Counterfactual Data Augmentation via Perspective Transition for Open-Domain Dialogues

Jiao Ou[1,2], **Jinchao Zhang**[3], **Yang Feng**[1,2]*, **Jie Zhou**[3]
[1] Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] University of Chinese Academy of Sciences
[3] Pattern Recognition Center, WeChat AI, Tencent Inc, China
{oujiao17b,fengyang}@ict.ac.cn, {dayerzhang,withtomzhou}@tencent.com

Abstract

The construction of open-domain dialogue systems requires high-quality dialogue datasets. The dialogue data admits a wide variety of responses for a given dialogue history, especially responses with different semantics. However, collecting high-quality such a dataset in most scenarios is labor-intensive and time-consuming. In this paper, we propose a data augmentation method to automatically augment high-quality responses with different semantics by counterfactual inference. Specifically, given an observed dialogue, our counterfactual generation model first infers semantically different responses by replacing the observed reply perspective with substituted ones. Furthermore, our data selection method filters out detrimental augmented responses. Experimental results show that our data augmentation method can augment high-quality responses with different semantics for a given dialogue history, and can outperform competitive baselines on multiple downstream tasks.

1 Introduction

Open-domain dialogue systems have attracted much attention (Chen et al., 2017; Huang et al., 2020; Ni et al., 2021; Fu et al., 2022) due to their potential applications. Generally, training open-domain dialogue systems requires high-quality dialogue datasets. The dialogue data admits a wide variety of responses for a given dialogue history (Hou et al., 2018). Specifically, a given dialogue history can exist many valid responses with different semantics, and the response of each semantic information can also have abundant alternative expressions (Li et al., 2019). However, manually collecting high-quality such datasets is usually labor-intensive and time-consuming in practice.

A feasible solution to address this problem is to use data augmentation techniques. Currently, some data augmentation methods have been used in open-domain dialogues (Sennrich et al., 2016; Niu and Bansal, 2019; Li et al., 2019; Cai et al., 2020; Zhang et al., 2020a; Xie et al., 2022) to augment data. However, the augmented data have limited semantic differences from the observed data based on the restrained changes. These existing methods only consider word- or sentence-level alternative expressions of the observed data without augmenting more valid responses with different semantics.

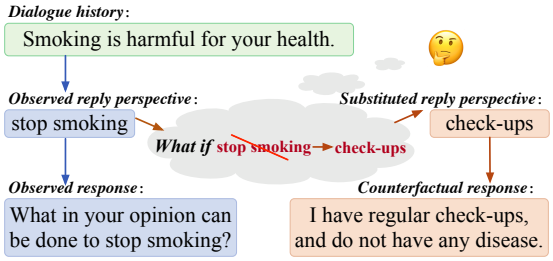


Figure 1: An example of a counterfactual response, which is a semantically different response re-inferred by changing the observed reply perspective.

In this paper, we propose to augment valid responses with different semantics for a given dialogue history. Imagine that when humans infer different-semantic responses, they may naturally ask a question: Given an observed dialogue, what the response would happen if we change the *reply perspective*, while keeping the current environment unchanged? Answering this question will infer a different response, given an example shown in Figure 1. The imagination of different responses under the current environment is so-called *counterfactual inference* (Pearl et al., 2000), which ensures the quality of inferred responses (Zhu et al., 2020).

Motivated by this, we propose a **C**ounterfactual data **A**ugmentation method via **P**erspective **T**ransition, CAPT for short, to generate counterfactual responses for a given observed dialogue. CAPT interprets a counterfactual generation model as a structural causal model (SCM), which de-

*Work done while Jiao Ou was interning at WeChat AI. Yang Feng is the corresponding author. Our code is public at https://github.com/ictnlp/CAPT.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1635–1648
December 7-11, 2022

scribes the generation process under the current environment. The current environment is modeled by unobserved variables in the SCM that capture all unobserved but relevant factors that affect response generation. Counterfactual responses are then generated by intervening in the reply perspective in the SCM, i.e., replacing the observed reply perspective with valid alternatives, while keeping these unobserved variables unchanged. To achieve an alternative, we first construct a shift graph based on all observed dialogues, which explicitly represents the shift associations between both focuses of attention on dialogue histories and their corresponding responses respectively. We then randomly choose a focus on the given dialogue history and regard its 1-hop neighbors in the shift graph as candidates. A valid alternative can be predicted from these candidates. After achieving all counterfactual augmented responses, the augmented data are further filtered using a data selection module. Finally, we merge the observed data with this augmented data as training data for downstream tasks.

Experiment results indicate that CAPT can augment high-quality responses with different semantics, and our augmented data contributes to the performance improvement of both retrieval-based and generation-based open-domain dialogue models. Our contributions are summarized as follows: (1) We propose a counterfactual data augmentation method via perspective transition to augment responses with different semantics for a given dialogue history. To the best of our knowledge, this is the first study to augment more valid responses with different semantics in open-domain dialogues. (2) Automatic and manual evaluation show that CAPT generates semantically different responses, which can be further used to improve the performance of downstream tasks. (3) Extensive experiments show that providing more responses with different semantics can further improve performance.

2 Background

In this section, we describe task definitions and review the concept of the structural causal model. Please see task definitions in Appendix A.

2.1 Structural Causal Model

Definition. A structural causal model (SCM) consists a set of observed variables $\mathbf{V} = \{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_m\}$ and a set of independent unobserved random variables $\mathbf{U} = \{\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m\}$ with distribution $P(\mathbf{U})$, which are connected by a set of functions $\mathbf{F} = \{f_1, \ldots, f_m\}$. Specifically, $\forall i$, $\boldsymbol{V}_i$ is caused by a set of parent variables $\mathbf{PA}_i$ and $\boldsymbol{U}_i$, i.e., $\boldsymbol{V}_i = f_i(\mathbf{PA}_i, \boldsymbol{U}_i)$, where $\mathbf{PA}_i \subseteq \mathbf{V} \setminus \boldsymbol{V}_i$ in the causal DAG (Buesing et al., 2019).

For the counterfactual generation model, it can be cast as an SCM with three observed variables, including *dialogue history* $\boldsymbol{X}$, *reply perspective* $\boldsymbol{Z}$ and *response* $\boldsymbol{Y}$. The counterfactual generation SCM turns the conditional distribution $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ into a deterministic function $\boldsymbol{Y} = f(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{U})$, where $\boldsymbol{U}$ captures all unobserved but influential factors of the current environment, such as speaking style. The function f is defined by the learned counterfactual generation model. Overall, SCM can infer counterfactual responses given the known function f and the posterior of the unobserved variable $P(\boldsymbol{U}|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{Y} = \boldsymbol{y})$.

Intervention. Before observing what the observed variable $\boldsymbol{V}_i$ would happen, an intervention would be given on a parent variable $\boldsymbol{V}_j$, $\boldsymbol{V}_j \in \mathbf{PA}_i$, where the intervention in the SCM is an action by changing the observed value. For the counterfactual generation SCM, the intervention is to replace the observed value $\boldsymbol{z}$ of the *reply perspective* $\boldsymbol{Z}$ with a different value $\tilde{\boldsymbol{z}}$.

Counterfactual Inference. Given an SCM and observed a variable $\boldsymbol{V}_i = \boldsymbol{v}_i$, counterfactual inference answers the question that what the observed variable $\boldsymbol{V}_i$ would have changed if a parent variable $\boldsymbol{V}_j$ has been intervened while keeping the current environment unchanged. Accordingly, generating a counterfactual response involves a query about what the response $\boldsymbol{Y}$ would have happened if an intervention is taken by setting $\boldsymbol{Z}$ as a different value $\tilde{\boldsymbol{z}}$, rather than the observed value $\boldsymbol{z}$.

Overall, to generate counterfactual responses, we can follow a three-step procedure (Pearl et al., 2016): (1) **Abduction**: Predict the "current environment of the SCM" , i.e., compute the posterior $P(\boldsymbol{U}|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{Y} = \boldsymbol{y})$ and sample $\boldsymbol{u}$ from it. (2) **Action**: Perform an intervention by replacing the observed value $\boldsymbol{z}$ of $\boldsymbol{Z}$ with a different value $\tilde{\boldsymbol{z}}$. (3) **Prediction**: Reason a counterfactual response $\tilde{\boldsymbol{y}}$, given the posterior sample $\boldsymbol{u}$ and the known function f.

3 Method

In this section, our goal is to take an input dialogue sample $(\boldsymbol{x}, \boldsymbol{y})$ and augment high-quality responses

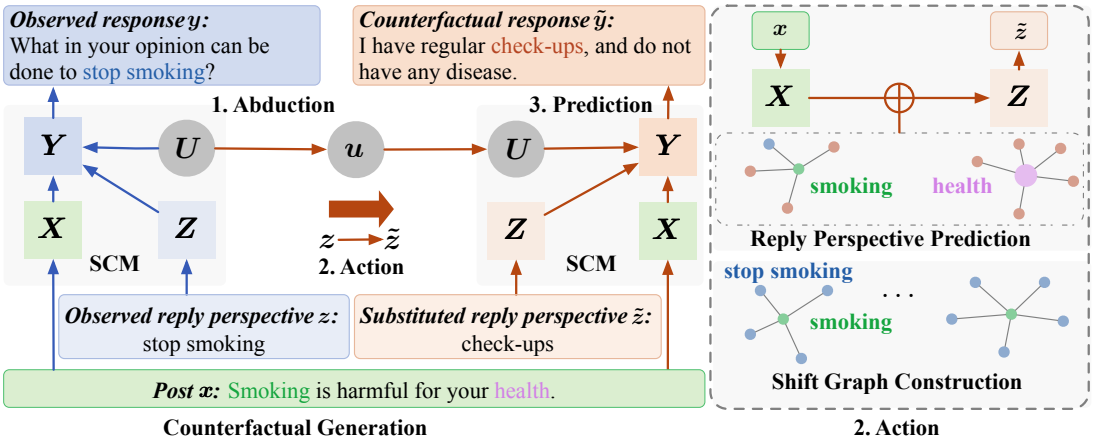


Figure 2: The three-step procedure of counterfactual generation: (1) Abduction: we estimate the "current environment of the SCM" $\boldsymbol{u}$ where the observed response $\boldsymbol{y}$ occurs. (2) Action: we perform an intervention on $\boldsymbol{Z}$ in the SCM by replacing $\boldsymbol{z}$ with $\tilde{\boldsymbol{z}}$. For obtaining $\tilde{\boldsymbol{z}}$, we first construct the shift graph by characterizing each observed shift between the focus (e.g., smoking) on $\boldsymbol{x}$ and the reply perspective $\boldsymbol{z}$ (e.g., stop smoking) in $\mathcal{D}$. We then randomly choose a focus on $\boldsymbol{x}$, e.g. health, and regard its 1-hop neighbors as candidates. Finally, we predict $\tilde{\boldsymbol{z}}$ from these candidates conditioned on the chosen focus of $\boldsymbol{x}$ and the post $\boldsymbol{x}$. (3) Prediction: the counterfactual response is generated based on the post $\boldsymbol{x}$ and the alternative $\tilde{\boldsymbol{z}}$ from the inferred $\boldsymbol{u}$.

that have different semantics from $\boldsymbol{y}$. To this end, in Section 3.1, we introduce a technique called *Counterfactual Generation via Perspective Transition* for intervening in the observed reply perspective to augment responses under the current environment. In Section 3.2, we describe how to train those models involved in Section 3.1, including the reply perspective predictor and the counterfactual generator. In Section 3.3, we design a data selection method, named *Bi-directional Perplexity Selection*, to select high-quality augmented data.

3.1 Counterfactual Generation via Perspective Transition

This paper mainly focuses on single-turn dialogues. Given a post-response pair $(\boldsymbol{x}, \boldsymbol{y})$, we use the SCM to generate a counterfactual response $\tilde{\boldsymbol{y}}$ following the three-step procedure shown in Figure 2.

1. Abduction. This step is to estimate the unobserved variable given the observed sample $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y})$ (for more details about $\boldsymbol{z}$ see the action step). Specifically, when generating the t-th token of $\boldsymbol{y}$, our counterfactual generator outputs a categorical distribution $P(Y_t|\boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Z}=\boldsymbol{z}, \boldsymbol{Y}_{<t}=\boldsymbol{y}_{<t})$, where $\boldsymbol{y}_{<t}$ is the token sequence generated in the previous time step. According to Oberst and Sontag (2019), the impact of the unobserved random variable $\boldsymbol{U}_t$ is simulated by introducing Gumbel random noises. Thus, we perform the Gumbel-Max Trick (Luce, 1959) for this categorical distribution as follows,

$$\begin{aligned} p_{tk} &= P(Y_t = k|\boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Z}=\boldsymbol{z}, \boldsymbol{Y}_{<t}=\boldsymbol{y}_{<t}), \\ y_t &= \arg\max_{k=1,\ldots,|V|}(\log p_{tk} + u_{tk}), \end{aligned} \tag{1}$$

where $u_{tk} \sim \text{Gumbel}(0, 1)$ and $|V|$ denotes the vocabulary size.

Consequently, our counterfactual generation SCM transforms into a Gumbel-Max SCM (Oberst and Sontag, 2019). The estimation of the unobserved variable is then to sample from the posterior distribution over these Gumbel random variables. Fortunately, a straightforward way to infer posterior (Maddison et al., 2014) is utilizing the properties of the shifted Gumbel variables $g_{tk} = \log p_{tk} + u_{tk}$: in the posterior, the maximum value is independent with the argmax of the shifted Gumbel variables and is distributed as a standard Gumbel. Thus, we first let $y_t = k^*$ (* denotes the observed token) and sample the maximum value g^*_{tk} from $\text{Gumbel}(0, 1)$. Secondly, we sample the remaining values g_{tk} from the shifted Gumbel distribution $\text{Gumbel}(\log p_{tk}, 1)$ truncated at g^*_{tk}. Then, for each index of k, a sample of u_{tk} is obtained by subtracting off the location parameter $\log p_{tk}$ from g_{tk}. Finally, the resulting sample $\boldsymbol{u}_t = [u_{t1}, \ldots, u_{t|V|}]$ is used to infer the counterfactual responses.

2. Action. This step is to replace the observed reply perspective $\boldsymbol{z}$ with a substituted reply per-

spective $\tilde{\boldsymbol{z}}$. However, two sub-problems need to be addressed: *representing the reply perspective* and *predicting a substituted valid reply perspective*. By observing human dialogues, we find that a reply perspective can be represented by a keyword, like "stop smoking" in Figure 2. It can be achieved based on the process that humans first naturally focus on a certain point of a given post like "smoking" and then would unconsciously shifting this focus point to another one. The focus point of the post can be similarly represented by a keyword. We name the focus point on the post and the shifted one as the *focus* and *reply perspective* respectively. When humans have different focuses (e.g., "health" in Figure 2) or different shifts on the same focus, they will obtain substituted reply perspectives.

To achieve valid alternatives, it is critical to make valid shifts from a focus. We build a shift graph based on all observed samples, where head and tail vertices are focuses and reply perspectives respectively, and edges represent observed shifts between focuses and reply perspectives. Inspired by Xu et al. (2020) and Zou et al. (2021), we can regard 1-hop neighbors of a given focus as candidates and predict a valid alternative from these candidates. It is based on the fact that the corresponding reply perspectives can be shared if posts containing the same focus have similar semantics.

We build the shift graph $\mathcal{G}$ with two steps: vertex construction and edge construction. For vertex construction, we first exploit a rule-based keyword extraction method (Campos et al., 2020) to identify salient keywords from utterances in the observed dialogue dataset $\mathcal{D}$. To further identify the focus $\boldsymbol{c}$ from all keywords of $\boldsymbol{x}$, we use guidance from the future information (i.e., response) to select the keyword that is semantically closest to $\boldsymbol{y}$. To identify the reply perspective $\boldsymbol{z}$, we select the keyword with the closest semantics to $\boldsymbol{c}$. More concretely, we use cosine similarity between their embedding via BERT (Devlin et al., 2019) as the measure of semantic closeness, where each embedding is achieved by taking the average of the hidden state of each token. For edge construction, we build an edge by connecting $\boldsymbol{c}$ with $\boldsymbol{z}$. In this way, we characterize all shift associations in $\mathcal{D}$.

Once the shift graph is built, we predict $\tilde{\boldsymbol{z}}$ as

$$\tilde{\boldsymbol{z}} = \arg\max_{\boldsymbol{z}} P(\boldsymbol{Z}|\boldsymbol{C}=\tilde{\boldsymbol{c}}, \boldsymbol{X}=\boldsymbol{x}, \boldsymbol{N}=\mathcal{N}(\tilde{\boldsymbol{c}})), \tag{2}$$

which is given by a trained reply perspective predictor. Note that $\tilde{\boldsymbol{c}}$ can be any keyword in the post $\boldsymbol{x}$ and $\mathcal{N}(\tilde{\boldsymbol{c}})$ denotes 1-hop neighbors of $\tilde{\boldsymbol{c}}$.

3. Prediction. This step is to generate the counterfactual response given the posterior sample $\boldsymbol{u}_t = [u_{t1}, \ldots, u_{t|V|}]$. Specifically, when generating the t-th token of the counterfactual response, our counterfactual generator computes the categorical distribution as follows,

$$\begin{aligned} \tilde{p}_{tk} &= P(Y_t = k|\boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Z}=\tilde{\boldsymbol{z}}, \boldsymbol{Y}_{<t}=\tilde{\boldsymbol{y}}_{<t}), \\ \tilde{y}_t &= \arg\max_{k=1,\ldots,|V|} (\log \tilde{p}_{tk} + u_{tk}), \end{aligned} \tag{3}$$

where $\tilde{\boldsymbol{z}}$ is the predicted reply perspective, $\tilde{\boldsymbol{y}}_{<t}$ is the token sequence generated in the previous step.

Overall, counterfactual generation via perspective transition can be used as an effective data augmentation method for open-domain dialogues to augment responses with wider semantic coverage. We show this method in Algorithm 1. The algorithm takes an observed sample $(\boldsymbol{x}, \boldsymbol{y})$ as an input and loop through every keyword of $\boldsymbol{x}$ as a different focus $\tilde{\boldsymbol{c}}$. For each $\tilde{\boldsymbol{c}}$, to sample multiple corresponding reply perspectives, we equally divide the candidate set $\mathcal{N}(\tilde{\boldsymbol{c}})$ into K sub-sets $\{\mathcal{N}_1(\tilde{\boldsymbol{c}}), \ldots, \mathcal{N}_K(\tilde{\boldsymbol{c}})\}$ for nested loop. At each iteration it predicts a different $\tilde{\boldsymbol{z}}$ for perspective transition to output a counterfactual sample $(\boldsymbol{x}, \tilde{\boldsymbol{y}})$.

3.2 Model Training

CAPT relies on the reply perspective predictor and the counterfactual generator, which greatly influence the quality of augmentation. Inspired by Yang et al. (2020); Schick and Schütze (2021), we choose a pre-trained encoder-decoder model BART (Lewis et al., 2020) as the backbone model.

Reply Perspective Predictor. We fine-tune BART on $\mathcal{D}$ to learn $P(\boldsymbol{Z}|\boldsymbol{C}, \boldsymbol{X}, \boldsymbol{N})$. In particular, the input is a concatenated text sequence consisting of the post $\boldsymbol{X}$, the focus $\boldsymbol{C}$, and the candidates $\boldsymbol{N}$. The output is the predicted reply perspective $\boldsymbol{Z}$. We maximize the objective as follows,

$$\mathcal{L}_p = -\sum_{t=1}^{|Z|} \log P(Z_t|[\boldsymbol{C}, \boldsymbol{X}, \boldsymbol{N}], \boldsymbol{Z}_{<t}), \tag{4}$$

where the bracket $[\cdot, \cdot, \cdot]$ denotes concatenation with the token [SEP]. The candidates $\boldsymbol{N}$ are also concatenated with commas. $\boldsymbol{Z}_{<t}$ is a prefix of the reply perspectives. $|Z|$ denotes the length of $\boldsymbol{Z}$.

Counterfactual Generator. We fine-tune BART on $\mathcal{D}$ to learn $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$. Specifically, the generator is trained to generate the response $\boldsymbol{Y}$ with

Algorithm 1: Data Augmentation

Input: $(\boldsymbol{x}, \boldsymbol{y})$: An observed sample
$\mathcal{C}$: All keywords $\{\tilde{\boldsymbol{c}}_1, \ldots, \tilde{\boldsymbol{c}}_{|\mathcal{C}|}\}$ of $\boldsymbol{x}$
$\mathcal{G}$: The shift graph

Output: A counterfactual sample $(\boldsymbol{x}, \tilde{\boldsymbol{y}})$

Get the observed reply perspective $\boldsymbol{z}$;
for $i \leftarrow 1$ **to** $|\mathcal{C}|$ **do**
  Get 1-hop neighbors $\mathcal{N}(\tilde{\boldsymbol{c}}_i)$ from $\mathcal{G}$
  Remove $\boldsymbol{z}$ from $\mathcal{N}(\tilde{\boldsymbol{c}}_i)$
  Equally divide $\mathcal{N}(\tilde{\boldsymbol{c}}_i)$ into $\{\mathcal{N}_1(\tilde{\boldsymbol{c}}_i), \ldots, \mathcal{N}_K(\tilde{\boldsymbol{c}}_i)\}$
  for $j \leftarrow 1$ **to** K **do**
    $\tilde{\boldsymbol{y}} \leftarrow \texttt{Trans}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \tilde{\boldsymbol{c}}_i, \mathcal{N}_j(\tilde{\boldsymbol{c}}_i))$
Function $\texttt{Trans}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \tilde{\boldsymbol{c}}, \mathcal{N}(\tilde{\boldsymbol{c}}))$:
  Infer $\boldsymbol{u}$ from $P(\boldsymbol{U}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$
  Predict $\tilde{\boldsymbol{z}}$ from $P(\boldsymbol{Z}|\boldsymbol{x}, \tilde{\boldsymbol{c}}, \mathcal{N}(\tilde{\boldsymbol{c}}))$
  Reason $\tilde{\boldsymbol{y}}$ from $P(\boldsymbol{Y}|\boldsymbol{x}, \tilde{\boldsymbol{z}})$ under the current environment $\boldsymbol{u}$
  return $\tilde{\boldsymbol{y}}$

the input prompt consisting of the post $\boldsymbol{X}$ and the reply perspective $\boldsymbol{Z}$. Similarly, we maximize the following objective:

$$\mathcal{L}_g = -\sum_{t=1}^{|Y|} \log P(Y_t|[\boldsymbol{X}, \boldsymbol{Z}], \boldsymbol{Y}_{<t}), \quad (5)$$

3.3 Bi-directional Perplexity Selection

Filtering out detrimental augmented samples can improve downstream performance (Bras et al., 2020). Existing methods (Axelrod et al., 2011; Xie et al., 2020; Zhang et al., 2020a) pick out samples that the model only trained on the observed data is most confident about. However, these models have only seen limited samples so they may not identify valid but unseen samples from the counterfactual-generated data. Inspired by Lee et al. (2021), we leverage a large-scale dialogue pre-trained language model DialoFlow (Li et al., 2021), utilizing its powerful ability of transfer learning. Since large-scale dialogues have been seen, it can identify valid but unseen samples like "an expert" via perplexity (PPL) scores. Nonetheless, the resulting samples might contain samples with generic responses. Inspired by Li et al. (2016), we further introduce backward PPL to rerank responses for prioritizing those valid and interesting samples.

Specifically, we independently fine-tune DialoFlow to learn $P(\boldsymbol{Y}|\boldsymbol{X})$ and $P(\boldsymbol{X}|\boldsymbol{Y})$ on $\mathcal{D}$ for calculating *forward* and *backward* PPL scores. Once we obtain the forward PPL scores for all samples, we find the best threshold η that separates valid samples from invalid samples. Inspired by Lee et al. (2021), we leverage the validation set to find the optimal single threshold parameter η, where we regard observed samples from the validation set as valid samples, and invalid samples are constructed by replacing the responses of valid samples with randomly-sampled responses. Furthermore, we rerank the responses of each post in the valid samples via backward PPL scores. Since the higher the backward PPL score, the more likely the response is dull (Li et al., 2016), we choose samples in order from low to high until the desired number of augmented samples are obtained.

4 Experimental Setup

4.1 Settings

The experiments are conducted on the Chinese Weibo corpus (Zhang et al., 2020a). Specifically, the dataset $\mathcal{D}$ contains training, validation, and test sets with 300K, 5K, and 10K post-response samples, respectively. Please see Appendix B for more details on data and method implementations.

4.2 Baselines

We compare CAPT with a set of baselines: (1) **Observed**, which only uses the observed data to fine-tune dialogue models. (2) **Augmented**, which only uses our augmented data to fine-tune dialogue models. (3) **Back-Trans** (Sennrich et al., 2016), which back-translates responses via Google Translate. (4) **MLM** (Cai et al., 2020), which fine-tunes the BERT-large model on $\mathcal{D}$ to substitute some words of responses. The substituting probability is 0.15. (5) **DL** (Zhang et al., 2020a), which constructs post-response pairs where both post and response are retrieved from the unpaired data. Augmented dialogues are further filtered by their ranking module. (6) **BM25** (Gangal et al., 2021), which uses the BM25 algorithm to retrieve the top-k similar post to the observed post, and the corresponding response of the retrieved post is regarded as the augmented response. (7) **BART** (Lewis et al., 2020), which fine-tunes the BART-large model that takes the post as the input to generate responses with different decode strategies, including greedy search, sampling with temperature 0.5, and top-k sampling (k=10,25). They are denoted as **BART-greedy**, **BART-samp**, **BART-k10**, and **BART-k25**, respectively. Augmented pairs generated by BM25 and

Focus	Reply Perspective	Counterfactual Response
Post: I am *sleepless* because of *coughing*. (咳嗽睡不着)		
sleepless (睡不着)	sleep (睡)	I *slept* badly. (我是没睡好。)
coughing (咳嗽)	doctor (医生)	Have you seen the *doctor* already?(去看医生了吗?)
	cold (感冒)	Honey, do you have a *cold* too?(亲爱的，你是不是也感冒了？)
Post: I have a *stomachache* every day. (最近每天胃痛唉)		
stomachache (胃痛)	spicy (辣的)	You can't eat *spicy* food. (不能吃辣的)
	check (检查)	You need to *check* your body. (去检查下吧)
	serious (严重)	What happened? Why is it so *serious*? (搞什么那么严重)

Figure 3: Real cases showing the generation process of responses with different semantics.

BART are filtered by our data selection method.

4.3 Evaluation Metrics

Automatic Evaluation. The following metrics are used to automatically evaluate *retrieval-based* models. (1) Mean Average Precision (**MAP**): the average of Average Precision (AP) over test samples. AP is the average of precision scores at the ranks where references are found; (2) $\boldsymbol{R_{10}@k}$: the percentage of references among the top-k selected responses (k=1,2,5) when given 10 candidates in total. The following metrics are used to evaluate *generation-based* models. (1) **BLEU**: the overlap of n-grams (n<4) between the generated response and the reference. (2) **Dist-n**: the ratio of unique n-grams (n=1,2) over all n-grams in the generated responses, which measures the n-gram diversity. As we sample 3 responses for each test post, evaluation is performed both within and among the sampled responses. **Intra-Dist** calculates that ratio within each sampled response, and **Inter-Dist** calculates that ratio among all three responses. (3) $\mathbf{BS}_f$: the F1-value of **BERTScore** (Zhang et al., 2020b), which measures the semantic similarity between each 2 responses in 3 sampled responses. Lower scores imply greater semantic diversity.

We also use **Dist-n** and $\mathbf{BS}_f$ to automatically evaluate the quality of augmented data, which evaluates the diversity among the generated responses. In addition, we introduce the following metrics to evaluate the diversity with respect to the original response. (1) **Novelty-n**: the ratio of new n-grams (n=1,2) in the augmented responses. **Intra-Novelty** similarly calculates the ratio within each augmented response, i.e., n-grams that are covered by the augmented response but not in the original response. **Inter-Novelty** calculates the ratio within the three augmented responses. (2) $\mathbf{BS}_{fo}$: the F1-value of **BERTScore**, which measures the semantic similarity between the augmented response and its corresponding original response.

Manual Evaluation. The following metrics are used to manually evaluate the quality of augmented data and generation-based models. Three annotators are employed to rate the samples. (1) **Fluency (Flu.)**: is the response fluent? (2) **Coherence (Coh.)**: is the response serve as a valid continuation of the preceding post? (3) **Interesting (Int.)**: is the response generic? (4) **Richness (Rich.)**: do the three sampled responses express different semantics? The rating scale is of 0 to 2, in which 0 means worst and 2 best.

5 Results and Discussion

5.1 Evaluating Augmented Data

We first evaluate the quality of augmented data. Specifically, we respectively select 900K augmented post-response pairs generated by these methods, on which automatic evaluation is performed. We further conduct manual evaluation on 600 samples, which contain 200 randomly-sampled posts and each post has 3 corresponding responses. The inter-annotator agreement is measured via the Fleiss's kappa κ (Randolph, 2005). The κ values for *Fluency*, *Coherence*, *Interesting* and *Richness* are 0.67 (moderate agreement), 0.46 (moderate agreement), 0.64 (moderate agreement) and 0.69 (moderate agreement), respectively.

The results are shown in Table 1 and 2, which indicates that our augmented data outperforms all the baselines. We further observe that: (1) Our augmented data achieve similar scores as the observed data over all the metrics, which indicates that our augmented data is high-quality. We present some cases of the augmented data to show the gen-

Method	Intra-Dist-1,2		Inter-Dist-1,2		BS_f	Intra-Novelty-1,2		Inter-Novelty-1,2		BS_{fo}
BART-greedy	93.34‡	98.37‡	64.83‡	81.81‡	66.46‡	84.42‡	95.54‡	60.54‡	80.38‡	58.12‡
BART-samp	94.14‡	98.79‡	70.85‡	89.27‡	63.60‡	84.24‡	95.99‡	65.84‡	87.74‡	58.11‡
BART-k10	93.08‡	98.63‡	70.60‡	90.07‡	63.15‡	85.00†	96.23†	67.36‡	89.11‡	58.08‡
BART-k25	93.74‡	98.77‡	74.63‡	91.98‡	61.61‡	85.76	96.43	71.01‡	90.90‡	57.83‡
CAPT	**94.64**	**98.90**	**79.91**	**94.79**	**59.59**	**85.84**	**96.63**	**74.47**	**92.98**	**57.31**‡
Observed	94.05	98.90	-	-	-	-	-	-	-	-

Table 1: Automatic evaluation on the quality of augmented data generated by different generation-based methods. The bottom row corresponds to the high-quality observed dialogue data in $\mathcal{D}$. Significance tests between CAPT and baselines are performed using t-test. † and ‡ indicate p-value < 0.05 and 0.01, respectively.

Method	Flu.	Coh.	Int.	Rich.
BART-greedy	1.921	1.507	1.222‡	0.611‡
BART-samp	1.833‡	1.383‡	1.500‡	0.926‡
BART-k10	1.853‡	1.461†	1.506‡	0.983‡
BART-k25	1.813‡	1.333‡	1.560†	1.182‡
CAPT	**1.953**	**1.653**	**1.707**	**1.660**
Observed	1.941	1.744	1.740	-

Table 2: Manual evaluation on augmented data. The bottom row corresponds to the high-quality observed dialogue data in $\mathcal{D}$. Significance tests between CAPT and baselines are performed using t-test. † and ‡ indicate p-value < 0.05 and 0.01, respectively.

Method	MAP	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
Observed	80.21‡	69.72‡	82.05‡	94.96†
Augmented	76.67‡	65.14‡	78.16‡	92.46‡
MLM	80.22‡	69.76‡	82.05‡	94.90†
Back-Trans	80.26‡	69.75‡	82.21‡	94.99
DL	80.47‡	70.05‡	82.41†	95.03
BM25	80.07‡	69.68‡	81.62‡	94.82‡
BART-greedy	80.37‡	70.03‡	82.17‡	94.75‡
BART-samp	80.42‡	70.17‡	82.03‡	94.88‡
BART-k10	80.38‡	70.06‡	82.15‡	94.79‡
BART-k25	80.53‡	70.30‡	82.21‡	94.91†
CAPT	**81.08**	**71.08**	**82.86**	**95.14**

Table 3: Automatic evaluation on different data augmentation methods for retrieve-based models. We repeatedly experiment 10 times with different seeds and report the averaged scores. † and ‡ indicate that the improvement of CAPT is significant at the level of 0.05 and 0.01 respectively (significance tests via t-test).

eration process of different-semantic responses in Figure 3. (2) Our augmented data achieve better scores of BS_f, BS_{fo} and Richness, which indicates that CAPT can augment more responses with different semantics. In particular, BART-samp vs. CAPT shows the effectiveness of intervention in the reply perspective. (3) BART-k10 achieves relatively good scores on all the metrics compared to other baselines. This indicates that the top-k sampling (k=10) is superior to the other decoding strategies. Thus, the top-k sampling (k=10) can be used for the following generation-based models.

5.2 Evaluating Dialogue Model

We further evaluate the benefit of our augmented data on retrieve-based and generation-based dialogue models. Specifically, we follow Zhang et al. (2020a) and select 300K augmented post-response samples for all methods for a fair comparison. We conduct automatic evaluation on 5K test data and manual evaluation on 600 samples that contain 200 randomly-sampled posts with 3 generated responses. The κ value for *Fluency*, *Coherence*, *Interesting* and *Richness* are 0.67 (moderate agreement) are 0.71 (substantial agreement), 0.59 (moderate agreement), 0.48 (moderate agreement) and 0.53(moderate agreement), respectively.

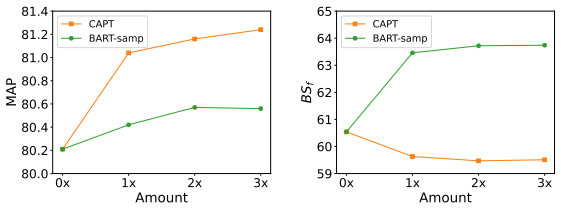
The results on retrieve-based and generation-based models are respectively shown in Table 3 and 4, which indicates that CAPT outperforms all the baselines on almost all the metrics for both dialogue models. This confirms the effectiveness of augmenting valid responses with different semantics. We can further observe that: (1) CAPT achieves higher scores for almost all the metrics compared to other BART-based methods, especially BART-samp. This demonstrates that intervention in the reply perspective is effective for improving the performance of dialogue models. (2) CAPT achieves higher BS_f and Richness ratings but a relatively lower BLEU score. We speculate that augmenting more semantically different samples enables dialogue models to generate more responses that differ from references.

5.3 Further Discussion

Further, we also investigate the impact of the amount of augmented responses and the effect of each component of CAPT.

Method	BLEU	Intra-Dist-1,2		Inter-Dist-1,2		BS_f	Flu.	Coh.	Int.	Rich.
Observed	2.22	91.11‡	98.21‡	73.83‡	93.18†	60.54‡	1.806‡	1.377‡	1.645	1.075‡
Augmented	1.85	92.29‡	98.16‡	77.86	93.28	59.76	1.848	1.363‡	1.652	1.320
MLM	2.16	91.19‡	98.25‡	74.41‡	93.37	60.50‡	1.813†	1.438	1.653	1.095‡
Back-Trans	2.21	91.26‡	98.26‡	74.66‡	93.49	60.45‡	1.791‡	1.443	1.657	1.115‡
DL	2.23	92.09‡	98.35‡	75.02‡	93.42	60.35‡	1.823†	1.462	1.665	1.135‡
BM25	1.68	91.55‡	98.14‡	76.51‡	92.02‡	60.17†	1.803‡	1.155‡	1.650	1.185†
BART-greedy	**3.54**	91.54‡	98.02‡	64.79‡	80.87‡	67.18‡	1.841	1.453	1.508‡	0.895‡
BART-samp	2.86	92.12‡	98.42‡	69.81‡	88.91‡	63.51‡	1.822†	1.448	1.582‡	0.910‡
BART-k10	2.72	91.71‡	98.53	70.51‡	90.02‡	63.45‡	1.835	1.480	1.584‡	0.925‡
BART-k25	2.70	91.93‡	98.45‡	71.29‡	90.46‡	62.81‡	1.812†	1.425†	1.623†	0.935‡
CAPT	2.11	**93.39**	**98.67**	**78.03**	**93.62**	**59.64**	**1.867**	**1.492**	**1.677**	**1.355**

Table 4: Automatic and manual evaluation on different data augmented methods for generation-based dialogue models. Significance tests between CAPT and baselines were performed using t-test, where booststrap resampling (Koehn, 2004) was applied for automatic evaluation. † and ‡ indicate p-value < 0.05 and 0.01, respectively.



(a) Retrieve-based Models (b) Generation-based Models

Figure 4: Performance changes on retrieve-based and generation-based dialogue models respectively by providing different amounts of augmented data generated by CAPT and BART-sampling. We use MAP and BS_f metrics to evaluate corresponding models.

Method	MAP	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
CAPT	81.08	71.08	82.86	95.14
-Predictor	80.63‡	70.33‡	82.65	94.96†
-Candidate	80.22‡	69.90‡	82.01‡	94.62‡
-Selection	80.41‡	69.92‡	82.44†	95.08
-Dial PLM	80.52‡	70.19‡	82.39†	94.98†
-Back PPL	80.68‡	70.41‡	82.51†	95.07
-Gumbel	80.83†	70.62†	82.76	95.02

Table 5: Ablation study on different components of CAPT on retrieve-based dialogue models. We repeatedly experiment 10 times with different seeds and report the averaged scores. † and ‡ indicate that the performance drop is significant at the level of 0.05 and 0.01 respectively (significance tests via t-test).

The Impact of Amount. We select 0x, 1x, 2x, 3x the amount of training samples to assess the impact of providing more responses and compare CAPT with the baseline, i.e., BART-samp. Note that 3x represents that 3*300K augmented post-response samples are selected. Considering that samples selected in order have different interesting degrees, we eliminate the impact of interesting by uniformly selecting 900K augmented samples and randomly select from them. The results are shown in Figure 4. We can observe that: (1) The MAP score on BART-samp reaches a peak at 2x and drops afterward, and BS_f keeps increasing from 0x to 3x augmentation. We speculate that BART-samp only outputs alternative expressions with diversified words, which have limited semantic differences. Augmentation of similar samples at high amounts would negatively affect training. (2) However, the MAP score on CAPT keeps increasing and BS_f does not increase. This indicates that CAPT can augment responses with different semantics, and providing more semantically different responses can further improve the performance of downstream tasks.

Ablation Study. We perform the following ablation tests to validate the effect of each component: (1) Randomly choose a keyword from candidates as the reply perspective without the prediction step (-Predictor); (2) Only take the post and the focus as the input to the predictor without 1-hop neighbors as candidates (-Candidate); (3) Do not filter out the augmented data via data selection (-Selection); (4) Leverage a general pre-trained language model GPT2, which does not see enough dialogue samples, to replace the dialogue pre-trained language model DialoFlow (-Dial PLM); (5) Only use the forward PPL scores to filter out invalid samples without ranking via the backward PPL scores (-Back PPL). (6) Generate responses not under the current environment, i.e, without the posterior Gumbel noises (-Gumbel). The results are shown in Table 5. We observe that ablating each component brings varying degrees of performance drop. This demonstrates the necessity of designing all these

components.

6 Related Work

Data Augmentation. Data augmentation has been widely used in various NLP tasks and surveyed by Shorten and Khoshgoftaar (2019); Wen et al. (2021); Feng et al. (2021); Ni et al. (2021); Chen et al. (2021). Overall, data augmentation methods either add slightly modified copies of existing data or create synthetic data. Some work propose to use heuristic rules (Du and Black, 2018) or paraphrasing-based methods (Niu and Bansal, 2019; Li et al., 2019; Cai et al., 2020; Zhang et al., 2020a; Xie et al., 2022; Cao et al., 2022). Another line of work (Chang et al., 2021; Yang et al., 2020; Schick and Schütze, 2021; Wang et al., 2021; Zheng et al., 2022) is exploiting large-scale pre-trained language models for data augmentation. However, these existing methods do not focus on creating semantically different responses.

Semantically Different Augmentation. Gangal et al. (2021) utilizes knowledge sources, including COMET (Bosselut et al., 2019) and corpus retrieval (Robertson et al., 1994) to augment semantically diverse references for dialogue evaluation. Both methods only pre-define limited augmented perspectives. In contrast, CAPT obtains richer reply perspectives by building a shift graph.

Counterfactual Inference. Our work is based on counterfactual inference (Pearl et al., 2000), which has shown promising results in various NLP tasks, including question answering (Paranjape et al., 2022; Yu et al., 2021), machine translation (Liu et al., 2021) and story generation (Qin et al., 2019; Hao et al., 2021; Chen et al.). In particular, Zhu et al. (2020) uses counterfactual inference for response generation, which explores potential responses via counterfactual off-policy training. However, CAPT focuses on *counterfactual data augmentation*, which can be used to improve the performance of multiple downstream tasks.

Graph Construction. Some researches (Xu et al., 2020; Zou et al., 2021) also build a graph to manage concept shifts for response generation, which aims to form a more coherent and controllable dialogue. In contrast, CAPT builds a shift graph to predict valid substituted reply perspectives, which are used to augment responses with different semantics. Due to the different purposes of use, our graph construction is different from these existing works.

7 Conclusion

This paper presents a counterfactual data augmentation method, CAPT, to augment more responses with different semantics for a given dialogue history. Specifically, CAPT employs counterfactual inference to generate counterfactual responses by intervening in the observed reply perspective, which replaces with different reply perspectives for generating semantically different responses. Experimental results show that CAPT can augment high-quality responses with different semantics, which can be further used to improve the performance of downstream tasks. In future work, we plan to explore an appropriate training strategy for further preventing dialogue models from being affected by noises in our augmented data, and extend CAPT on multi-turn dialogues. We hope that CAPT will encourage future research for other generation tasks.

Limitations

CAPT works well in scenarios with a certain amount of observed data. A small amount of observed data would lead to a small-scale shift graph. Thus, it is difficult to provide enough candidates to pick out more valid reply perspectives, and then augment sufficient valid post-response samples. In addition, CAPT may be more suitable for open-domain dialogue augmentation in some languages that require good-quality keyword extraction methods and pre-trained models for that language. e.g., Chinese and English. When transferred to different languages, e.g., English, the modifications are required as follows: (1) use the English-version keyword extraction method and keyword/sentence encoder when building the graph; (2) use the English-version pre-trained model as the backbone model for the reply perspective predictor and the counterfactual generator.

Ethics Statement

In this work, we employ three annotators to manually evaluate the quality of augmented data and generation-based dialogue models. We pay $0.2 to each annotator for each sample.

Acknowledgement

We sincerely thank the anonymous reviewers for their thorough reviewing and valuable suggestions.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779. Association for Computational Linguistics.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1078–1088. PMLR.

Lars Buesing, Theophane Weber, Yori Zwols, Nicolas Heess, Sébastien Racanière, Arthur Guez, and Jean-Baptiste Lespiau. 2019. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *7th International Conference on Learning Representations, ICLR 2019*.

Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. 2020. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, pages 257–289.

Yu Cao, Wei Bi, Meng Fang, Shuming Shi, and Dacheng Tao. 2022. A model-agnostic data manipulation method for persona-based dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Ernie Chang, Xiaoyu Shen, Dawei Zhu, Vera Demberg, and Hui Su. 2021. Neural data-to-text generation with LM-based text augmentation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2021. An empirical survey of data augmentation for limited data learning in nlp. *ArXiv*, abs/2106.07499.

Jiangjie Chen, Chun Gan, Sijie Cheng, Hao Zhou, Yanghua Xiao, and Lei Li. Unsupervised editing for counterfactual stories. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10473–10481.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Wenchao Du and Alan Black. 2018. Data augmentation for neural online chats response selection. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, Brussels, Belgium. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988. Association for Computational Linguistics.

Tingchen Fu, Shen Gao, Xueliang Zhao, Ji rong Wen, and Rui Yan. 2022. Learning towards conversational ai: A survey. *AI Open*, pages 14–28.

Varun Gangal, Harsh Jhamtani, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Improving automated evaluation of open domain dialog via diverse reference augmentation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4079–4090. Association for Computational Linguistics.

Changying Hao, Liang Pang, Yanyan Lan, Yan Wang, Jiafeng Guo, and Xueqi Cheng. 2021. Sketch and customize: A counterfactual story generator. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12955–12962.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245. Association for Computational Linguistics.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.*, pages 21:1–21:32.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

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.

Juntao Li, Lisong Qiu, Bo Tang, Dongmin Chen, Dongyan Zhao, and Rui Yan. 2019. Insufficient data can also rock! learning to converse using smaller data with augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6698–6705.

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.

Qi Liu, Matt Kusner, and Phil Blunsom. 2021. Counterfactual data augmentation for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 187–197. Association for Computational Linguistics.

R. Duncan Luce. 1959. *Individual Choice Behavior: A Theoretical analysis*. Wiley, New York, NY, USA.

Chris J Maddison, Daniel Tarlow, and Tom Minka. 2014. A* sampling. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, Vinay Adiga, and Erik Cambria. 2021. Recent advances in deep learning based dialogue systems: A systematic survey. *arXiv preprint arXiv:2105.04387*.

Tong Niu and Mohit Bansal. 2019. Automatically learning data augmentation policies for dialogue tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics.

Michael Oberst and David Sontag. 2019. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. PMLR.

Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. 2022. Retrieval-guided counterfactual generation for QA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1670–1686. Association for Computational Linguistics.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. Causal inference in statistics: A primer. 2016. *Internet resource*.

Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19:2.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053. Association for Computational Linguistics.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *TREC*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*.

Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. 2021. Time series data augmentation for deep learning: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4653–4660. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Han Xiao. 2018. bert-as-service. `https://github.com/hanxiao/bert-as-service`.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc.

Shufang Xie, Ang Lv, Yingce Xia, Lijun Wu, Tao Qin, Tie-Yan Liu, and Rui Yan. 2022. Target-side input augmentation for sequence to sequence generation. In *International Conference on Learning Representations*.

Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Conversational graph grounded policy learning for open-domain conversation generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1835–1845. Association for Computational Linguistics.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025.

Sicheng Yu, Hao Zhang, Yulei Niu, Qianru Sun, and Jing Jiang. 2021. COSY: COunterfactual SYntax for cross-lingual understanding. 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 577–589. Association for Computational Linguistics.

Rongsheng Zhang, Yinhe Zheng, Jianzhi Shao, Xiaoxi Mao, Yadong Xi, and Minlie Huang. 2020a. Dialogue distillation: Open-domain dialogue augmentation using unpaired data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, and Minlie Huang. 2022. Augesc: Large-scale data augmentation for emotional support conversation with pretrained language models. *CoRR*.

Qingfu Zhu, Wei-Nan Zhang, Ting Liu, and William Yang Wang. 2020. Counterfactual off-policy training for neural dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3438–3448. Association for Computational Linguistics.

Yicheng Zou, Zhihua Liu, Xingwu Hu, and Qi Zhang. 2021. Thinking clearly, talking fast: Concept-guided non-autoregressive generation for open-domain dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2215–2226. Association for Computational Linguistics.

A Task Definitions

Response Selection. Given a dataset $\mathcal{D} = \{(\boldsymbol{x}^i, \boldsymbol{y}^i, l^i)_{i=1}^N\}$, the retrieval-based dialogue model learns a matching function to correctly identify the positive response from a set of negative responses. Specifically, the matching function $P_\theta(l^i|\boldsymbol{x}^i, \boldsymbol{y}^i)$ predicts whether the response $\boldsymbol{y}^i$ matches the dialogue history $\boldsymbol{x}^i$. $l^i \in \{0, 1\}$ denotes a matching label, which indicates that $\boldsymbol{y}^i$ is a proper response for $\boldsymbol{x}^i$ if $l^i = 1$, otherwise $l^i = 0$. The model parameters θ can be learned by minimizing the loss function that is formulated as

$$\begin{aligned}\mathcal{L}_{sel} = -\sum_{i=1}^{N} [l^i \log P_\theta(l^i = 1|\boldsymbol{x}^i, \boldsymbol{y}^i) \\ + (1 - l^i) \log P_\theta(l^i = 0|\boldsymbol{x}^i, \boldsymbol{y}^i)]\end{aligned} \tag{6}$$

Generally, the training negative responses are randomly selected from the dataset $\mathcal{D}$.

Response Generation. Given a dataset $\mathcal{D} = \{(\boldsymbol{x}^i, \boldsymbol{y}^i)_{i=1}^N\}$, the generation-based dialogue model learns to model the distribution $\mathcal{P}_\phi(\boldsymbol{y}^i|\boldsymbol{x}^i)$ of the response $\boldsymbol{y}^i$ given the dialogue history $\boldsymbol{x}^i$. The model parameters ϕ can be learned by minimizing the following loss:

$$\mathcal{L}_{gen} = -\sum_{i=1}^{N} \log P_\phi(\boldsymbol{y}^i|\boldsymbol{x}^i) \tag{7}$$

However, a dialogue dataset that admits multiple semantically different responses for each dialogue history is usually expensive to collect, as it requires annotators to write a large variety of valid responses. Although such a dataset can be crawled from social networks, it will contain many noisy and meaningless responses. It is also expensive to pick out sufficient high-quality dialogues that meet requirements. Thus, counterfactual data augmentation aims to further augment different-semantic responses $\tilde{\boldsymbol{y}}^i$ for $\boldsymbol{x}^i$ in $\mathcal{D}$ without manually collecting new data. In the following sections, we will omit the superscript i for simplicity.

B Experimental Details

B.1 Data

The experiments are conducted on the Chinese Weibo corpus (Zhang et al., 2020a). Specifically, the dataset $\mathcal{D}$ contains training, validation, and test sets with 300K, 5K, and 10K post-response samples, respectively. To build the shift graph, we apply YAKE (Campos et al., 2020) that relies on the statistical features of the text to automatically extract the most important keywords of each utterance in the training data. Keywords are limited to nouns, adjectives and verbs. The number of keyword vertices and edges are $77,439$ and $202,266$ respectively. Furthermore, we randomly sample 200 post-response samples and employ three human annotators to evaluate the appropriateness of both keywords of focus and reply perspective. About 86% keyword pairs are accepted by the annotators. The average number of candidate keywords at training and augmentation times are 102 and 124 respectively. After achieving augmented data, we similarly evaluate whether the responses share similar core semantics with the given reply perspectives. About 96.5% responses are accepted by the annotators.

B.2 Implementation Details

CAPT. For graph construction, we pursue bert-as-service (Xiao, 2018) to achieve the embedding by mapping a variable-length text sequence to a fixed-length vector. Our predictor and generator are independently fine-tuned on the BART-large model (Shao et al., 2021) using the loss in Eq. 4 and 5 for ten epochs, with the batch size of 64, the learning rate of 1e-5. The other hyper-parameter setting follows that of Shao et al. (2021). The maximum sequence length is set to 512. We thus limit the maximum candidate size of our predictor to 100. If the candidate size is greater than 100, we randomly sample 100 candidates. We then filter out those samples whose candidate size is less than 5. For data selection, we implement the score functions by fine-tuning the pre-trained DialoFlow (Li et al., 2021) model with $\mathcal{D}$ for two epochs, with the batch size of 64 and the learning rate of 1e-5. The best threshold η is 10.

At augmentation time, we also limit the range of the candidate size from 5 to 100. Thus, we divide the whole candidate set into K sub-sets and set the candidate size of each sub-set $N_{\tilde{c}} = \max(\min(\frac{|\mathcal{N}(\tilde{\boldsymbol{c}})|}{K}, 100), 5)$, where K is initialized by 20. We further update $K = \frac{|\mathcal{N}(\tilde{\boldsymbol{c}})|}{N_{\tilde{c}}}$. The predictor outputs reply perspectives with greedy search. The generator samples counterfactual responses from posterior Gumbel noises, the temperature is set to 0.5.

Retrieve-based Model. The retrieve-based model is built by fine-tuning the pre-trained BERT-base (Devlin et al., 2019) for two epochs,

with the learning rate of 2e-5, the batch size of 64, and the max sequence length of 512. we adopt the last checkpoint for evaluation.

Generation-based Model. The generation-based model is built by fine-tuning the pre-trained BART-large (Shao et al., 2021) for five epochs, with the learning rate of 1e-5, the batch size of 64, and the max sequence length of 512. At inference time, we use the top-k sampling (k=10), and the maximum decoded length is set to 50. we adopt the last checkpoint for evaluation.

Training and Evaluation. We train retrieve-based dialogue models with 4 GPUs, generation-based models with 8 GPUs, the reply perspective predictor with 8 GPUs, and the counterfactual generator with 8 GPUs. We use Nvidia Tesla V100 GPUs. The training time for retrieve-based models, generation-based models, the reply perspective predictor, and the counterfactual generator is approximately 2h, 4h, 4h and 5h, respectively. At augmentation time, it takes 55min to predict reply perspectives and 1h to generate counterfactual responses for all augmented samples. When calculating the forward and backward PPL scores, it takes 40min respectively.