Title
Unsupervised Editing for Counterfactual Stories

Permalink
https://escholarship.org/uc/item/3t54r6s6

Authors
Chen, Jiangjie
Gan, Chun
Cheng, Sijie
et al.

Publication Date
2021-12-10
Unsupervised Editing for Counterfactual Stories

Jiangjie Chen\textsuperscript{a\textdagger}, Chun Gan\textsuperscript{\dagger}, Sijie Cheng\textsuperscript{b}, Hao Zhou\textsuperscript{\dagger}, Yanghua Xiao\textsuperscript{b\textdagger\textdagger}, Lei Li\textsuperscript{\dagger\textdagger}

\textsuperscript{a}Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
\textsuperscript{b}ByteDance AI Lab, JD.com
\textsuperscript{\dagger}University of California, Santa Barbara
\textsuperscript{\dagger\textdagger}Fudan-Aishu Cognitive Intelligence Joint Research Center

\{jjchen19, sjcheng20, shawyh\}@fudan.edu.cn,
cgan5@wisc.edu, zhouhao.nlp@bytedance.com, lilei@cs.ucsb.edu

Abstract

Creating what-if stories requires reasoning about prior statements and possible outcomes of the changed conditions. One can easily generate coherent endings under new conditions, but it would be challenging for current systems to do it with minimal changes to the original story. Therefore, one major challenge is the trade-off between generating a logical story and rewriting with minimal-edits. In this paper, we propose EDUCAT, an editing-based unsupervised approach for counterfactual story rewriting. EDUCAT includes a target position detection strategy based on estimating causal effects of the what-if conditions, which keeps the causal invariant parts of the story. EDUCAT then generates the stories under fluency, coherence and minimal-edits constraints. We also propose a new metric to alleviate the shortcomings of current automatic and human evaluation. The resources of EDUCAT are available at: https://github.com/jiangjiechen/EDUCAT

1 Introduction

Counterfactual reasoning is a hypothetical thinking process to assess possible outcomes by modifying certain prior conditions. It is commonly known as “what-if” analysis — “what will happen if . . .”. It is a big challenge to build an intelligent system with counterfactual reasoning capabilities (Pearl 2009; Pearl and Mackenzie 2018). Counterfactual reasoning relies on the ability to find the causal invariance in data, i.e. the factors held constant with the change of conditions in a series of events (Slioman and Lagnado 2004).

In this paper, we study unsupervised counterfactual story rewriting, a concrete instance of counterfactual reasoning. We focus on unsupervised methods for this task, since humans do not need supervised learning to imagine alternative futures. The task is to create plausible alternative endings given small modifications to the story context.

\textsuperscript{\dagger}Work is done during internship at ByteDance AI Lab.
\textsuperscript{\dagger\textdagger}Corresponding authors.
\textsuperscript{\textdagger\textdagger}Work is done while at ByteDance AI Lab.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
event, an ideal system can generate a plausible ending with few edits to the variant events.

Most of the existing methods (Li, Ding, and Liu 2018; Xu et al. 2018; Guan, Wang, and Huang 2019; Guan et al. 2020) focus on the story generation in an auto-regressive manner. These approaches keep the story logical mainly by exploiting the language modeling ability of LMs such as the GPTs (Radford et al. 2018, 2019; Brown et al. 2020). Few of them (Qin et al. 2019, 2020) deal with the reasoning balance in counterfactual text generation, which requires balancing between coherence and minimal-edits. For example, Qin et al. (2020) propose to keep the balance by constraining the decoding on new endings with a sentence-level similarity scorer with the original ones. However, LMs are known to be hard to control, often leading to over-editing.

In this paper, we propose EDUCAT, an EDiting-based Unsupervised Counterfactual generATion method for counterfactual story rewriting. Given the original story and a modified condition statement, the challenge is to locate which part to retain (i.e. causal invariance) and which to modify (i.e. causal variance) while maintaining coherence to the context after editing. Inspired by causal analysis research (Hernández 2004), we quantify the potential outcome after intervention using the ratio between consistencies with the counterfactual and initial conditions, which can be computed by an off-the-shelf model. EDUCAT employs a Markov chain Monte Carlo sampling framework (Metropolis et al. 1953) for unsupervised generation by iteratively generating token modifications (Miao et al. 2019). With desired properties and guidance from the estimated potential outcome, EDUCAT generates fluent and coherent alternative story endings with minimal edits.

The contributions of this work are as follows:

• We first solve the counterfactual story rewriting task using unsupervised discrete editing method based on MCMC sampling.
• We draw inspiration from causal analysis and propose two counterfactual reasoning components that quantify the outcomes of context changes.
• We conduct experiments to verify that EDUCAT achieves the best trade-off between coherence and minimal-edits for unsupervised methods.

2 Task Formulation with Causal Model

In counterfactual story rewriting task, given a story consisting of a premise $z$, a story context $x$ and an ending $y$, we intervene by altering $x$ into a counterfactual context $x'$ and hope to predict new ending $y'$.

This problem naturally fits to be formulated with a Causal Model, a directed acyclic graph used to encode assumptions on the data generating process. As presented in the Figure 2, the left part shows a simple example of a causal model with treatment ($X$), effect ($Y$) and confounder ($Z$), respectively. In causal inference, a confounder is a random variable that influences both the treatment and effect variables, causing a spurious correlation (Pearl 2009). Note that in this problem, $z$ consists of both observed confounder $s_1$ and unobserved commonsense knowledge, where the latter is very difficult to explicitly model.

The counterfactual inference can be formulated with a $\text{do}$-operator. As shown in Figure 2, we can intervene on the $X$ variable by applying $\text{do}(X) = x'$ to set its value to the counterfactual without changing the rest. The arrow pointing from $Z$ to $X$ in the causal model is deleted since $X$ no longer depends on $Z$ after the intervention, resulting in a new graphical model. Consequently, the problem of counterfactual story generation can be formally restated as a counterfactual inference problem as follows: given $(z, x, y)$, what would the potential outcome of $y$ be if one changes the story context from $x$ to $x'$?

3 Proposed Approach: EDUCAT

In this section, we present an overview and details of EDUCAT. In general, the rewriting process works as follows: starting with an original full story, EDUCAT performs the following procedures iteratively:

1. Conflict Detection, it finds possible chunks in current story endings contradictory to counterfactual conditions;
2. Edits Proposal, it proposes an edited ending and decides its acceptance based on fluency and coherence scores.

The above steps repeat multiple rounds. Each proposal is either accepted or rejected based on desired properties $\pi(y)$, which is defined as the score product of each property score:

$$\pi(y) \propto \lambda^0_c(y) \cdots \lambda^n_c(y)$$

Finally, we pick the best one according to a ranking function as the output. An illustrative example is given in Figure 1.

However, the challenge remains for the quantification of these desired properties for ideal story rewriting. Inspired by causal analysis research, we can quantitatively calculate the difference of story endings’ quality given different conditions with the Causal Risk Ratio (CRR) (Hernández and Robins 2020). CRR is defined as follows:

$$\text{CRR} = \frac{P(Y = y \mid \text{do}(X = x'), Z = z)}{P(Y = y \mid \text{do}(X = x), Z = z)}$$

Figure 2: Formulating counterfactual story rewriting with intervention on causal model, where $z$ is the common premise of the story, $x, y$ denote the original story, and $x', y'$ are the counterfactual story.
The value goes up when the new ending is more consistent with the counterfactual condition. However, it is difficult to explicitly calculate both observed and unobserved confounders ($z^*$) in $P(Y = y|do(X = x))$ as follows:

$$P(Y = y|do(X = x)) = \sum_{z^*} P(Y = y|X = x, Z = z^*)P(Z = z^*)$$

We make a causal sufficiency assumption that only observed confounder ($z$) is considered:

$$P(Y = y|do(X = x)) = P(Y = y|X = x, Z = z)$$

(4)

So CRR can be calculated by

$$\text{CRR} = \frac{P(Y = y|X = x^*, Z = z)}{P(Y = y|X = x, Z = z)}$$

5)

In this way, we can roughly estimate the influence on possible endings brought by a changed condition. Next, we will elaborate on the details of EDUCAT.

### 3.1 Constrained Generation via MCMC

In EDUCAT, we direct the Markov chain Monte Carlo (MCMC) sampling process with counterfactual reasoning ability brought by conflict token detection and desired properties as sampling constraints.

EDUCAT directly samples from the sentence space with three local operations: token replacement, deletion and insertion. During sampling, after an edit position is found, the operation is randomly chosen with equal probability. Finally, the proposed new sentence will either be accepted or rejected according to the acceptance rate computed by desired properties $\pi(y)$. The above process is repeated till convergence.

Specifically, Metropolis-Hasting sampling (MH) algorithm moves the current sentence $y_t$ to the next sentence $y_{t+1}$ by generating from the proposal distribution $g(y_{t+1}|y_t)$ and accepting it based on an acceptance rate. The sample distribution in MCMC will converge to the stationary distribution $\pi(y)$ in the Markov chain under mild conditions. The acceptance rate $\alpha$ at the $t$-th iteration is defined as follows,

$$\alpha(y_{t+1}|y_t) = \min \left\{1, \frac{\pi(y_{t+1})g(y_t|y_{t+1})}{\pi(y_t)g(y_{t+1}|y_t)} \right\}$$

(6)

$T$ is a temperature controlled by a cooling schedule $T = 0.95^{t/4}$ in our implementation.

Next, we will describe in detail the design of stationary distribution $\pi(y)$ (3.2) and transition proposal distribution $g(y_{t+1}|y_t)$ (3.3).

### 3.2 Desired Properties for Story Rewriting

Aside from the basic fluency property, the original CGMH framework is designed with properties such as similarity and keywords constraints. These simple properties cannot directly sample with counterfactual reasoning ability. Instead, we want the generated new endings to be not only fluent in terms of storytelling, but also logically coherent with $X'$ instead of $X$. In EDUCAT, we define two score functions in story rewriting, namely, a fluency score function $\lambda_{LM}$ and a coherence score function $\lambda_{Coh}$. Thus, the stationary distribution $\pi(y)$ is defined as the product of fluency score and the coherence score as follows:

$$\pi(y) \propto \lambda_{LM}(y) \cdot \lambda_{Coh}(y)$$

(7)

**Fluency Score** We compute the probability of the generated ending based on a pre-trained language model, e.g., GPT-2 (Radford et al. 2019). This is important and in line with previous work to guarantee the fluency and readability of the generated sentence. The likelihood is computed autoregressively as:

$$\lambda_{LM}(y^*) = \prod_{i=1}^{N} P_{LM}(y_i^*|z, x', y_{<i}^*)$$

(8)

We denote $y^*$ as the proposed ending at the current stage, and $y_{<i}^*$ as the $i$-th token in the ending.

**Coherence Score** Intuitively, we want to punish proposed endings contradictory to the counterfactual conditions but consistent with the initial ones. Therefore, the purpose of coherence score function $\lambda_{Coh}$ is to encourage the model to rewrite the original endings. The value of $\lambda_{Coh}$ should be larger than 1 if the generated ending is more causally related to counterfactual context than the initial one. Inspired by the definition of Causal Risk Ratio, the coherence score function $\lambda_{Coh}$ is defined as follows:

$$\lambda_{Coh}(y^*) = \frac{P_{Coh}(Y = y^*|z, x')}{P_{Coh}(Y = y^*|z, x)}$$

(9)

where the formulation for $P_{Coh}$ is fit for any model for quantification that measures the coherence between an ending and a story context. In our implementation, we employ conditional sentence probability calculated by a pre-trained language model (e.g., a GPT-2) to measure the coherence within a story in an unsupervised way. Note that we hope to solve this task in an unsupervised way. But $P_{Coh}$ is fully extendable for better story coherence checking models.

### 3.3 Editing Proposal Design

Regularized by the desired properties, we can make editing proposals by solving two questions: 1) Where to edit? and 2) Edit with what?

**Where to Edit: Conflict Detection** It is critical to know where to edit the original stories to write natural counterfactual stories with only minimal edits. Namely, we need to identify tokens that contradict with the counterfactual context (Hao et al. 2021). Meanwhile, causal invariant information is kept in the unchanged tokens.

Also inspired by the calculation of Causal Risk Ratio, we estimate the potential outcome of changing the contexts to find the most likely contradictory tokens. Let $y^*$ be the current ending to edit (initialized with $y$) and $y_i^*$ be the tokens, we define the conflicting probability $P_{cf}(y_i^*)$ on the $i$-th token in $y^*$ as follows,

$$P_{cf}(y_i^*) = \text{softmax}\left(\frac{P_{LM}(y_i^*|z, x, y_{<i}^*)}{P_{LM}(y_i^*|z, x', y_{<i}^*)}\right)$$

(10)
The token-level likelihood is computed via a language model. According to the definition, $P_{\text{MLM}}(y^*_m)$ is larger if $y^*_m$ is more causally related to the initial context than the counterfactual one. Those tokens are more likely to contradict with counterfactual conditions at each iteration. They should have a higher priority to be edited.

**Edit with What: Modification Action** We randomly sample from three token-level modification actions (replacement, deletion, and insertion) with equal probability to find what to use to edit the endings given editing positions. Let $y_t$ be the current sentence, the proposal distribution is defined as $g(y_{t+1}|y_t)$. The expectation of transition proposal from $y_t$ to $y_{t+1}$ is given by

$$g(y_{t+1}|y_t) = \frac{1}{3} \sum_{op \in \{r,d,i\}} g_{op}(y_{t+1}|y_t)$$

(11)

where $g_r$, $g_d$, $g_i$ correspond to the replacement, deletion and insertion proposals, respectively. For replacement, let $y_t = [w_1, \ldots, w_m, \ldots, w_n]$, the replacement action replaces the token $w_m$ with $w^c$, where $w^c$ is sampled from a pre-selected candidate set $Q$. Let $y_{t+1} = [w_1, \ldots, w^c, \ldots, w_n]$, then the proposal for replacement is

$$g_r(y_{t+1}|y_t) = \mathbb{1}(w^c \in Q) \cdot P_{\text{MLM}}(w^c = w^c|x_{-m})$$

(12)

Here $\mathbb{1}(w^c \in Q)$ is the indicator function which equals 1 if $w^c \in Q$ and 0 otherwise. $P_{\text{MLM}}(w^c = w^c|x_{-m})$ is the probability of the selected token given the rest of the sentence $x_{-m}$. It is computed using a masked language model (MLM), e.g. BERT (Devlin et al. 2019) or RoBERTa (Liu et al. 2019).

The transition function for deletion is rather simple: $g_d(y_{t+1}|y_t) = 1$ if and only if $y_{t+1} = [w_1, \ldots, w_{m-1}, w_{m+1}, \ldots, w_n]$, and 0 for others. The insertion operation consists of two steps. First, a mask token is inserted into the position and then a replacement operation is performed on the inserted token.

### 4 Experiments

#### 4.1 Experimental Setup

**Dataset** We experiment EDUCAT on TIME TRAVEL (Qin et al. 2019), a standard counterfactual story rewriting dataset. TIME TRAVEL is built on ROCStories (Mostafazadeh et al. 2016), which consists of a large set of five-sentence stories $S = s_{1:5}$. The first sentence $s_1$ denotes the premise of a story, $s_2$ sets up the initial context, and the last three sentences $s_{3:5}$ are the story endings. Using causal language we described above, $s_1$, $s_2$, $s_{3:5}$ correspond to $Z = z$, $X = x$, $Y = y$, respectively. In TIME TRAVEL, the initial context was rewritten by humans into a counterfactual context $s'_2$, followed with edited endings $s'_{3:5}$. They correspond to $X = x'$ and $Y = y'$ in the causal graphical model. As EDUCAT is unsupervised and thus does not need training, we run EDUCAT directly on the test set.

The statistics of TIME TRAVEL are reported in Table 1. Only part of the training set is annotated with the edited endings. Each sample in the development and test set is annotated with 3 and 4 rewritten endings respectively, which explains the difference between # of $x'$ and # of $y'$ in the development and test set in Table 1. Note that the fourth edited ending in test set is not included in evaluation as ground truth ending, but only serves as human baseline.

**Baselines** Following previous work, we categorize the baselines into three classes: 1) **Unsupervised zero-shot baselines**, with only off-the-shelf pre-trained models for generation, including pre-trained GPT-2 (generating with $s_1, s'_2$) and DELOREAN (Qin et al. 2020). Moreover, in comparisons with unsupervised editing-based methods, we add CGMH (Miao et al. 2019), which is EDUCAT without conflict detection and coherence score; 2) **Unsupervised training baselines**, GPT-2 + Recon+CF (Qin et al. 2019), which is trained with domain data $S$ and $<s_1, s'_2>$ (i.e. without $s'_{3:5}$); 3) **Supervised training baselines**, with a GPT-2 + SUP (Qin et al. 2019) trained for predicting $s'_{3:5}$ from $S$ and $s'_2$ in the form of $<S, [SEP], s_1, s'_2>$.

Note that in our paper, we aim at using only off-the-shelf pre-trained models for story rewriting, which makes the previous SOTA method DELOREAN our major baseline. DELOREAN iteratively revises the generated tokens by updating their hidden representations during decoding. The update is constrained by minimizing the sentence-level KL divergence between the generated and original endings, followed by a BERT to re-rank the generated candidates with the next sentence prediction task.

**Implementation Details** All of the pre-trained checkpoints are inherited from the implementations of Huggingface (Wolf et al. 2020). Consistent with previous work, we adopt GPT-2, Medium (24 layers) or Small (12 layers), for causal language modeling. We use pre-trained RoBERTa-base as the unsupervised masked language model for token proposal. We keep the first 100 tokens MLM predicts as candidates. We randomly sample one token as the proposed token based on normalized probabilities. In the experiments, we run EDUCAT and its variants for 100 steps.

#### 4.2 Evaluation Metrics

**Automatic Evaluation Metrics** Following previous work, we adopt BLEU-4 (Papineni et al. 2002) and BERTSCORE (Zhang et al. 2020) as automatic metrics, which are referenced metrics. Given ground-truth endings and the generated endings, BLEU computes the number of overlapping n-grams, and BERTSCORE computes their semantic similarity using BERT. As reported in Qin et al. (2019), BLEU measures the minimal-edits property well, but correlates poorly with human judgements w.r.t. coherence.

For assessing the coherence with the counterfactual conditions, we propose a simple, unreferenced, and model-based metric ENTSCORE (ENTS). Inspired by researches

| # counterfactual context ($x'$) | Train | Dev | Test |
|-------------------------------|-------|-----|------|
|                               | 96,867| 1,871| 1,871|
| # edited endings ($y'$)       | 16,752| 5,613| 7,484|

Table 1: Statistics of TIME TRAVEL dataset.
on natural language inference [Kang et al. 2018; Dziri et al. 2019], we fine-tune a RoBERTa (base or large) with binary classification objective to check whether a story context entails a story ending. We use 28,363 stories with annotated edited endings in TIME TRAVEL to train the metric, leading to 113,452 training samples, i.e., x′ contradicts with y but entails by y′ and x contradicts with y′ but entails y. The best metrics achieve the F1 scores of 73.07 (base) and 81.64 (large) in the test set. We take the predicted probability of whether an ending is entailed by the counterfactual context as the output of EAN.

To better evaluate the subtle trade-off in this task, we calculate a harmonic mean of EAN and BLEU to represent the trade-off between coherence and minimal-edits, defined as \( \text{H\text{\textsc{mean}} = \frac{2 \cdot \text{BLEU} \cdot \text{EAN}}{\text{BLEU} + \text{EAN}} } \).

**Human Evaluation Metrics** We also conduct human evaluation to compensate for these automatic metrics and assess their ability for this task. Following [Qin et al. 2020], our human evaluation mainly focuses on two primary criteria: i) coherence, the logical consistency between the counterfactual context \((s_1, s_2)\) and generated endings, and ii) minimal-edits, the extent of minimal-revision between two endings. We calculate the pairwise comparison as human metrics. Annotators are asked to score from 0 to 3 and choose the better one or both between two generated outputs from EDUCAT and baselines without knowledge of their origins. We arrange a training session before annotation session, where the annotators annotate some cases and resolve their disputes through discussion. Then, we randomly select 100 samples from the test set. Each sample was rated by three graduate students, paid with local minimum wage. The final decision is made based on the majority vote.

**Human Correlation with Metrics** Before automatic evaluation, we show the ability of these automatic metrics by performing correlation analysis using the scores produced by human annotators on the generated endings. We calculate three coefficients, including Pearson’s \( r \), Spearman’s \( \rho \) and Kendall’s \( \tau \). Pearson’s \( r \) measures linear correlation, and the latter two measure monotonic correlation, where Spearman’s \( \rho \) is more sensitive to abnormal values. According to Table 2, H\text{\textsc{mean}} proves to be the best metric among them in terms of correlation with human judgements for this task, which is also our primary metric in the experiments.

### Table 2: The correlation between automatic metrics and human judgements in coherence. H\text{\textsc{mean}} is the harmonic mean between E\text{\textsc{nts}} (large) and BLEU. All of these numbers are statistically significant at \( p < 0.01 \).

| Metrics     | Pearson’s \( r \) | Spearman’s \( \rho \) | Kendall’s \( \tau \) |
|-------------|-------------------|------------------------|----------------------|
| BLEU        | 0.2619            | 0.2454                 | 0.1758               |
| BERTSCORE   | 0.3252            | 0.3332                 | 0.2385               |
| ENTS (base) | 0.3937            | 0.3973                 | 0.2865               |
| ENTS (large)| 0.4685            | 0.4732                 | 0.3389               |
| H\text{\textsc{mean}} (large) | 0.4995          | 0.4996                 | 0.3662               |

In terms of correlation with human judgements for this task, which is also our primary metric in the experiments.

#### Table 3: Automatic evaluation results in the test set of TIME TRAVEL. These methods use GPT-2\(_M\) by default. E\text{\textsc{nts}}\(_l\) is short for ENTS\(_{\text{core}}\) (large).

| Method               | BLEU   | BERT   | E\text{\textsc{nts}}\(_l\) | H\text{\textsc{mean}} |
|----------------------|--------|--------|---------------------------|----------------------|
| **Supervised Training** |        |        |                           |                      |
| GPT-2\(_M\) + SUP    | 76.35  | 81.72  | 35.06                     | 48.05                |
| **Unsupervised Training** |        |        |                           |                      |
| GPT-2\(_M\) + FT      | 3.90   | 53.00  | 52.77                     | 7.26                 |
| Recon+CF              | 76.37  | 80.20  | 18.00                     | 29.13                |
| **Off-the-shelf Pre-trained Models** |        |        |                           |                      |
| GPT-2\(_M\)          | 1.39   | 47.13  | 54.21                     | 2.71                 |
| DELOREAN              | 23.89  | 59.88  | 51.40                     | 32.62                |
| CGMH                  | 41.34  | 73.82  | 29.80                     | 34.63                |
| EDUCAT                | 44.05  | 74.06  | 32.28                     | 37.26                |

| Human                | 64.76  | 78.82  | 80.56                     | 71.80                |

#### 4.3 Results

**Automatic Evaluation** Table 3 shows our results w.r.t. automatic metrics. In general, we observe that BLEU and BERTSCORE indicate the trade-off between minimal edits and coherence in this task. Models that generate coherent endings can also cause excessive edits. Among them, EDUCAT achieves the best trade-off in terms of H\text{\textsc{mean}}, which is also the metric that has the best correlation with human judgements, as shown in Table 2.

For supervised and unsupervised training methods, we find Recon+CF scores high on BLEU and BERTSCORE but low on ENTS\(_{\text{core}}\), suggesting that the endings it generates are not coherent with counterfactual contexts but paraphrased from original endings [Qin et al. 2019]. Moreover, the gap remains between supervised methods and unsupervised ones.

Interestingly, zero-shot GPT-2\(_M\) and DELOREAN perform very well in ENTS\(_{\text{core}}\) but poorly on BLEU and BERTSCORE. ENTS\(_{\text{core}}\) draws the decision boundary based on the change of conditions \((s_2, s_2')\). Therefore, as long as the ending follows the counterfactual condition, where large-scale language models such as GPT-2 excel, ENTS\(_{\text{core}}\) will produce a high score. Zero-shot GPT-2\(_M\) does not constrain the generation on minimal-edits to the original endings and hallucinates from the original story during the generation. Hence, it generates fluent endings thanks to the language modeling ability of GPT-2 with over-editing. The same is true for DELOREAN, but it alleviates this problem by constraining on the KL-divergence with original endings. Indeed, it is easy to generate coherent endings with massive edits, as even a zero-shot GPT-2 can achieve a high score in coherence. However, this task puts forward higher demands on the model’s ability to do it under minimal edits to find the causal invariance.

**Human Evaluation** We first show manual evaluation results in Table 4. In general, EDUCAT outperforms CGMH.
shows the potential of ENT in the story domain. Results indicate that using E\textsubscript{NT} ability as the coherence measurement might be weak for grade from national sentence probability as the coherence function. This result suggests that random edit token BLEU and E\textsubscript{NT} smaller GPT-2 this task according to Table 5 in all metrics. Even with proposed modules. We find both components are beneficial to in their minds to connect two events, thus small but critical edits can still result in a logical ending to a human mind.

### Ablation Study
We perform an ablation study for the proposed modules. We find both components are beneficial to this task according to Table 3 in all metrics. Even with smaller GPT-2\textsubscript{s} as the backbone causal language model, EduCat still outperforms unsupervised baselines. In particular, we find a considerable performance drop in BLEU and ENT\textsubscript{SCORE} for EduCat without conflict detection module. This result suggests that random edit token finding is inefficient to find the causal invariance, So the method prefers the editing actions that generate fluent endings instead of ones that balance the trade-off well, which puts forth higher demands to the system.

We observe a mild performance boost in the trade-off (H\textsubscript{MEAN}) by introducing \( x_{\text{Coh}} \) with unsupervised conditional sentence probability as the coherence function \( P_{\text{Coh}} \). What if EduCat has more powerful coherence guidance from \( x_{\text{Coh}} \)? To test the limit of our method, we also upgrade \( x_{\text{Coh}} \) by directly replacing the original \( P_{\text{Coh}} \) with ENT\textsubscript{SCORE} (base), since the unsupervised sentence probability as the coherence measurement might be weak for the story domain. Results indicate that using ENT\textsubscript{SCORE} in \( x_{\text{Coh}} \) leads to a clear boost in coherence (+30.20% in ENT\textsubscript{SCORE}) and the trade-off (+14.95% in H\textsubscript{MEAN}). This shows the potential of EduCat framework for this task given a robust discriminator, which is also similar to the benefits of a strong reward function in reinforcement learning. Nevertheless, to keep this method solely unsupervised with only off-the-shelf models, we claim scores achieved by EduCat with the original \( x_{\text{Coh}} \) as our major results, but with much room for improvement.

### 4.4 Case Study
Finally, we show some of the samples produced by EduCat against baselines in Figure 1 to make an intuitive comparison and explore our method’s limitations. Although Delorean also generates fluent counterfactual stories, it struggles with maintaining the balance between minimal-edits and logical consistency to the counterfactual context, and makes massive edits. In contrast, the discrete editing strategy EduCat works far better than the gradient update-based method in Delorean in terms of minimal edits.

In both cases, EduCat and CGMH conduct a handful of edits to the original endings and yield fluent endings. In the first one, EduCat makes crucial and logical lexical edits, e.g., the sun’s position should be low since it is evening in the altered condition \( z' \), while CGMH and Delorean do not. EduCat shows some commonsense knowledge, as one needs no air conditioning as the weather was starting to cool off, and park is a good place to go in the evening (maybe for a walk). In the second one, Delorean does not generate valid story endings. CGMH makes mistakes by changing “bad sport” to “head coach”, whereas EduCat paraphrases it to “dirty player”.

### 5 Related Work

#### Constrained Text Generation
Many research efforts have been made to control the generation with various desired properties. Most studies (Hu et al. 2018, Tan et al. 2020) train supervised models to inject constraints into generation. In this work, we focus on unsupervised constrained generation, which is much more difficult. Recent unsupervised generation relies heavily on pre-trained language models (PLMs) (Radford et al. 2019, Keskar et al. 2019). Daithathri et al. (2020) control the generation using an external attribute model that affects token decoding through back-propagation. Qin et al. (2020) adopt this idea and adjust for this task by optimizing the sentence generation as a whole through iterative forward and backward passes.
S1: Gina had done everything she could think of to beat the heat. And it was only noon. The sun was still high in the sky. She decided she needed to go where there was air conditioning. She went inside a nearby cafe.

S2: Luckily, it was evening and starting to cool off.

S3: The sun had gotten lower in the sky. She decided she needed to go outside and get some fresh air. She went to the beach.

S4: She decided she needed to go somewhere where there was air conditioning. She decided next time it was so hot she needed to go where there was air conditioning. So she planned to go inside a nearby cafe.

S5: The sun was high in the sky. She decided she needed to go somewhere where there was air. She went to the beach.

S6: She decided that she needed to go somewhere where there was no air conditioning. She decided she needed to go somewhere where there was air conditioning. The sun was low in the sky. She decided that she needed to go somewhere where there was no air conditioning. She headed to the park.

Figure 3: Two samples from the test set of TimeTravel. We present the predictions of EduCat and baselines. Text in red denotes the mistakes these models make.

Another line of unsupervised constrained generation is search-based methods, including methods with constrained beam search (Hokamp and Liu 2017; Lu et al. 2021) and stochastic search. The former line of work is restricted to lexical constraints, while the latter is more extensible. Miao et al. (2019) first introduce Metropolis-Hastings sampling into text generation and constrain the generation with stationary distributions. Zhang et al. (2020a) extend CGMH by designing combinatorial constraints. Liu et al. (2020) model the constraint generation as a discrete optimization problem, which is solved with simulated annealing. To find edit positions, Sha (2020) define differentiable score functions and use gradients to find edit positions and sample actions, while He and Li (2021) train a position finding classifier with X-LNet (Yang et al. 2019) for lexically constrained sentence generation. In this paper, we mainly explore this line of work to non-monotonic reasoning and generation tasks with insights from causal analysis.

Causal Inference and NLP There is a recent surge of interest in how NLP methodology can evaluate and estimate causal effects and how causal inference can enhance current natural language understanding and generation. Researchers have studied how text can be used as a mediator, confounder, treatment, or outcome (Grimmer, Messing, and Westwood 2017; Wood-Doughty, Shipter, and Dredze 2018; Wu et al. 2020; Feder et al. 2021) to estimate causal effect under different contexts such as gender bias, etc. Another line of research attempts to equip the current text generation mechanism with counterfactual reasoning ability. For instance, Kaushik, Hovy, and Lipton (2020); Zeng et al. (2020) augment existing datasets to include counterfactual samples and demonstrate better out of domain generalization ability on tasks as sentimental classification, NER, etc. In terms of work more related to ours (Zhu et al. 2020; Qin et al. 2019; 2020), they explored the counterfactual text generation tasks such as counterfactual dialogue and story generation. Our work adapts idea from both lines of research.

6 Conclusion and Future Work

In this paper, we aim to balance the trade-off between logic and minimal-edits in order to detect causal invariance in the story rewriting task, which demands causal reasoning skills. We propose EduCat, an editing-based unsupervised counterfactual story rewriter using MCMC sampling. For detecting causal invariance, EduCat is equipped with the ability of conflict detection and scores for coherence to control the edit proposals based on causal risk ratio, a measure of causal effects. Experiments on the TimeTravel dataset show that EduCat substantially outperforms unsupervised SOTA methods in both automatic and human evaluation metrics, indicating the superiority of editing-based methods in this task. Further ablation study stresses the importance of the proposed causal reasoning components. Although this work makes an attempt on automatic evaluation of this task by proposing EntScore, we highlight that future research should prioritize on the automatic metrics for this task, especially for unreferenced metrics.

Acknowledgements

We thank Changzhi Sun, Xinbo Zhang, Yuxuan Song, Chao Wang and the anonymous reviewers for the discussions for the manuscript. We also thank Lianhui Qin for providing baseline results. This work was supported by National Key Research and Development Project (No. 2020AAA0109302), Shanghai Science and Technology Innovation Action Plan (No.19511120400) and Shang-
References

Andrieu, C.; De Freitas, N.; Doucet, A.; and Jordan, M. I. 2003. An introduction to MCMC for machine learning. *Machine learning*, 50(1): 5–43.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; et al. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, NeurIPS 2020, December 6-12, 2020, virtual.

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dziri, N.; Kamalloo, E.; Mathewson, K.; and Zaiane, O. 2019. Evaluating Coherence in Dialogue Systems using Entailment. 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)*, 3806–3812. Minneapolis, Minnesota: Association for Computational Linguistics.

Feder, A.; Keith, K. A.; Manzoor, E.; Pryzant, R.; Sridhar, D.; Wood-Doughty, Z.; Eisenstein, J.; Grimmer, J.; Reichart, R.; Roberts, M. E.; Stewart, B. M.; Veitch, V.; and Yang, D. 2021. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. arXiv:2109.00725.

Grimmer, J.; Messing, S.; and Westwood, S. J. 2017. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4): 413–434.

Gu, J.; Huang, F.; Zhao, Z.; Zhu, X.; and Huang, M. 2020. A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation. *Transactions of the Association for Computational Linguistics*, 8: 93–108.

Gu, J.; Wang, Y.; and Huang, M. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6473–6480.

Hao, C.; Pang, L.; Lan, Y.; Wang, Y.; Guo, J.; and Cheng, X. 2021. Sketch and Customize: A Counterfactual Story Generator. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14): 12955–12962.

He, X.; and Li, V. O. 2021. Show Me How To Revise: Improving Lexically Constrained Sentence Generation with XLNet. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14): 12989–12997.

Hernán, M. A. 2004. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, 58(4): 265–271.

Hernán, M. A.; and Robins, J. M. 2020. Causal inference: what if.

Hokamp, C.; and Liu, Q. 2017. Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1535–1546. Vancouver, Canada: Association for Computational Linguistics.

Hu, Z.; Yang, Z.; Salakhutdinov, R.; Qin, L.; Liang, X.; Dong, H.; and Xing, E. P. 2018. Deep Generative Models with Learnable Knowledge Constraints. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, 10522–10533.

Kang, D.; Khot, T.; Sabharwal, A.; and Hovy, E. 2018. AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2418–2428. Melbourne, Australia: Association for Computational Linguistics.

Kausik, D.; Hovy, E.; and Lipton, Z. C. 2020. Learning the Difference that Makes a Difference with Counterfactual Augmented Data. *International Conference on Learning Representations (ICLR)*.

Keskar, N. S.; McCann, B.; Varshney, L. R.; Xiong, C.; and Socher, R. 2019. Ctrl: A conditional transformer language model for controllable generation. arXiv preprint arXiv:1909.05858.

Li, Z.; Ding, X.; and Liu, T. 2018. Generating Reasonable and Diversified Story Ending Using Sequence to Sequence Model with Adversarial Training. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1033–1043. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Liu, X.; Mou, L.; Meng, F.; Zhou, H.; Zhou, J.; and Song, S. 2020. Unsupervised Paraphrasing by Simulated Annealing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 302–312. Online: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Lu, X.; West, P.; Zellers, R.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2021. NeuroLogic Decoding: (Un)supervised
Neural Text Generation with Predicate Logic Constraints. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4288–4299. Online: Association for Computational Linguistics.

Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; and Teller, E. 1953. Equation of state calculations by fast computing machines. The journal of chemical physics, 21(6): 1087–1092.

Miao, N.; Zhou, H.; Mou, L.; Yan, R.; and Li, L. 2019. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 6834–6842.

Mostafazadeh, N.; Chambers, N.; He, X.; Purikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; and Allen, J. F. 2016. A Corpus and Evaluation Framework for Deeper Understanding of Commonsense Stories. CoRR, abs/1604.01696.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Pearl, J. 2009. Causality. Cambridge university press.

Pearl, J.; and Mackenzie, D. 2018. The book of why: the new science of cause and effect. Basic books.

Qin, L.; Bosselut, A.; Holtzman, A.; Bhagavatula, C.; Clark, E.; and Choi, Y. 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), 5043–5053. Hong Kong, China: Association for Computational Linguistics.

Qin, L.; Shwartz, V.; West, P.; Bhagavatula, C.; Hwang, J. D.; Le Bras, R.; Bosselut, A.; and Choi, Y. 2020. Backpropagation-based Decoding for Unsupervised Counterfactual and Abductive Reasoning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 794–805.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. OpenAI Blog, 1(8): 9.

Sha, L. 2020. Gradient-guided Unsupervised Lexically Constrained Text Generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 8692–8703. Online: Association for Computational Linguistics.

Sloman, S.; and Lagnado, D. A. 2004. Causal invariance in reasoning and learning. Psychology of learning and motivation, 44: 287–326.

Tan, B.; Qin, L.; Xing, E.; and Hu, Z. 2020. Summarizing Text on Any Aspects: A Knowledge-Informed Weakly-Supervised Approach. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 6301–6309. Online: Association for Computational Linguistics.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. Online: Association for Computational Linguistics.

Wood-Doughty, Z.; Shipter, I.; and Dredze, M. 2018. Challenges of using text classifiers for causal inference. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, volume 2018, 4586. NIH Public Access.

Wu, Y.; Kuang, K.; Zhang, Y.; Liu, X.; Sun, C.; Xiao, J.; Zhuang, Y.; Si, L.; and Wu, F. 2020. De-biased Court’s View Generation with Causality. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 763–780.

Xu, J.; Ren, X.; Zhang, Y.; Zeng, Q.; Cai, X.; and Sun, X. 2018. A Skeleton-Based Model for Promoting Coherence Among Sentences in Narrative Story Generation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 4306–4315. Brussels, Belgium: Association for Computational Linguistics.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems, 5753–5763.

Zeng, X.; Li, Y.; Zhai, Y.; and Zhang, Y. 2020. Counterfactual Generator: A Weakly-Supervised Method for Named Entity Recognition. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 7270–7280.

Zhang, M.; Jiang, N.; Li, L.; and Xue, Y. 2020a. Language Generation via Combinatorial Constraint Satisfaction: A Tree Search Enhanced Monte-Carlo Approach. In Findings of the Association for Computational Linguistics: EMNLP 2020, 1286–1298. Online: Association for Computational Linguistics.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020b. BERTScore: Evaluating Text Generation with BERT. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Zhu, Q.; Zhang, W.-N.; Liu, T.; and Wang, W. Y. 2020. Counterfactual Off-Policy Training for Neural Dialogue Generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 3438–3448. Online: Association for Computational Linguistics.