Unsupervised Editing For Counterfactual Stories

Jiangjie Chen[1,3], Chun Gan[2], Sijie Cheng[1], Hao Zhou[3], Yanghua Xiao[1] and Lei Li[4]

[1]Fudan University [2]JD.com [3]ByteDance AI Lab [4]University of California, Santa Barbara

Background & Introduction

❖**Task: Counterfactual Story Ending Rewriting**

❖*What if I had done something different? What would be the difference in the following events?*

❖**Goal: Counterfactual Reasoning**

❖A hypothetical thinking process to assess possible outcomes by modifying certain prior conditions

❖**Research Questions:**

❖**The trade-off: Minimal-edits vs. Coherence —** Can we rewrite a coherent new story ending with minimal edits?

❖**Humans do not need training to imagine possible futures! —** Can we achieve it without supervision?

Motivation & Contribution

❖**How can we ensure minimal edits?**

❖We first solve the counterfactual story rewriting task using unsupervised discrete editing method based on MCMC sampling.

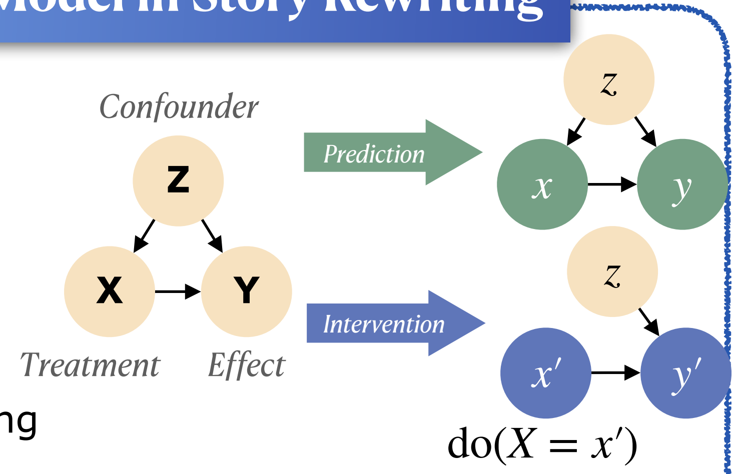
❖**How can we ensure coherence?**

❖We draw inspiration from causal analysis and propose two counterfactual reasoning components that quantify the outcomes of condition changes.

Structured Causal Model in Story Rewriting

❖**SCM in story rewriting**:

1. z: Story premise (one of the observed confounders)
2. x: Initial condition
3. y: Story ending
4. x': Counterfactual condition
5. y': Counterfactual story ending



❖**Intervention**: changing the initial condition to a counterfactual one.

Estimating Potential Outcome After Intervention

❖**A Metric in Causal Analysis — Causal Risk Ratio**

❖The larger CRR is, the more causally related it is to the counterfactual condition than to the initial one.

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

$$P(Y = y \mid do(X = x')) = \sum_z P(Y = y \mid X = x', Z = z) P(Z = z)$$

❖**Causal Sufficiency Assumption**

❖Difficult to enumerate all *unobserved confounders*

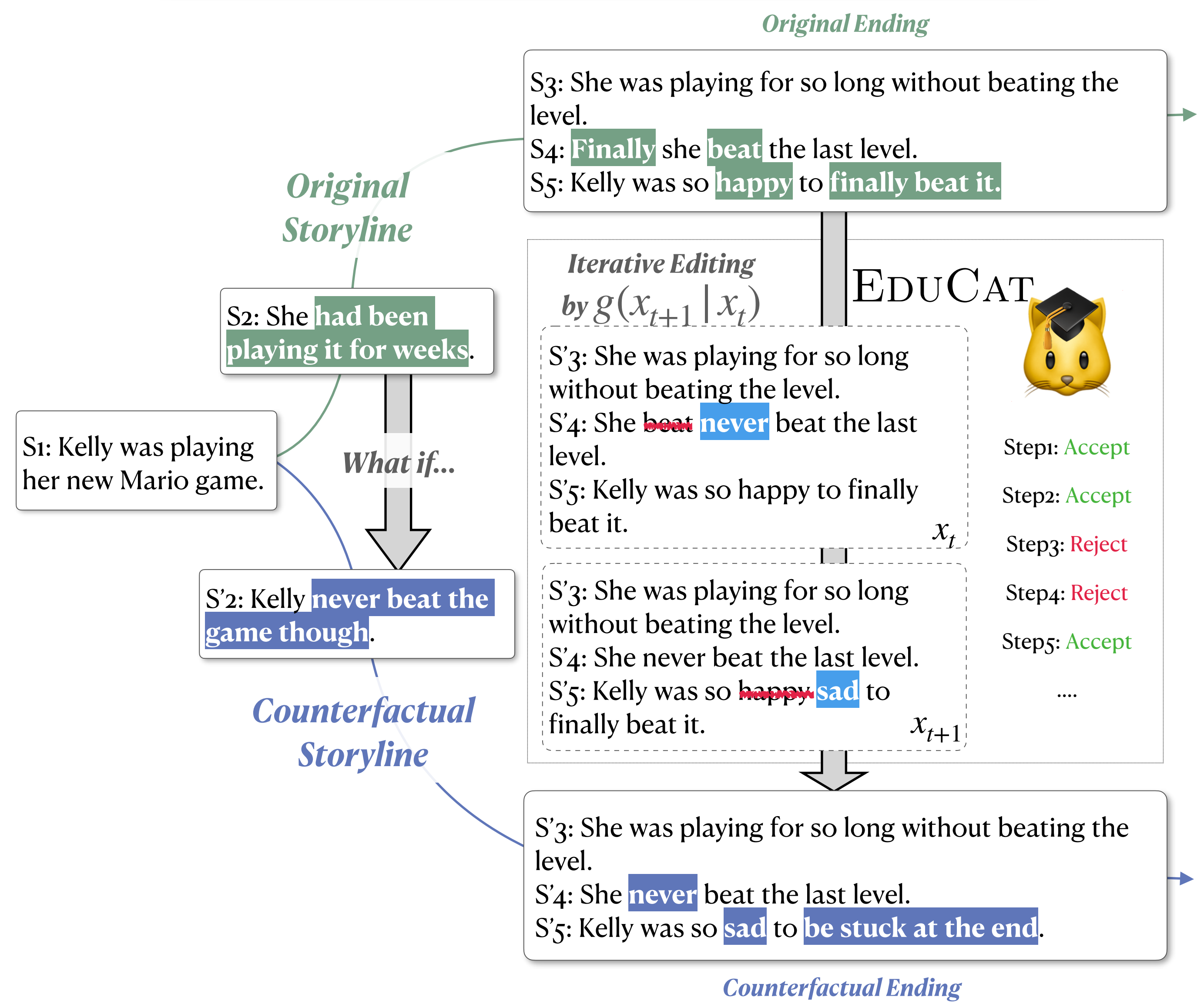
❖Assume only observed confounders (premise z)

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

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

Now we can calculate CRR!

EDUCAT: Editing a New Story Ending



❖**EDUCAT: Unsupervised Constrained Editing via MCMC Sampling**

1. Define desired properties as stationary distribution $\pi(\mathrm{y})$
2. Move y_t to y_{t+1} by generating from the proposal distribution $g(\mathrm{y}_{t+1} \mid \mathrm{y}_t)$
3. Accept a proposal with acceptance rate $\boxed{\alpha(\mathrm{y}_{t+1} \mid \mathrm{y}_t)}$
4. Iterate until convergence
5. Rank the accepted ones with $\pi(\cdot)$

$$\min\left\{1, \frac{\pi(\mathrm{y}_{t+1})^{1/T} g(\mathrm{y}_t \mid \mathrm{y}_{t+1})}{\pi(\mathrm{y}_t)^{1/T} g(\mathrm{y}_{t+1} \mid \mathrm{y}_t)}\right\}$$

❖**Desired properties for stationary distribution** $\pi(y)$

1. **Fluency**: sentence probability from GPT-2;
$$\mathcal{X}_{\mathrm{LM}}(y^*) = \prod_{i=1}^{N} P_{\mathrm{LM}}(y_i^* \mid z, x', y_{<i}^*)$$
2. **Coherence**: Punish proposed endings contradictory to the counterfactual conditions but consistent with the initial ones.
$$\mathcal{X}_{\mathrm{Coh}}(y^*) = \frac{P_{\mathrm{Coh}}(Y = y^* \mid z, x')}{P_{\mathrm{Coh}}(Y = y^* \mid z, x)}$$

CRR

❖**Make an Edit Proposal**

1. *Where to edit? — conflict token detection in* y_t
$$P_{\mathrm{cf}}(y_i^*) = \mathrm{softmax}\left(\frac{P_{\mathrm{LM}}(y_i^* \mid z, x, y_{<i}^*)}{P_{\mathrm{LM}}(y_i^* \mid z, x', y_{<i}^*)}\right)$$

CRR

2. *Edit with what? — modification actions*
   - *Replace*: mask-predict with an MLM (e.g., BERT)
     - $g_r(\mathrm{y}_{t+1} \mid \mathrm{y}_t) = \mathbb{1}(w^c \in \mathcal{Q}) \cdot P_{\mathrm{MLM}}(w_m^* = w^c \mid \mathrm{x}_{-m})$
     - Sample from $P_{\mathrm{MLM}}(\cdot)$
   - *Insert*: insert a [MASK], then do *Replace*
   - *Delete*: reverse of *Insert*

More details in the paper!

Experiments

❖**Dataset**

❖TimeTravel [Qin et al. 2019]

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

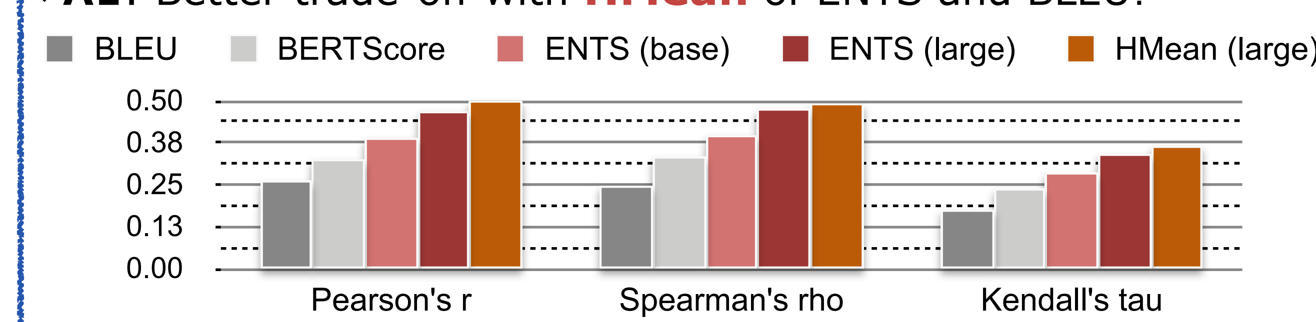
Unsupervised, only test set used!

❖**Metrics**

- BLEU
- BERTScore
- **EntScore**: a model-based metric for coherence
  - Leaning towards initial or counterfactual? Binary classification with RoBERTa
- **HMean**: Harmonic Mean of EntScore and BLEU
  - For the *trade-off*

❖**RQ1: How are these metrics correlates with humans**

❖**A1**: Better trade-off with **HMean** of ENTS and BLEU!



❖**RQ2: Performance of EDUCAT?**

❖**A1**: **Competitive** against baselines under automatic and human evaluation.

Method	BLEU	BERT	ENTS$_l$	HMEAN
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

Methods	Coherence		
	Win	Tie	Lose
EDUCAT vs. DELOREAN	45%	32%	23%
EDUCAT vs. CGMH	32%	51%	17%
EDUCAT vs. Human	12%	24%	64%
	Min-edits		
EDUCAT vs. DELOREAN	64%	27%	9%
EDUCAT vs. CGMH	26%	49%	25%
EDUCAT vs. Human	16%	40%	44%

❖**RQ3: The source of performance gain?**

❖**A1**: Both conflict detection and coherence objective.

Ablation	BLEU	BERT	ENTS$_l$	HMEAN
EDUCAT (GPT-2$_S$)	39.82	72.35	31.72	35.31
EDUCAT (GPT-2$_M$)	44.05	74.06	32.28	37.26
– $\mathcal{X}_{\mathrm{Coh}}$	**44.20**	**74.27**	31.44	36.74
– *conflict detection*	40.96	73.61	30.79	35.16
– *both*	41.34	73.82	29.80	34.63
+ $\mathcal{X}_{\mathrm{Coh}}$ w/ ENTS$_b$	43.65	74.09	**42.03**	**42.83**

More details in the paper!

Case Study

S1: Gina had done everything she could think of to beat the heat.
S2: And it was only noon.
S3: The sun was still high in the sky.
S4: She decided she needed to go where there was air conditioning.
S5: She went inside a nearby cafe.

S'2: Luckily, it was evening and starting to cool off.

S'3: The sun was still high in the sky.
S'4: She decided she needed to go outside and get some fresh air.
S'5: She went inside and got some fresh air.
DELOREAN

S'3: The sun was low in the sky.
S'4: She decided that she needed to go somewhere where there was no air conditioning.
S'5: She headed to the park.
EDUCAT

If you have any questions, please email: jjchen19@fudan.edu.cn