responses in various dialogue datasets. Xu et al. (2022), Lewis et al. (2020), Izacard and Grave (2021), and Qu et al. (2021) propose retrieval-based dialog systems that select relevant utterances from history as input. However, such methods select utterances based on semantic relevance, which may still suffer from spurious correlation in input. Whang et al. (2021), Niu and Bansal (2018), Lee and Choi (2022), and Akama et al. (2020) seek to first generate or retrieve response candidates, then select final responses using dialog–response binary classifier. Such binary classifiers are trained to identify relevance or irrelevance. However, relevance includes causation and spurious correlation, which cannot be identified by those classifiers.

## 7 Conclusion

We conduct the first study from a causal view to investigate and tackle spurious correlations in dialogues. Inspired by constraint-based causal discovery algorithms, we propose a novel constrained self-training method to build a CI classifier by using a small corpus CGDialog, which is manually annotated with causal graphs by us. The CI classifier is applied to filter out spuriously correlated utterances in conversation histories before training a response generation model. That classifier also serves as a scoring function during inference to select the best response from all generated candidates. By identifying conditionally dependencies between utterances and responses, our model agnostic approach significantly improves the overall generation quality of response models in terms of relevance, informativeness and fluency.

## Acknowledgments

We thank the action editor and the anonymous reviewers for their constructive feedback. This material is based on research sponsored by Air Force Research Laboratory and DARPA under agreement numbers FA8750-19-2-0501 and HR001122C0029. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The computational resources of this work are supported by the Multi-modal Australian ScienceS Imaging and Visualisation Environment (MASSIVE).

## References

Reina Akama, Sho Yokoi, Jun Suzuki, and Kentaro Inui. 2020. Filtering noisy dialogue corpora by connectivity and content relatedness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 941–958, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.229

Reinald Kim Amplayo and Mirella Lapata. 2021. Informative and controllable opinion summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2662–2672, Online. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.9

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift Reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839. https://doi.org/10.1609/icwsm.v14i1.7347

Alexis Bellot and Mihaela van der Schaar. 2019. Conditional independence testing using generative adversarial networks. https://doi.org/10.48550/arXiv.1907.04068

Anja Belz and Eric Kow. 2010. Comparing rating scales and preference judgements in language evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2301

Guendalina Caldarini, Sardar Jaf, and Kenneth McGarry. 2022. A literature survey of recent advances in chatbots. *Information*, 13(1). https://doi.org/10.3390/info13010041

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics. https://doi.org/10.3115/1626355.1626373

Richárd Csáky, Patrik Purgai, and Gábor Recski. 2019. Improving neural conversational models with entropy-based data filtering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5650–5669, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1567

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents.

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart. 2020. *Statistical Significance Testing for Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. https://doi.org/10.1007/978-3-031-02174-9

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.74

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2074

Nyoungwoo Lee and Ho-Jin Choi. 2022. Toward robust response selection model for cross negative sampling condition. In *2022 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 395–397.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995. Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 128–138, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.11

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.269

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. https://doi.org/10.48550/arXiv.1907.11692

David Lopez-Paz and Maxime Oquab. 2017. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*.

Jordan J. Louviere, Terry N. Flynn, and Anthony Alfred John Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press. https://doi.org/10.1017/CBO9781107337855

Brady Neal. 2020. *Introduction to Causal Inference from a Machine Learning Perspective*.

Tong Niu and Mohit Bansal. 2018. Adversarial over-sensitivity and over-stability strategies for dialogue models. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 486–496, Brussels, Belgium. Association for Computational Linguistics.

Ana Rita Nogueira, João Gama, and Carlos Abreu Ferreira. 2021. Causal discovery in machine learning: Theories and applications. *Journal of Dynamics and Games*, 8(3):203–231. https://doi.org/10.3934/jdg.20210085

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2012

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing SystemsDecember 2019 Article No.: 721,* pages 8026–8037.

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

Judea Pearl and Thomas Verma. 1991. A theory of inferred causation. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, KR'91, pages 441–452, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332. https://doi.org/10.1007/s12559-021-09925-7

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9:510–527. https://doi.org/10.1162/tacl_a_00381

Lisong Qiu, Juntao Li, Wei Bi, Dongyan Zhao, and Rui Yan. 2019. Are training samples correlated? Learning to generate dialogue responses with multiple references. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3826–3835, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1372

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.466

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. https://doi.org/10.18653/v1/D16-1264

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A

new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1534

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.24

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? An empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1004

Rajat Sen, Karthikeyan Shanmugam, Himanshu Asnani, Arman Rahimzamani, and Sreeram Kannan. 2018. Mimic and classify: A meta-algorithm for conditional independence testing. https://doi.org/10.48550/arXiv.1806.09708

Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G. Dimakis, and Sanjay Shakkottai. 2017. Model-powered conditional independence test. *Advances in Neural Information Processing Systems*, 30.

Herbert A. Simon. 1954. Spurious correlation: A causal interpretation. *Journal of the American Statistical Association*, 49(267):467–479. https://doi.org/10.1080/01621459.1954.10483515

Peter Spirtes, Clark N. Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, Prediction, and Search*, MIT Press.

Julius Steen and Katja Markert. 2021. How to evaluate a summarizer: Study design and statistical analysis for manual linguistic quality evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1861–1875, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.160

Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A. Efros. 2019. Unsupervised domain adaptation through self-supervision.

Xiangru Tang, Alexander Fabbri, Haoran Li, Ziming Mao, Griffin Adams, Borui Wang, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. Investigating crowdsourcing protocols for evaluating the factual consistency of summaries. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5680–5692, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.160

Haohan Wang, Zeyi Huang, and Eric Xing. 2021. Learning robust models by countering spurious correlations.

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021. In *Do response selection models really know what's next? Utterance manipulation strategies for multi-turn response selection*. volume 35, pages 14041–14049. https://doi.org/10.1609/aaai.v35i16.17653

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on*

*Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A transfer learning approach for neural network based conversational agents. https://doi.org/10.48550/arXiv.1901.08149

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.356

Xi Yang, Xing He, Hansi Zhang, Yinghan Ma, Jiang Bian, and Yonghui Wu. 2020. Measurement of semantic textual similarity in clinical texts: Comparison of transformer-based models. *JMIR Medical Informatics*, 8(11): e19735. https://doi.org/10.2196/19735, PubMed: 33226350

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1205

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.30

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. CoMAE: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.72

Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is more: Learning to refine dialogue history for personalized dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5808–5820, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.426

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. In *Emotional chatting machine: Emotional conversation generation with internal and external memory*. *Proceedings of the AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v32i1.11325

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. 2019. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991. https://doi.org/10.1109/ICCV.2019.00608