国际人工智能会议
AAAI 2021论文北京预讲会

Sketch and Customize: A Counterfactual Story Gene

Changying Hao[1,2], Liang Pang[1,2*], Yanyan Lan[1,2], Yan Wang[3], Jiafeng Guo[1,2], Xueqi Cheng[1,2]

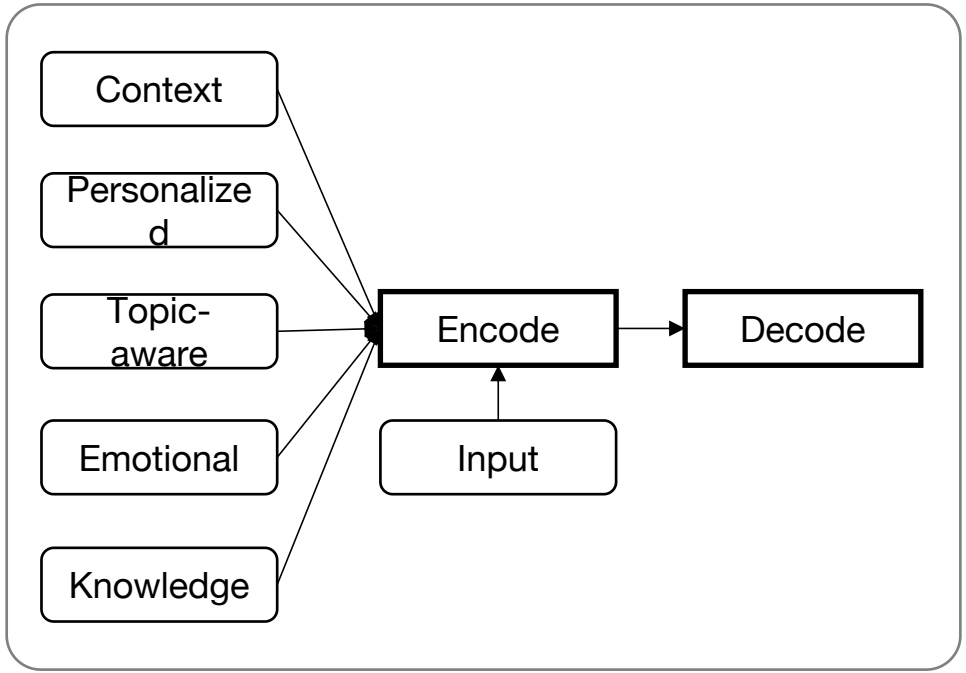
1. CAS Key Lab of Network Data Science and Technology,

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

2. University of Chinese Academy of Sciences, Beijing, China

3. Tencent AI Lab, Shenzhen, China

Conditional Text Generation



- Conditional text generation has been a research hotspot in recent years.
- Most of those tasks focus on how to generate good texts under certain conditions, while few work concentrates on how the consequence changes when the condition is modified.
- The causal relationship between condition and its corresponding generation is not well studied in these tasks.

INTRODUCTION

Counterfactual Story Generation Task -- test the causal reasoning ability for text generation models

Cares about how to revise an original story ending guided by a modified condition.

Premise

Mary loved flowers.

Original Story

Original Condition

She decided to go to the **park**.

Original Ending

She smelled all the flowers growing in the **field**.
She **picked** a few flowers and brought them home.
She placed the flowers in a vase with water.

Counterfactual Story

Counterfactual Condition

She decided to go to the **florist**.

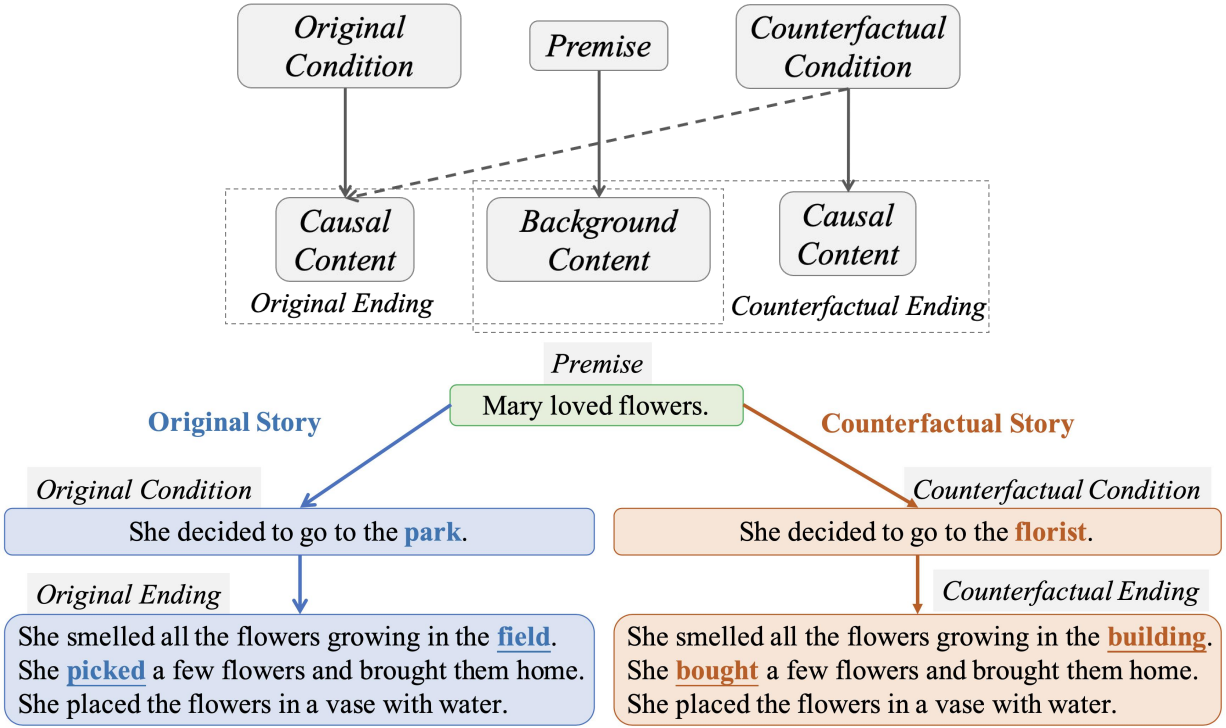
Counterfactual Ending

She smelled all the flowers growing in the **building**.
She **bought** a few flowers and brought them home.
She placed the flowers in a vase with water.

INTRODUCTION

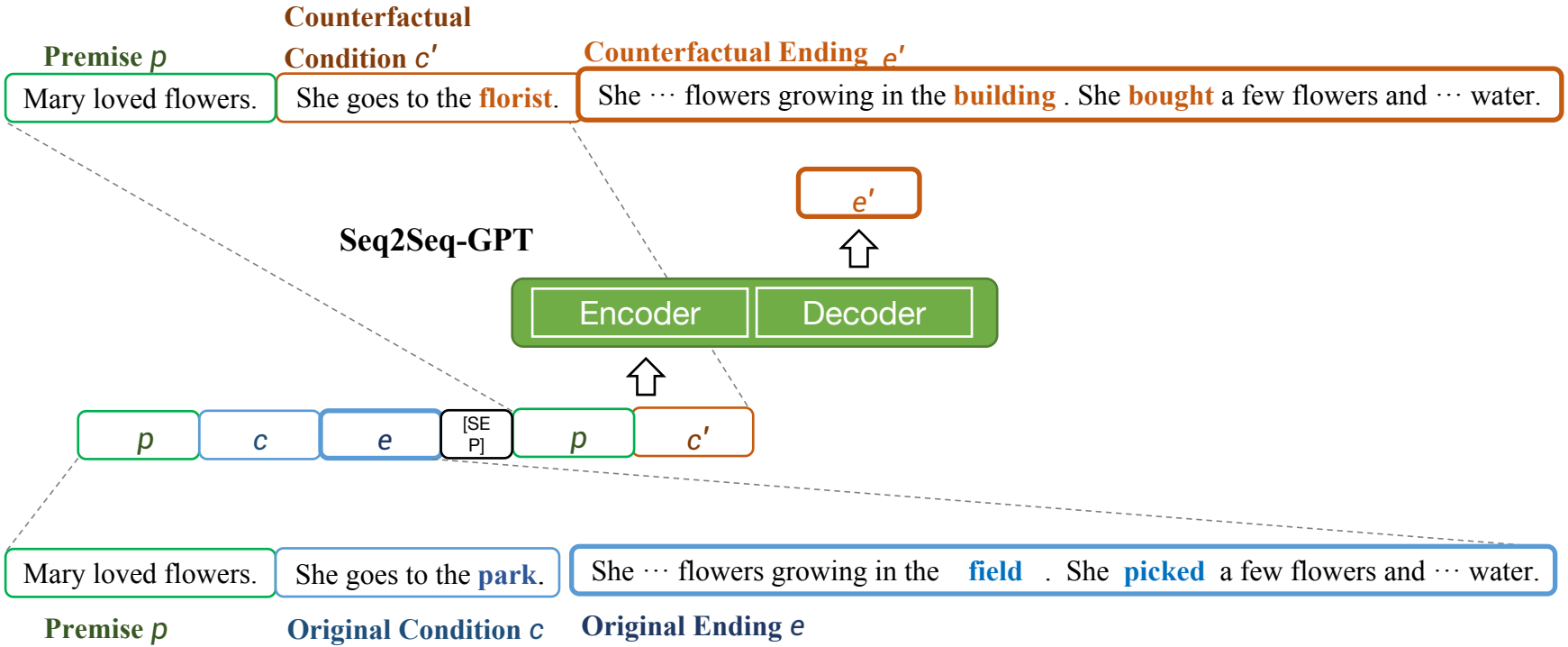
How to generate consistent counterfactual endings?

Our Idea: Explore the causality in text generation.



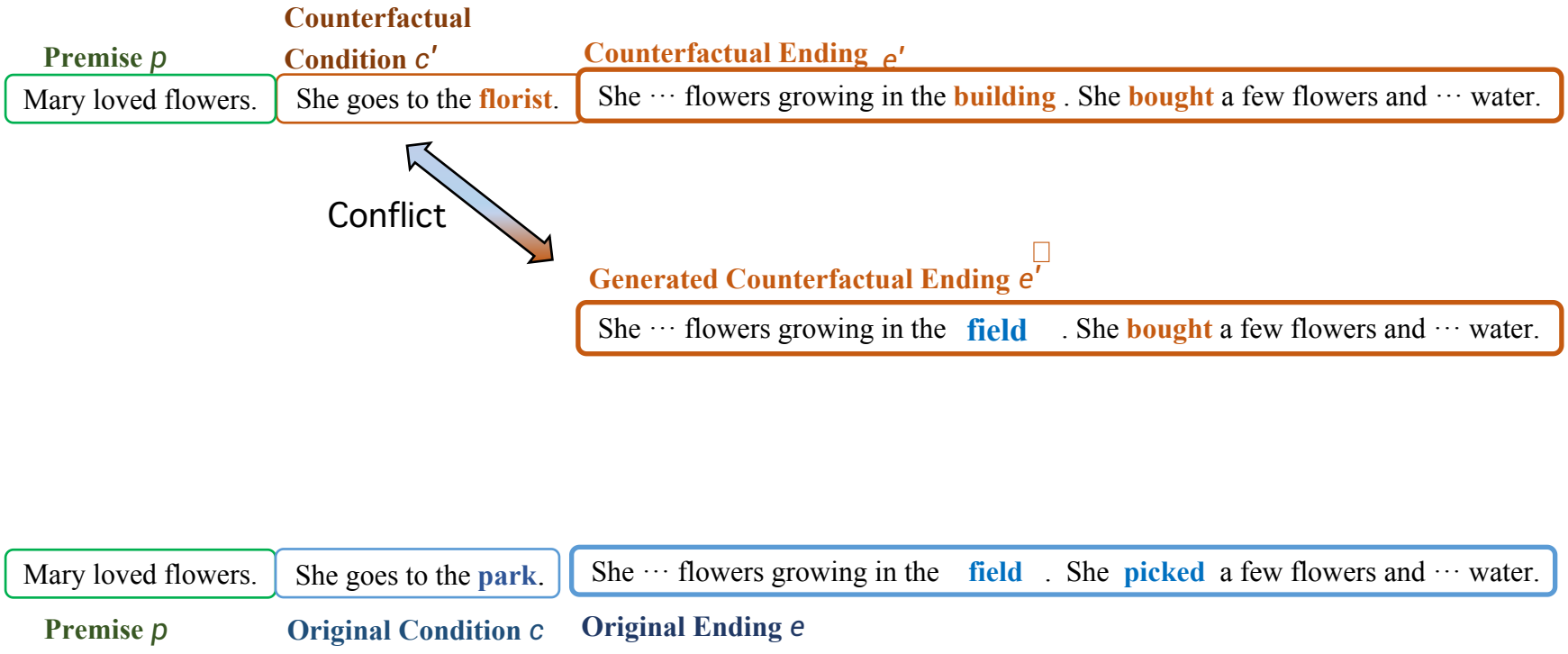
INTRODUCTION

How do previous works do such task?



INTRODUCTION

Problem



INTRODUCTION

Reasons

Premise p | **Counterfactual Condition** c' | **Counterfactual Ending** e'

Mary loved flowers. | She goes to the **florist**. | She ⋯ flowers growing in the **building** . She **bought** a few flowers and ⋯ water.

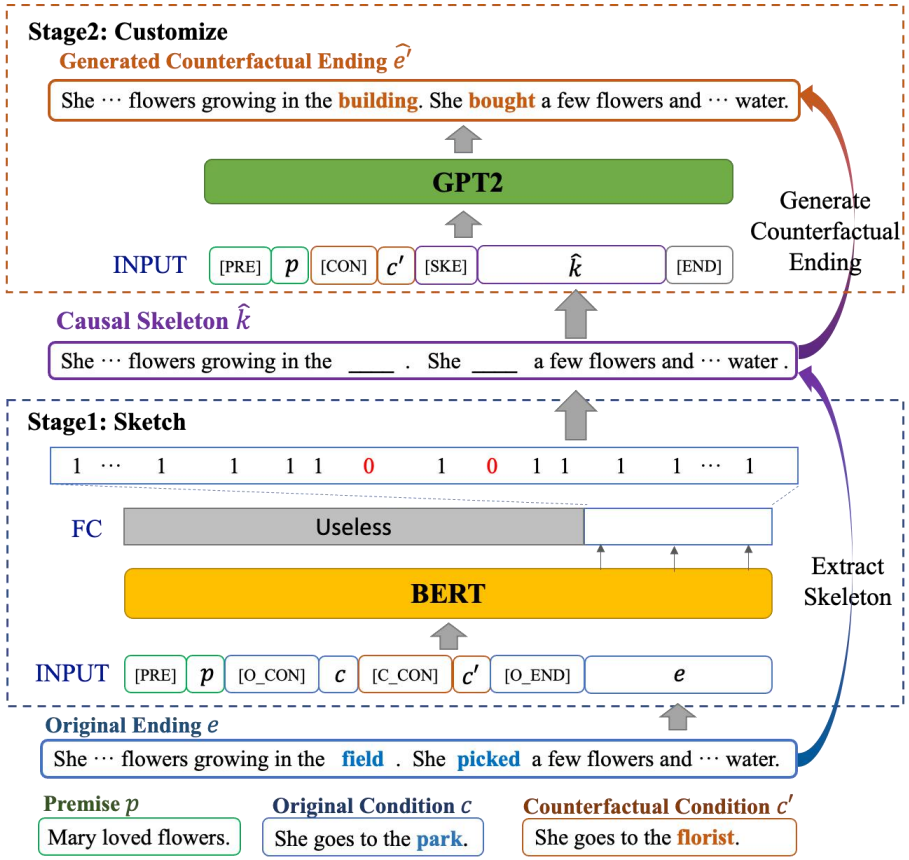
- Most words are the same. Copying will lead to a low MLE loss.
- Have difficulty in distinguishing words of different importance.
- Captures some spurious correlations between the two endings, instead of the causal relations between

Mary loved flowers. | She goes to the **park**. | She ⋯ flowers growing in the **field** . She **picked** a few flowers and ⋯ water.

Premise p | **Original Condition** c | **Original Ending** e

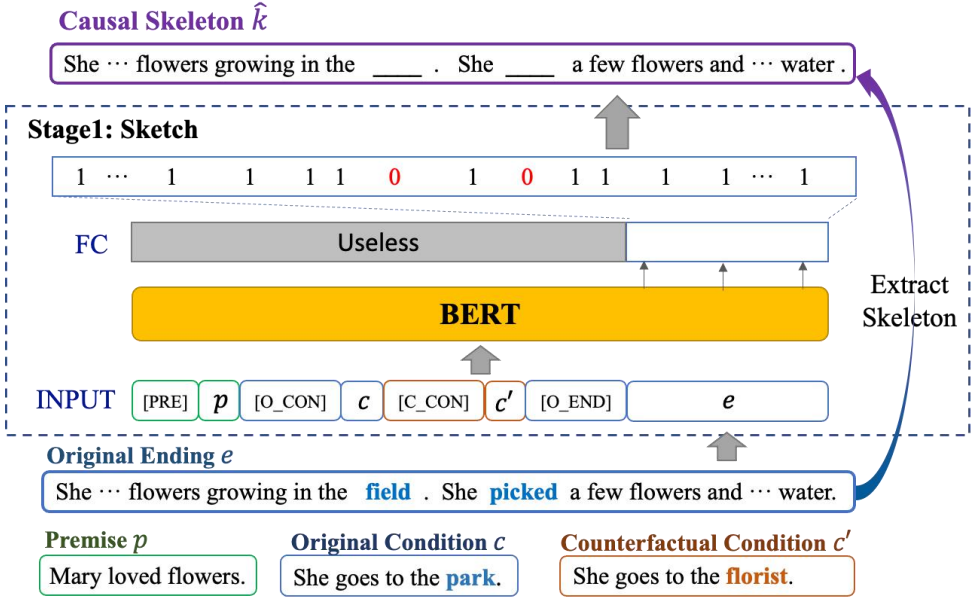
OUR APPROACHES

Two-stage Model -- Sketch & Customize



OUR APPROACHES

Sketch – Causality Detection with BERT



- BERT representation

$$R = \{r_1, \cdots, r_{N_r}\} \in \mathbb{R}^{N_r \times E}$$

- Sequence Labelling

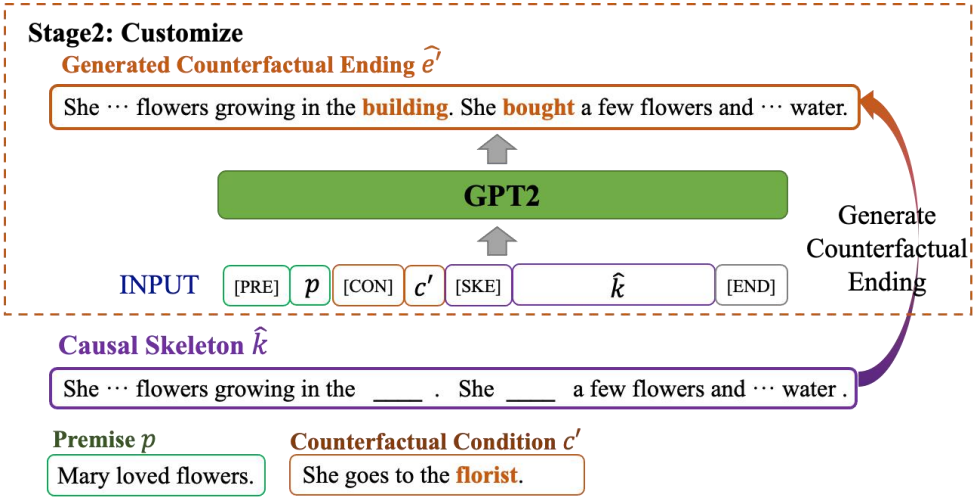
$$p_1(l_i|S) = \text{softmax}(Wr_i + b),$$

$$l_i = \begin{cases} 0, & i \in e_{causal} \\ 1, & i \notin e_{causal} \end{cases}, \ i \in [N_e, N_r]$$

- Replace the causal content with blanks, then we get the causal skeleton.

OUR APPROACHES

Customize – Counterfacutal Ending Generation with GPT2



- Ending Generation

$$p_2(y_t|x, y_{<t}) = \mathrm{GPT2}(x, y_{<t})$$

- Original Condition is not provided.
  Model can generate consistent words in
  the skeleton.

OUR APPROACHES

Training & Inference

- Training
  1. Sequence labelling -- weighted cross-entropy loss

$$\mathcal{L}_{seq} = -\sum_{i=N_e}^{N_r} [\lambda \log p_1(l_i = 0|S) + (1-\lambda) \log p_1(l_i = 1|S)]$$

  2. Ending generation – negative log likelihood loss

$$\mathcal{L}_{gen} = -\sum_{t=1}^{m} \log[p_2(e'_t|p, c', k, e'_{<t})]$$

  Causal Skeleton Augmentation for the Customize stage

  1) randomly replacing 20% of background words with blanks,
  2) randomly replacing 20% of background words with words sampled from the vocabulary,
  3) randomly shuffling the order of 20% of background words

- Inference

  1. Predict the label for each word in the original ending

$$\hat{l_i} = \underset{l_i \in \{0,1\}}{\arg\max}\ p_1(l_i|S)$$

  2. Merge the consecutive blanks into one blank and get the predicted skeleton

  3. Generate the counterfactual ending using top-k sampling

$$\hat{e'_t} = \underset{e'_t \in V}{\text{sample}}\ p_2(e'_t|p, c', \hat{k}, e'_{<t})$$

EXPERIMENTAL RESULTS

Human Evaluation

	PRE	CF	PLOT	Avg.
Seq2Seq-GPT	2.558	1.985$^{\downarrow}$	2.170	2.238
Random&C	2.572	1.905$^{\downarrow}$	2.132	2.203
LCS&C	2.542	2.083	2.145	2.257
S&C-0.5	**2.650**	1.668$^{\downarrow}$	**2.425**$^{\uparrow}$	2.248
S&C-0.8	2.590	**2.130**	2.120	**2.280**
S&C-w/o-Aug	2.458$^{\downarrow}$	2.030	1.845$^{\downarrow}$	2.111
Human	2.610	2.217	2.252$^{\uparrow}$	2.360

PRE: Consistency and relevance to the premise.
CF: Consistency to the counterfactual condition.
Plot: Similarity to the plot of the original ending.

- S&C-0.8 model get the highest CF score among all the methods, it is 0.145 higher than Seq2Seq- GPT. It can generate more consistent counterfactual endings.
- S&C-0.8 outperforms Seq2Seq-GPT on PRE, it can generate consistent and relevant endings to the premise.
- S&C-0.8 model is evaluated less similar with the original ending compared to Seq2Seq-GPT baseline because of the copy strategy of Seq2Seq-GPT.

EXPERIMENTAL RESULTS

Human Evaluation

	PRE	CF	PLOT	Avg.
Seq2Seq-GPT	2.558	1.985$^{\downarrow}$	2.170	2.238
Random&C	2.572	1.905$^{\downarrow}$	2.132	2.203
LCS&C	2.542	2.083	2.145	2.257
S&C-0.5	**2.650**	1.668$^{\downarrow}$	**2.425**$^{\uparrow}$	2.248
S&C-0.8	2.590	**2.130**	2.120	**2.280**
S&C-w/o-Aug	2.458$^{\downarrow}$	2.030	1.845$^{\downarrow}$	2.111
Human	2.610	2.217	2.252$^{\uparrow}$	2.360

PRE: Consistency and relevance to the premise.
CF: Consistency to the counterfactual condition.
Plot: Similarity to the plot of the original ending.

- Causal skeleton is important.

1) S&C outperform Random&C which using random skeletons.

EXPERIMENTAL RESULTS

Human Evaluation

	PRE	CF	PLOT	Avg.
Seq2Seq-GPT	2.558	1.985$^{\downarrow}$	2.170	2.238
Random&C	2.572	1.905$^{\downarrow}$	2.132	2.203
LCS&C	2.542	2.083	2.145	2.257
S&C-0.5	**2.650**	1.668$^{\downarrow}$	**2.425**$^{\uparrow}$	2.248
S&C-0.8	2.590	**2.130**	2.120	**2.280**
S&C-w/o-Aug	2.458$^{\downarrow}$	2.030	1.845$^{\downarrow}$	2.111
Human	2.610	2.217	2.252$^{\uparrow}$	2.360

PRE: Consistency and relevance to the premise.
CF: Consistency to the counterfactual condition.
Plot: Similarity to the plot of the original ending.

- Causal skeleton is important.

1) S&C outperform Random&C which uses random skeletons.

2) S&C-0.8 behaves similarly to the LCS&C which uses LCS skeletons.

EXPERIMENTAL RESULTS

Human Evaluation

	PRE	CF	PLOT	Avg.
Seq2Seq-GPT	2.558	1.985$^{\downarrow}$	2.170	2.238
Random&C	2.572	1.905$^{\downarrow}$	2.132	2.203
LCS&C	2.542	2.083	2.145	2.257
S&C-0.5	**2.650**	1.668$^{\downarrow}$	**2.425**$^{\uparrow}$	2.248
S&C-0.8	2.590	**2.130**	2.120	**2.280**
S&C-w/o-Aug	2.458$^{\downarrow}$	2.030	1.845$^{\downarrow}$	2.111
Human	2.610	2.217	2.252$^{\uparrow}$	2.360

PRE: Consistency and relevance to the premise.
CF: Consistency to the counterfactual condition.
Plot: Similarity to the plot of the original ending.

- S&C-0.8 outperforms S&C-0.5 on the CF metric significantly, λ is important for the task.

1) Help solve the label imbalance problem

2) Higher loss weight causes skeletons with more blanks, leaving more spaces for generation model.

EXPERIMENTAL RESULTS

Human Evaluation

	PRE	CF	PLOT	Avg.
Seq2Seq-GPT	2.558	1.985$^{\downarrow}$	2.170	2.238
Random&C	2.572	1.905$^{\downarrow}$	2.132	2.203
LCS&C	2.542	2.083	2.145	2.257
S&C-0.5	**2.650**	1.668$^{\downarrow}$	**2.425**$^{\uparrow}$	2.248
S&C-0.8	2.590	**2.130**	2.120	**2.280**
S&C-w/o-Aug	2.458$^{\downarrow}$	2.030	1.845$^{\downarrow}$	2.111
Human	2.610	2.217	2.252$^{\uparrow}$	2.360

PRE: Consistency and relevance to the premise.
CF: Consistency to the counterfactual condition.
Plot: Similarity to the plot of the original ending.

- Skeleton augmentation is important.

1) S&C-0.8 achieves better scores in all of these three aspects than S&C-w/o-Aug

2) Using only the LCS skeletons to train the customize stage leads to overfitting and limits model generation capabilities.

CONCLUSION

- Revisit the text generation task in a causal perspective, where the generated text is split into the background and causal parts, which is related to the premise and the changed condition respectively;
- Propose a Sketch and Customize framework for improving the causal reasoning ability of the text generation models;
- Conduct experiments on a counterfactual story rewriting task to verify the performance of our proposed framework.

THANKS

Name: Changying Hao

Email: haochangying18s@ict.ac.cn

2020.12.19