Unsupervised Editing for Counterfactual Stories

Jiangjie Chen\textsuperscript{1,3}, Chun Gan\textsuperscript{2}, Sijie Cheng\textsuperscript{1}, Hao Zhou\textsuperscript{3}, Yanguhua Xiao\textsuperscript{1}, Lei Li\textsuperscript{4}
Automatic Story Writing
“I want some steak!”
“It’s a sunny day, let’s go out😊!”
“Nice steak they have😊!”

Photo taken @Shanghai
Automatic Story Re-Writing

“What if 🌧️?”

Photo taken @Shanghai
“Oh☹, I hate rainy days.”

What if…?
“What should I do?

What if...?
“I might as well cook it myself 🍳!”

What if...?
Counterfactual Story Rewriting for Creative NLG

What if...?
Counterfactual Reasoning

- A hypothetical thinking process to assess possible outcomes by modifying certain prior conditions.
Counterfactual Reasoning

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

Judea Pearl’s “Ladder of Causality”

1. ASSOCIATION
ACTIVITY: Seeing, Observing
QUESTIONS: What if I see X? How are the variables related? How would seeing X change my belief in Y?
EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?

2. INTERVENTION
ACTIVITY: Doing, Intervening
QUESTIONS: What if I do X? How? What would Y be if I do X? How can I make Y happen?
EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

3. COUNTERFACTUALS
ACTIVITY: Imagining, Retrospection, Understanding
QUESTIONS: What if I had done ...? Why? Was it X that caused Y? What if X had not occurred? What if I had acted differently?
EXAMPLES: Was the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?
Counterfactual Reasoning

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

Judea Pearl’s “Ladder of Causality”

**Association: What if I see...?**
Counterfactual Reasoning

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

Judea Pearl’s “Ladder of Causality”

1. ASSOCIATION
   ACTIVITY: Seeing, Observing
   QUESTIONS: What if I see...? 
   (How are the variables related? 
   How would seeing X change my belief in Y?)
   EXAMPLES: What does a symptom tell me about a disease? 
   What does a survey tell us about the election results?

2. INTERVENTION
   ACTIVITY: Doing, Intervening
   QUESTIONS: What if I do...? I have? 
   (What would Y be if I do X? 
   How can I make Y happen?)
   EXAMPLES: If I take aspirin, will my headache be cured? 
   What if we ban cigarettes?

3. COUNTERFACTUALS
   ACTIVITY: Imagining, Respecifying, Understanding
   QUESTIONS: What if I had done...? Why? 
   (Was it X that caused Y? What if X had not occurred? What if I had acted differently?)
   EXAMPLES: Was it the aspirin that stopped my headache? 
   Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

Intervention: What if I do...?
Counterfactual Reasoning

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

Judea Pearl’s “Ladder of Causality”

3. COUNTERFACTUALS
ACTIVITY: Imagining, Reflection, Understanding
QUESTIONS: What if I had done ...? Why?
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)
EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

Counterfactuals: What if I had done...?
Counterfactual Reasoning

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

Judea Pearl’s “Ladder of Causality”

Counterfactuals: What if I had done...?

• Challenge: Causal Invariance

• the factors that hold constant with the change of conditions in a series of events
Counterfactual Reasoning

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

Judea Pearl’s “Ladder of Causality”

• **Challenge:** Causal Invariance
  • the factors that hold constant with the change of conditions in a series of events
The *Trade-off*: Minimal-edits vs. Coherence

Can we rewrite a new story ending with *minimal edits*?

**Original Storyline**

S1: Kelly was playing her new Mario game.

S2: She had been playing it for weeks.

S2': Kelly never beat the game though.

**Original Ending**

S3: She was playing for so long without beating the level.

S4: Finally she beat the last level.

S5: Kelly was so happy to finally beat it.

**Counterfactual Storyline**

S3': She was playing for so long without beating the level.

S4': She never beat the last level.

S5': Kelly was so sad to be stuck at the end.

**Counterfactual Ending**
The Trade-off: Minimal-edits vs. Coherence

Can we rewrite a new story ending with minimal edits?

Original Storyline

S1: Kelly was playing her new Mario game.

S2: She had been playing it for weeks.

Original Ending

S3: She was playing for so long without beating the level.
S4: Finally she beat the last level.
S5: Kelly was so happy to finally beat it.

Counterfactual Storyline

S’2: Kelly never beat the game though.

Counterfactual Ending

S’3: She was playing for so long without beating the level.
S’4: She never beat the last level.
S’5: Kelly was so sad to be stuck at the end.

For pre-trained LMs, massive editing can almost certainly lead to a coherent ending.
The Trade-off: Minimal-edits vs. Coherence

Can we rewrite a new story ending with minimal edits?

Also do it without supervision!

Humans do not need training to imagine possible futures!

For pre-trained LMs, massive editing can almost certainly lead to a coherent ending.
How does Previous Method Solve this Problem?

Qin, Lianhui, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. EMNLP 2020
How does Previous Method Solve this Problem?

Qin, Lianhui, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. EMNLP 2020
How does Previous Method Solve this Problem?

**Input:**
Ray hung a tire on a rope to make his daughter a swing.

**Past context X**

**Future constraint Z**

**Output:**
She hit the rope and the tire fell on top of her.

\[ Y : y_1, y_2, ..., y_N = \text{Sampling}(\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_N) \]

**Constraints backprop to the pre-trained LMs**

**Still massive edits?**
**EDUCAT: Edit a Story Ending**

**Original Ending**

S3: She was playing for so long without beating the level.
S4: Finally she beat the last level.
S5: Kelly was so happy to finally beat it.

**Counterfactual Ending**

S’3: She was playing for so long without beating the level.
S’4: She never beat the last level.
S’5: Kelly was so sad to finally beat it.

**Iterative Editing**

\[ g(x_{t+1} \mid x_t) \]

**Step 1:** Accept
**Step 2:** Accept
**Step 3:** Reject
**Step 4:** Reject
**Step 5:** Accept

**What if...**

S1: Kelly was playing her new Mario game.
S2: She had been playing it for weeks.
S’2: Kelly never beat the game though.

**Original Storyline**

S1: Kelly was playing her new Mario game.
S2: She had been playing it for weeks.
S3: She was playing for so long without beating the level.
S4: Finally she beat the last level.
S5: Kelly was so happy to finally beat it.

**Counterfactual Storyline**

S1: Kelly was playing her new Mario game.
S2: She never beat the game though.
S3: She was playing for so long without beating the level.
S4: She never beat the last level.
S5: Kelly was so sad to finally beat it.
Structured Causal Model

**Confounder**

- \( Z \)
- \( X \) -> \( Y \)

**Treatment**

- \( X \)

**Effect**

- \( Y \)

**Prediction**

- \( x \) -> \( y \)

**Premise**

- \( z \)

**Condition**

- \( z \) -> \( x \) -> \( y \)

**Ending**

- Premise
- Condition
- Ending

- \( 🥩 \)
- 🚶
- 🏨
Structured Causal Model

Confounding

\[ \text{Confounder} \]

\[
\begin{align*}
X & \rightarrow Z & \rightarrow Y \\
\text{Treatment} & & \text{Effect}
\end{align*}
\]

Prediction

\[ \text{Prediction} \]

\[
\begin{align*}
x & \rightarrow y \\
\end{align*}
\]

Intervention

\[ \text{Intervention} \]

\[
\text{do}(X = x')
\]

\[
\begin{align*}
x' & \rightarrow y' \\
\end{align*}
\]
Estimating Potential Outcome After Intervention — Causal Risk Ratio

Causal Risk Ratio:

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

$$P(Y = y \mid \text{do}(X = x')) = \sum_z P(Y = y \mid X = x', Z = z)P(Z = z)$$
Estimating Potential Outcome After Intervention — Causal Risk Ratio

Causal Risk Ratio:

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

\[
P(Y = y \mid \text{do}(X = x')) = \sum_z P(Y = y \mid X = x', Z = z)P(Z = z)
\]

Causal Sufficiency Assumption

\[
P(Y = y \mid \text{do}(X = x)) = P(Y = y \mid X = x, Z = z)
\]
Estimating Potential Outcome After Intervention — Causal Risk Ratio

Causal Risk Ratio:

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

\[
P(Y = y \mid \text{do}(X = x')) = \sum_z P(Y = y \mid X = x', Z = z)P(Z = z)
\]

Causal Sufficiency Assumption

\[
P(Y = y \mid \text{do}(X = x)) = P(Y = y \mid X = x, Z = z)
\]

\[
CRR = \frac{P(Y = y \mid X = x', Z = z)}{P(Y = y \mid X = x, Z = z)}
\]
Unsupervised Constrained Editing via MCMC Sampling

- CGMH: sentence generation with **Metropolis-Hastings Sampling**. [Miao et al. 2019]
• CGMH: sentence generation with Metropolis-Hastings Sampling. [Miao et al. 2019]
  – Define desired properties as stationary distribution $\pi(y)$
Unsupervised Constrained Editing via MCMC Sampling

- CGMH: sentence generation with Metropolis-Hastings Sampling. [Miao et al. 2019]
  - Define desired properties as stationary distribution $\pi(y)$
  - Move $y_t$ to $y_{t+1}$ by generating from the proposal distribution $g(y_{t+1} \mid y_t)$
Unsupervised Constrained Editing via MCMC Sampling

• CGMH: sentence generation with **Metropolis-Hastings Sampling**. [Miao et al. 2019]
  - Define desired properties as stationary distribution $\pi(y)$
  - Move $y_t$ to $y_{t+1}$ by generating from the proposal distribution $g(y_{t+1} | y_t)$
  - Accept a proposal with acceptance rate $\alpha(y_{t+1} | y_t)$
Unsupervised Constrained Editing via MCMC Sampling

- CGMH: sentence generation with **Metropolis-Hastings Sampling**. [Miao et al. 2019]
  - Define desired properties as stationary distribution $\pi(y)$
  - Move $y_t$ to $y_{t+1}$ by generating from the proposal distribution $g(y_{t+1} | y_t)$
  - Accept a proposal with acceptance rate $\alpha(y_{t+1} | y_t)$
  - Iterate until convergence
Unsupervised Constrained Editing via MCMC Sampling

- CGMH: sentence generation with **Metropolis-Hastings Sampling**. [Miao et al. 2019]
  - Define desired properties as stationary distribution \( \pi(y) \)
  - Move \( y_t \) to \( y_{t+1} \) by generating from the proposal distribution \( g(y_{t+1} | y_t) \)
  - Accept a proposal with acceptance rate \( \alpha(y_{t+1} | y_t) \)
  - Iterate until convergence
  - Rank the accepted ones with \( \pi(\cdot) \)
Unsupervised Constrained Editing via MCMC Sampling

- CGMH: sentence generation with Metropolis-Hastings Sampling. [Miao et al. 2019]
  - Define desired properties as stationary distribution $\pi(y)$
  - Move $y_t$ to $y_{t+1}$ by generating from the proposal distribution $g(y_{t+1} | y_t)$
  - Accept a proposal with acceptance rate $\alpha(y_{t+1} | y_t)$
  - Iterate until convergence
  - Rank the accepted ones with $\pi(\cdot)$

$$\alpha(y_{t+1} | y_t) = \min \left\{ 1, \frac{\pi(y_{t+1})^{1/T} g(y_t | y_{t+1})}{\pi(y_t)^{1/T} g(y_{t+1} | y_t)} \right\}$$
Unsupervised Constrained Editing via MCMC Sampling

- CGMH: sentence generation with **Metropolis-Hastings Sampling**. [Miao et al. 2019]
  - Define desired properties as stationary distribution $\pi(y)$
  - Move $y_t$ to $y_{t+1}$ by generating from the proposal distribution $g(y_{t+1} | y_t)$
  - Accept a proposal with acceptance rate $\alpha(y_{t+1} | y_t)$
  - Iterate until convergence
  - Rank the accepted ones with $\pi(\cdot)$

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

$$\pi(y) \propto \mathcal{X}_{LM}(y) \cdot \mathcal{X}_{Coh}(y)$$

coherence & fluency
Desired Properties: Fluency and Coherence

- **Fluency Score**
  - Sentence probability from a PLM (e.g., GPT-2)

\[ \mathcal{X}_{LM}(y^*) = \prod_{i=1}^{N} P_{LM}(y_i^* | z, x', y_{<i}) \]
Desired Properties: Fluency and Coherence

**Fluency Score**
- Sentence probability from a PLM (e.g., GPT-2)

\[
\mathcal{X}_{\text{LM}}(y^*) = \prod_{i=1}^{N} P_{\text{LM}}(y^*_i | z, x', y^*_{<i})
\]

**Coherence Score**
- **Punish** proposed endings contradictory to the counterfactual conditions but consistent with the initial ones
- Inspired by CRR
- \(P_{\text{Coh}}\) could be changed from a PLM to more sophisticated ones

\[
\mathcal{X}_{\text{Coh}}(y^*) = \frac{P_{\text{Coh}}(Y = y^* | z, x')}{P_{\text{Coh}}(Y = y^* | z, x)}
\]

CRR = \[
\frac{P(Y = y | X = x', Z = z)}{P(Y = y | X = x, Z = z)}
\]
Make an Edit Proposal — Where to Edit?

- Conflict token detection
Conflict token detection

\[ P_{cf}(y^*_i) = \text{softmax} \left( \frac{P_{LM}(y^*_i | z, x, y^*_i)}{P_{LM}(y^*_i | z, x', y^*_i)} \right) \]

\[ \text{CRR} = \frac{P(Y = y | X = x', Z = z)}{P(Y = y | X = x, Z = z)} \]

S5: Kelly was so happy to finally beat it.
Make an Edit Proposal — Edit with What?

- Modification actions

\[ g(y_{t+1} \mid y_t) = \frac{1}{3} \sum_{\text{op} \in \{r,d,i\}} g_{\text{op}}(y_{t+1} \mid y_t) \]

- Replace: mask-predict with an MLM (e.g., BERT)
  - \[ g_r(y_{t+1} \mid y_t) = 1(w^c \in Q) \cdot P_{\text{MLM}}(w^*_m = w^c \mid x_{-m}) \]
  - Sample from \( P_{\text{MLM}}(\cdot) \)

- Insert: insert a [MASK], then do Replace

- Delete: reverse of Insert
EDUCAT: Edit a Story Ending

Original Ending

S1: Kelly was playing her new Mario game.

S2: She had been playing it for weeks.

S3: She was playing for so long without beating the level.

S4: Finally she beat the last level.

S5: Kelly was so happy to finally beat it.

Original Storyline

What if...

Counterfactual Storyline

S’2: Kelly never beat the game though.
EDUCAT: Edit a Story Ending

Original Storyline

S1: Kelly was playing her new Mario game.

S2: She had been playing it for weeks.

S3: She was playing for so long without beating the level.
S4: Finally she beat the last level.
S5: Kelly was so happy to finally beat it.

Counterfactual Storyline

S2: Kelly never beat the game though.

What if...

Iterative Editing by $g(x_{t+1} | x_t)$

S’3: She was playing for so long without beating the level.
S’4: She beat never beat the last level.
S’5: Kelly was so happy to finally beat it.
EDUCAT: Edit a Story Ending

Original Storyline

S1: Kelly was playing her new Mario game.
S2: She had been playing it for weeks.
S3: She was playing for so long without beating the level.
S4: Finally she beat the last level.
S5: Kelly was so happy to finally beat it.

Counterfactual Storyline

S’2: Kelly never beat the game though.
S’3: She was playing for so long without beating the level.
S’4: She beat never beat the last level.
S’5: Kelly was so happy to finally beat it.

Iterative Editing

by \( g(x_{t+1} | x_t) \)

Step 1: Accept
Step 2: Accept
Step 3: Reject
Step 4: Reject
Step 5: Accept
....
EDUCAT: Edit a Story Ending

Original Storyline

S1: Kelly was playing her new Mario game.
S2: She had been playing it for weeks.
S3: She was playing for so long without beating the level.
S4: Finally she beat the last level.
S5: Kelly was so happy to finally beat it.

Counterfactual Storyline

S’2: Kelly never beat the game though.
S’3: She was playing for so long without beating the level.
S’4: She never beat the last level.
S’5: Kelly was so happy to finally beat it.

Iterative Editing

by \( g(x_{t+1} | x_t) \)

S’3: She was playing for so long without beating the level.
S’4: She never beat the last level.
S’5: Kelly was so happy to finally beat it.

Step1: Accept
Step2: Accept
Step3: Reject
Step4: Reject
Step5: Accept

…. 

EDUCAT: Edit a Story Ending

Original Ending

S3: She was playing for so long without beating the level.
S4: Finally she **beat** the last level.
S5: Kelly was so happy to **finally** beat it.

Counterfactual Ending

S’3: She was playing for so long without beating the level.
S’4: She **never** beat the last level.
S’5: Kelly was so happy to **finally** beat it.

Iterative Editing

by $g(x_{t+1} | x_t)$

Original Storyline

S1: Kelly was playing her new Mario game.
S2: She **had been** playing it for weeks.

Counterfactual Storyline

S’2: Kelly **never** beat the game though.

What if...

Step 1: Accept
Step 2: Accept
Step 3: Reject
Step 4: Reject
Step 5: Accept

Counterfactual Ending

S’3: She was playing for so long without beating the level.
S’4: She never **beat** the last level.
S’5: Kelly was so happy to **finally** beat it.

Counterfactual Ending

S’3: She was playing for so long without beating the level.
S’4: She never **beat** the last level.
S’5: Kelly was so happy to **finally** beat it.

Step 1: Accept
Step 2: Accept
Step 3: Reject
Step 4: Reject
Step 5: Accept

…. 
Experiments: Dataset and Metrics

• Dataset
  – TimeTravel

• Metrics
  – BLEU
  – BERTScore

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td># counterfactual context ($x'$)</td>
<td>96,867</td>
<td>1,871</td>
<td>1,871</td>
</tr>
<tr>
<td># edited endings ($y'$)</td>
<td>16,752</td>
<td>5,613</td>
<td>7,484</td>
</tr>
</tbody>
</table>

Table 1: Statistics of TIMETRAVEL dataset.
Experiments: Dataset and Metrics

- **Dataset**
  - TimeTravel

- **Metrics**
  - BLEU
  - BERTScore
    - **EntScore**: a model-based discriminative metric
      - Initial or counterfactual? Binary classification with RoBERTa
      - For coherence
Experiments: Dataset and Metrics

**• Dataset**
- TimeTravel

**• Metrics**
- BLEU
- BERTScore
- **EntScore: a model-based discriminative metric**
  - Initial or counterfactual? Binary classification with RoBERTa
  - For coherence
- **HMean: Harmonic Mean of EntScore and BLEU**
  - For the trade-off

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td># counterfactual context ($x'$)</td>
<td>96,867</td>
<td>1,871</td>
<td>1,871</td>
</tr>
<tr>
<td># edited endings ($y'$)</td>
<td>16,752</td>
<td>5,613</td>
<td>7,484</td>
</tr>
</tbody>
</table>

Table 1: Statistics of TIMETRavel dataset.
## Quality of Metrics: Correlation with Humans

<table>
<thead>
<tr>
<th>Metric</th>
<th>Pearson's r</th>
<th>Spearman's rho</th>
<th>Kendall's tau</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>0.38</td>
<td>0.38</td>
<td>0.13</td>
</tr>
<tr>
<td>ENTS (large)</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>BERTScore</td>
<td>0.38</td>
<td>0.38</td>
<td>0.13</td>
</tr>
<tr>
<td>ENTS (base)</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>HMean (large)</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
</tbody>
</table>

### Chart:
- **X-axis:** Metrics (Pearson's r, Spearman's rho, Kendall's tau)
- **Y-axis:** Correlation Values (0.00 to 0.50)
- **Legend:**
  - BLEU
  - ENTS (large)
  - BERTScore
  - ENTS (base)
  - HMean (large)
Quality of Metrics: Correlation with Humans

- **Pearson's r**
- **Spearman's rho**
- **Kendall's tau**

**Quality of Metrics:**
- Correlation with Humans

Diagram showing correlations with different metrics:
- **BLEU**
- **BERTScore**
- **ENTS (large)**
- **ENTS (base)**
- **HMean (large)**
Quality of Metrics: Correlation with Humans

- Pearson's $r$
- Spearman's $\rho$
- Kendall's $\tau$
- BLEU
- BERTScore
- ENTS (base)
- ENTS (large)
- HMean (large)

Bar chart showing correlation with humans for different metrics.
Quality of Metrics: Correlation with Humans

- Pearson's $r$
- Spearman's $\rho$
- Kendall's $\tau$
- BLEU
- BERTScore
- ENTS (base)
- ENTS (large)
- HMean (large)

Graph showing the correlation of different metrics with humans.
Quality of Metrics: Correlation with Humans

Better trade-off with HMean of ENTS and BLEU!
Automatic and Human Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>BERT</th>
<th>EntS&lt;sub&gt;L&lt;/sub&gt;</th>
<th>HMean</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supervised Training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPT-2&lt;sub&gt;M&lt;/sub&gt; + SUP</td>
<td>76.35</td>
<td>81.72</td>
<td>35.06</td>
<td>48.05</td>
</tr>
<tr>
<td><strong>Unsupervised Training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPT-2&lt;sub&gt;M&lt;/sub&gt; + FT</td>
<td>3.90</td>
<td>53.00</td>
<td>52.77</td>
<td>7.26</td>
</tr>
<tr>
<td>Recon+CF</td>
<td>76.37</td>
<td>80.20</td>
<td>18.00</td>
<td>29.13</td>
</tr>
<tr>
<td><strong>Off-the-shelf Pre-trained Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPT-2&lt;sub&gt;M&lt;/sub&gt;</td>
<td>1.39</td>
<td>47.13</td>
<td><strong>54.21</strong></td>
<td>2.71</td>
</tr>
<tr>
<td>DELOREAN</td>
<td>23.89</td>
<td>59.88</td>
<td>51.40</td>
<td>32.62</td>
</tr>
<tr>
<td>CGMH</td>
<td>41.34</td>
<td>73.82</td>
<td>29.80</td>
<td>34.63</td>
</tr>
<tr>
<td>EDUCAT</td>
<td><strong>44.05</strong></td>
<td><strong>74.06</strong></td>
<td>32.28</td>
<td><strong>37.26</strong></td>
</tr>
<tr>
<td>Human</td>
<td>64.76</td>
<td>78.82</td>
<td>80.56</td>
<td>71.80</td>
</tr>
</tbody>
</table>

Table 3: Automatic evaluation results in the test set of TIME-TRAVEL. These methods use GPT-2<sub>M</sub> by default. EntS<sub>L</sub> is short for EntSCORE (large).
Automatic and Human Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>BERT</th>
<th>EntS_t</th>
<th>HMEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supervised Training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPT-2_M + SUP</td>
<td>76.35</td>
<td>81.72</td>
<td>35.06</td>
<td>48.05</td>
</tr>
<tr>
<td><strong>Unsupervised Training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPT-2_M + FT</td>
<td>3.90</td>
<td>53.00</td>
<td>52.77</td>
<td>7.26</td>
</tr>
<tr>
<td>Recon+CF</td>
<td>76.37</td>
<td>80.20</td>
<td>18.00</td>
<td>29.13</td>
</tr>
<tr>
<td><strong>Off-the-shelf Pre-trained Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPT-2_M</td>
<td>1.39</td>
<td>47.13</td>
<td>54.21</td>
<td>2.71</td>
</tr>
<tr>
<td>DELOREAN</td>
<td>23.89</td>
<td>59.88</td>
<td>51.40</td>
<td>32.62</td>
</tr>
<tr>
<td>CGMH</td>
<td>41.34</td>
<td>73.82</td>
<td>29.80</td>
<td>34.63</td>
</tr>
<tr>
<td>EDUCAT</td>
<td>44.05</td>
<td>74.06</td>
<td>32.28</td>
<td>37.26</td>
</tr>
<tr>
<td>Human</td>
<td>64.76</td>
<td>78.82</td>
<td>80.56</td>
<td>71.80</td>
</tr>
</tbody>
</table>

Table 3: Automatic evaluation results in the test set of TIME-TRAVEL. These methods use GPT-2_M by default. EntS_t is short for ENTSCORE (large).

- EDUCAT is competitive against baselines but falls far behind humans.
Automatic and Human Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>BERT</th>
<th>EntSi</th>
<th>HMean</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supervised Training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPT-2M + SUP</td>
<td>76.35</td>
<td>81.72</td>
<td>35.06</td>
<td>48.05</td>
</tr>
<tr>
<td><strong>Unsupervised Training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPT-2M + FT</td>
<td>3.90</td>
<td>53.00</td>
<td>52.77</td>
<td>7.26</td>
</tr>
<tr>
<td>Recon+CF</td>
<td>76.37</td>
<td>80.20</td>
<td>18.00</td>
<td>29.13</td>
</tr>
<tr>
<td><strong>Off-the-shelf Pre-trained Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPT-2M</td>
<td>1.39</td>
<td>47.13</td>
<td><strong>54.21</strong></td>
<td>2.71</td>
</tr>
<tr>
<td>DELOREAN</td>
<td>23.89</td>
<td>59.88</td>
<td>51.40</td>
<td>32.62</td>
</tr>
<tr>
<td>CGMH</td>
<td>41.54</td>
<td>73.82</td>
<td>29.80</td>
<td>34.63</td>
</tr>
<tr>
<td>EDUCAT</td>
<td><strong>44.05</strong></td>
<td><strong>74.06</strong></td>
<td>32.28</td>
<td><strong>37.26</strong></td>
</tr>
<tr>
<td>Human</td>
<td>64.76</td>
<td>78.82</td>
<td>80.56</td>
<td>71.80</td>
</tr>
</tbody>
</table>

Table 3: Automatic evaluation results in the test set of TIME-TRAVEL. These methods use GPT-2M by default. EntSi is short for EntSCORE (large).

- EDUCAT is competitive against baselines but falls far behind humans.
- With massive edits, even a pre-trained GPT-2 can write coherent endings.

(Please check the paper for details.)
- EDUCAT is competitive against baselines but falls far behind humans.
- With massive edits, even a pre-trained GPT-2 can write coherent endings.
- EDUCAT is competitive in coherence and minimal-edits under human evaluation.

(Please check the paper for details.)

Table 3: Automatic evaluation results in the test set of TIME-TRAVEL. These methods use GPT-2$_M$ by default. ENTS$_l$ is short for ENTSCORE (large).

Table 4: Manual evaluation results, with scores denoting the percentage of Win, Lose or Tie when comparing EDUCAT with baselines.
Both conflict detection and coherence objective work for the task.

Ablation Study

<table>
<thead>
<tr>
<th>Ablation</th>
<th>BLEU</th>
<th>BERT</th>
<th>ENT$S_I$</th>
<th>HMEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDUCAT (GPT-2$_S$)</td>
<td>39.82</td>
<td>72.35</td>
<td>31.72</td>
<td>35.31</td>
</tr>
<tr>
<td>EDUCAT (GPT-2$_M$)</td>
<td>44.05</td>
<td>74.06</td>
<td>32.28</td>
<td>37.26</td>
</tr>
<tr>
<td>− $X_{Coh}$</td>
<td><strong>44.20</strong></td>
<td><strong>74.27</strong></td>
<td>31.44</td>
<td>36.74</td>
</tr>
<tr>
<td>− conflict detection</td>
<td>40.96</td>
<td>73.61</td>
<td>30.79</td>
<td>35.16</td>
</tr>
<tr>
<td>− both</td>
<td>41.34</td>
<td>73.82</td>
<td>29.80</td>
<td>34.63</td>
</tr>
<tr>
<td>+ $X_{Coh}$ w/ ENT$S_b$</td>
<td>43.65</td>
<td>74.09</td>
<td><strong>42.03</strong></td>
<td><strong>42.83</strong></td>
</tr>
</tbody>
</table>

Table 5: Ablation study of EDUCAT in terms of conflict detection module and coherence score $X_{Coh}$. We also change the $P_{Coh}$ in $X_{Coh}$ to the trained discriminative metric ENTSCORE.

(Please check the paper for details.)
Ablation Study

<table>
<thead>
<tr>
<th>Ablation</th>
<th>BLEU</th>
<th>BERT</th>
<th>ENTS\textsubscript{f}</th>
<th>HMEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDU\textsc{cat} (GPT-2\textsubscript{S})</td>
<td>39.82</td>
<td>72.35</td>
<td>31.72</td>
<td>35.31</td>
</tr>
<tr>
<td>EDU\textsc{cat} (GPT-2\textsubscript{M})</td>
<td>44.05</td>
<td>74.06</td>
<td>32.28</td>
<td>37.26</td>
</tr>
<tr>
<td>(\mathcal{X}_{\text{Coh}})</td>
<td>44.20</td>
<td>74.27</td>
<td>31.44</td>
<td>36.74</td>
</tr>
<tr>
<td>(-\text{conflict detection})</td>
<td>40.96</td>
<td>73.61</td>
<td>30.79</td>
<td>35.16</td>
</tr>
<tr>
<td>(-\text{both})</td>
<td>41.34</td>
<td>73.82</td>
<td>29.80</td>
<td>34.63</td>
</tr>
<tr>
<td>(+\mathcal{X}<em>{\text{Coh}}\text{ w/ ENTS}</em>{b})</td>
<td>43.65</td>
<td>74.09</td>
<td><strong>42.03</strong></td>
<td><strong>42.83</strong></td>
</tr>
</tbody>
</table>

Table 5: Ablation study of EDU\textsc{cat} in terms of conflict detection module and coherence score \(\mathcal{X}_{\text{Coh}}\). We also change the \(P_{\text{coh}}\) in \(\mathcal{X}_{\text{Coh}}\) to the trained discriminative metric ENTS\textsc{core}.

- Both conflict detection and coherence objective work for the task.
- Can be further improved with a more sophisticated coherence checking model for \(P_{\text{coh}}(\cdot)\) in \(\mathcal{X}_{\text{coh}}(\cdot)\).

(Please check the paper for details.)
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 had gotten lower in the sky.
S'4: She decided next time it was so hot she needed to go where there was air conditioning.
S'5: So she planned to go inside a nearby cafe.
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 had gotten lower in the sky.
S'4: She decided next time it was so hot she needed to go where there was air conditioning.
S'5: So she planned to go inside a nearby cafe.

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.
### 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 had gotten lower in the sky. |
| S'4: She decided next time it was so hot she needed to go where there was air conditioning. |
| S'5: So she planned to go inside a nearby cafe. |

| 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. |

| S'3: The sun was **high** in the sky. |
| S'4: She decided she needed to go somewhere where there was **air**. |
| S'5: She went to the **beach**. |
### 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 had gotten lower in the sky. | S'4: She decided next time it was so hot she needed to go where there was air conditioning. |
| S'5: So she planned to go inside a nearby cafe. |

| 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. |

| S'3: The sun was **high** in the sky. | S'4: She decided she needed to go somewhere where there was **air**. |
| S'5: She went to the **beach**. |

| 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. |
Takeaways

• Editing-based methods are also well-suited for generative counterfactual reasoning.

• Better components for EDUCAT will benefit unsupervised story rewriting.
  – e.g., better coherence score, more desired properties, etc.

• Better metrics should be studied for evaluating this task!
Have Fun with EDUCAT!

Checkout our code at GitHub!
https://github.com/jiangjiechen/EDUCAT