Unsupervised Editing For Counterfactual Stories

Jiangjie Chen¹,², Chun Gan², Sijie Cheng¹, Hao Zhou³, Yanghua Xiao¹ and Lei Li⁴
¹Fudan University ²JD.com ³ByteDance AI Lab ⁴University 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. y: Initial condition
   3. y: Story ending
   4. x: Counterfactual condition
   5. y: Counterfactual story ending

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

EDUCAT: Editing a New Story Ending

❖ EDUCAT: Unsupervised Constrained Editing via MCMC Sampling
   1. Define desired properties as stationary distribution \( x(y) \)
   2. Move \( y \) to \( y' \) by generating from the proposal distribution \( P(y'|y) \)
   3. Accept a proposal with acceptance rate \( \alpha(y,y') = \frac{1}{1 + \exp(-\alpha(y,y'))} \)
   4. Iterate until convergence
   5. Rank the accepted ones with \( \pi(t) \)

❖ Desired properties for stationary distribution \( x(y) \)
   1. Fluency: sentence probability from GPT-2
   2. Coherence: Punish proposed endings contradictory to the counterfactual conditions but consistent with the initial ones.

❖ Make an Edit Proposal
   1. Where to edit? — conflict token detection in \( y \)
   2. Edit with what? — modification actions
      ❖ Replace: mask-predict with an MLM (e.g., BERT)
      ❖ Insert: insert a [MASK], then do Replace
      ❖ Delete: reverse of Insert

Dataset

❖ TimeTravel [Qin et al. 2019]

❖ 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.

Case Study

❖ 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.
   - Therefore, she was very happy.

❖ S2: She went inside a nearby cafe.
   - The sun was still high in the sky.
   - She decided she needed to go outside and get some fresh air.
   - She went inside and got some fresh air.

❖ S3: She was playing for so long without beating the level.
   - S4: She never beat the last level.
   - S5: She was so happy to finally beat it.

❖ S3: She was playing for so long without beating the level.
   - S4: She never beat the last level.
   - S5: She was so happy to finally beat it.

❖ S3: She was playing for so long without beating the level.
   - S4: She never beat the last level.
   - S5: She was so happy to finally beat it.

❖ S3: She was playing for so long without beating the level.
   - S4: She never beat the last level.
   - S5: She was so happy to finally beat it.

❖ S3: She was playing for so long without beating the level.
   - S4: She never beat the last level.
   - S5: She was so happy to finally beat it.

❖ S3: She was playing for so long without beating the level.
   - S4: She never beat the last level.
   - S5: She was so happy to finally beat it.