Research Progress on Story Generation

Wei Wang[1,2]

1. Graduate School at Shenzhen, Tsinghua University

2. Tencent AI Lab, NLP Center

2019.09

Overview

- Do Massively Pretrained Language Models Make Better Storytellers? (CoNLL 2019)
- Counterfactual Story Reasoning and Generation (EMNLP 2019)

Do Massively Pretrained Language Models Make Better Storytellers?

Abigail See, Aneesh Pappu*, Rohun Saxena*, Akhila Yerukola*, Christopher D. Manning
Stanford University
{abisee,apappu,rohun,akhilay,manning}@cs.stanford.edu

Introduction

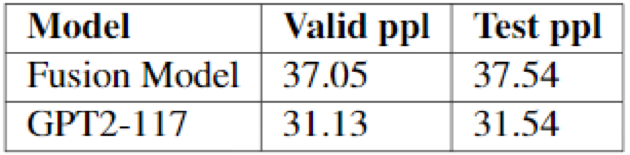
Do Massively Pretrained Language Models Make Better Storytellers? (CoNLL 2019)

- Large neural language models trained on massive amounts of text have emerged as a formidable strategy for Natural Language Understanding tasks. However, the strength of these models as Natural Language Generators is less clear.
- In this work, we compare the performance of an extensively pretrained model, OpenAI GPT2-117 (Radford et al., 2019), to a state-of-the-art neural story generation model (Fan et al., 2018).
- we prioritize evaluating text across the whole k spectrum, and measuring many different automatic metrics, rather than a few human metrics.

Experiment

Do Massively Pretrained Language Models Make Better Storytellers? (CoNLL 2019)

- **WritingPrompts dataset.** WritingPrompts (Fan et al., 2018) is a story generation dataset containing 303,358 human-written (prompt, story) pairs collected from the /r/WritingPrompts subreddit.
- **The Fusion Model**. The Fusion Model is a state-of-the-art neural story generation architecture trained on the WritingPrompts dataset (Fan et al., 2018).
- **GPT2-117**. GPT2 (Radford et al., 2019) is a large Transformer language model trained on WebText, a diverse corpus of internet text (not publicly released) containing over 8 million documents equalling 40GB of text in total.



Model	Valid ppl	Test ppl
Fusion Model	37.05	37.54
GPT2-117	31.13	31.54

Table 1: Word-level perplexities on WritingPrompts-1024 for the Fusion Model and finetuned GPT2-117.

Experiment-Story-prompt relatedness

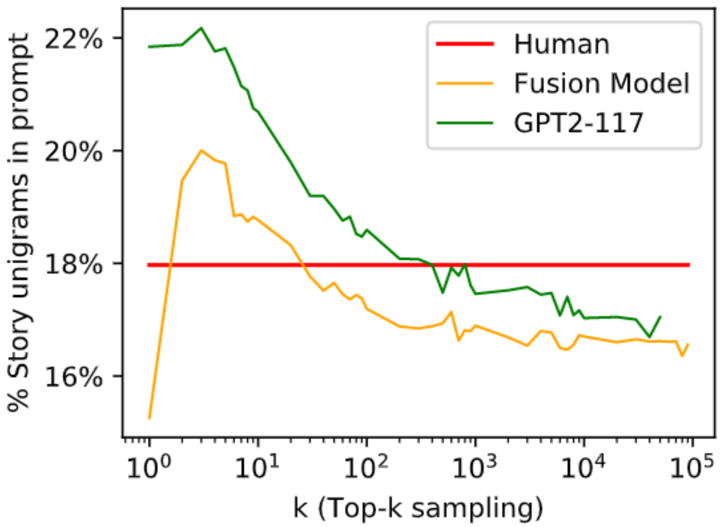
Do Massively Pretrained Language Models Make Better Storytellers? (CoNLL 2019)

Prompt ranking accuracy.

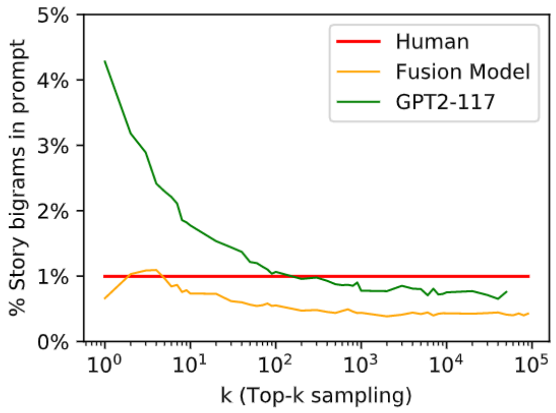
- The prompt ranking accuracy of a model is the percentage of cases in which the model assigns a higher probability to the story under its true prompt than under all of the other nine.
- We find that GPT2-117 scores 80.16% on this task, while the Fusion Model scores 39.8%.5 Random chance scores 10%. GPT2-117 conditions on the prompt much more strongly than the Fusion Model.

N-gram similarity

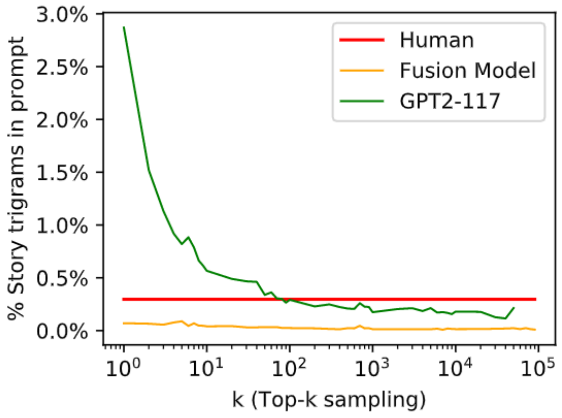
- For n = 1; 2; 3, we measure the percentage of generated n-grams that also appear in the prompt.
- For all n and k, we find that GPT2-117 has a higher overlap (i.e. copies more from the prompt) than the Fusion Model



(a) Percent of all story unigrams that are in the prompt.



(b) Percent of all story bigrams that are in the prompt.



(c) Percent of all story trigrams that are in the prompt.

Experiment-Story-prompt relatedness

Do Massively Pretrained Language Models Make Better Storytellers? (CoNLL 2019)

Sentence embedding similarity

- To capture a higher-level notion of semantic similarity, we measure story-prompt sentence similarity – the cosine similarity of story-prompt sentence pairs, averaged by taking the mean over all pairs
- GPT2-117 generates sentences that are more similar to the prompt than the Fusion Model for all k, and both models' prompt similarity decreases as k increases.

Named entity usage

- GPT2-117 uses more of the prompt named entities than the Fusion Model (as well as more named entities overall), but both models use fewer named entities than humans when k is less than vocabulary size

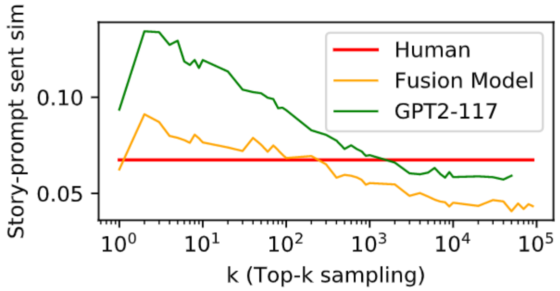
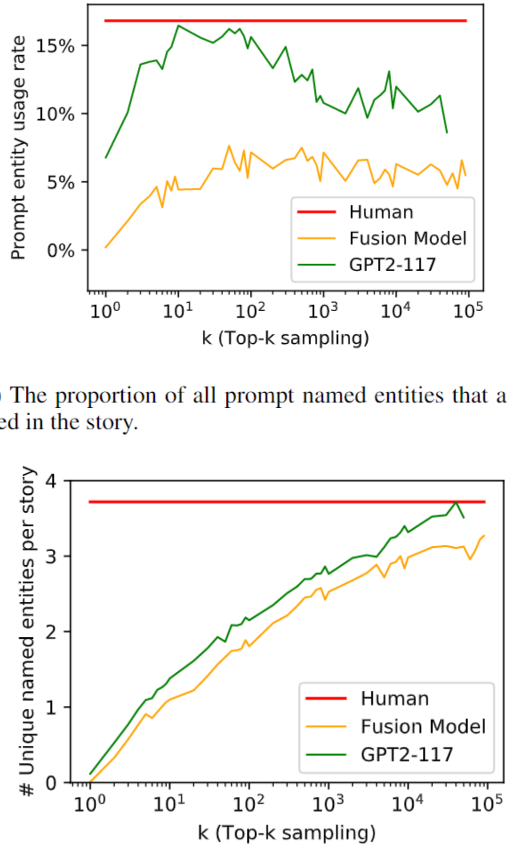


Figure 1: Compared to the Fusion Model, GPT2-117 produces stories that are more semantically similar to the prompt. Similarity decreases as k increases.



(a) The proportion of all prompt named entities that are used in the story.

(b) The number of unique named entities that appear in the story.

Experiment-Coherence

Do Massively Pretrained Language Models Make Better Storytellers? (CoNLL 2019)

- measuring its ability to rank shuffled human written text as less coherent than the original unshuffled text.
- Both models perform well on this task – the Fusion Model has an error rate of 3.44% and GPT2-117 an error rate of 2.17%. This 36.92% error reduction indicates that GPT2-117 is more sensitive to ordering of events.
- This shows that both models are less sensitive to out-of-order sentences that occur at the beginning of the text, than those occurring later.

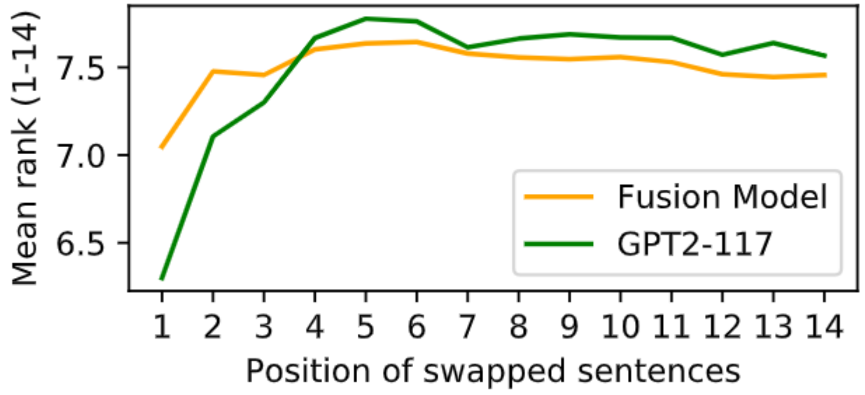


Figure 2: Sensitivity of the models to swapped sentences in different positions. A higher mean rank indicates higher sensitivity (i.e. the model assigns lower probability) relative to other positions. Both models are less sensitive to swapped sentences at the beginning of the text, compared to later. GPT2-117 shows this pattern more strongly, indicating greater use of context.

Experiment-Repetition and rareness

Do Massively Pretrained Language Models Make Better Storytellers? (CoNLL 2019)

N-gram repetition

- The distinct-n metric of a piece of text is the number of unique n-grams divided by the total number of generated n-grams (Li et al., 2016). We measure distinct-n of the generated stories for n = 1; 2; 3.
- both models' unigram diversity is far below that of human text when k is small. distinct-n increases as k increases, converging to a value close to the human level as k approaches vocabulary size. Though GPT2-117 has a slightly higher distinct-n than the Fusion Model for most values of k, the difference is negligible compared to the influence of k.

Rare word usage

- We compute the mean log unigram probability of the words in the generated story – a high value indicates using fewer rare words while a low value indicates more rare words.
- word rareness is primarily governed by k – however, GPT2-117 has a lower mean log unigram probability (i.e., uses more rare words) than the Fusion Model for all equal values of k ⩾ 2.

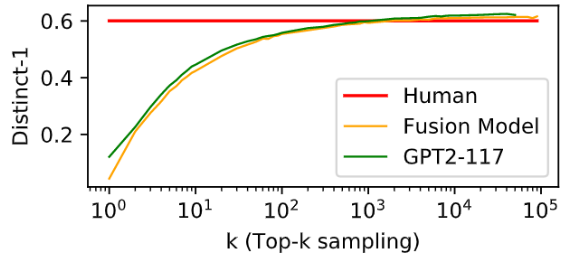
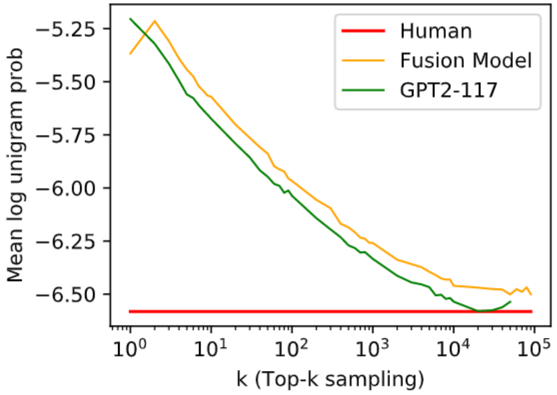


Figure 3: Repetition (low distinct-1) is primarily caused by choice of decoding algorithm (here low k), not insufficient training data. GPT2-117 is trained on 45× more data than the Fusion Model, but is similarly repetitive for all k.

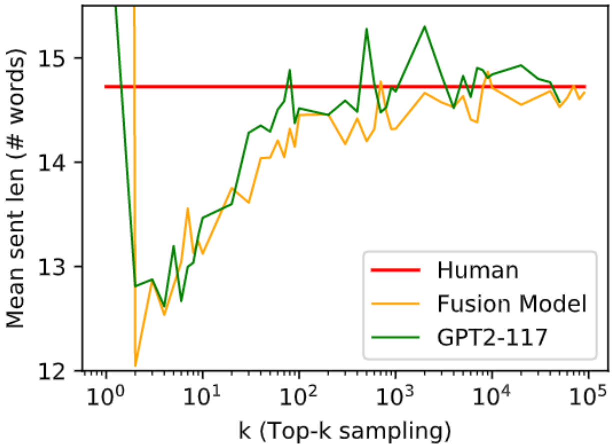


(a) The mean log unigram probability of generated words. Higher values indicate using fewer rare words while lower values indicate using more rare words.

Choice of decoding algorithm is a primary factor in diversity and repetition problems, with likelihood-maximizing algorithms the main culprit. the difference is small compared to the effect of *k*, indicating that training data alone is unlikely to solve these problems

Experiment-Syntactic style and complexity

Do Massively Pretrained Language Models Make Better Storytellers? (CoNLL 2019)

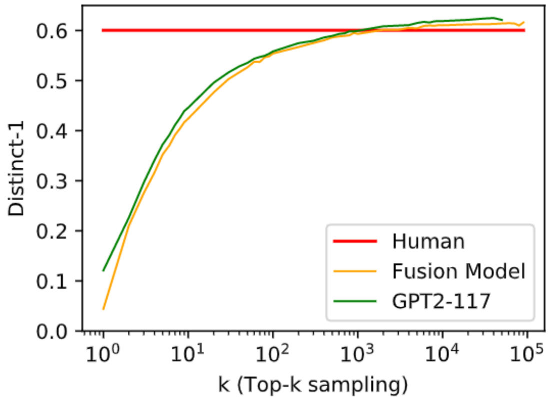


Sentence length Sentence

- length is a simple but effective feature to estimate readability and syntactic complexity of text.
- both models generate sentences that are on average shorter than human sentences when k is small, but converge to approximately human length as k increases.

Part-of-speech usage

- It has been shown that the distribution of parts-of-speech (POS), and more generally the distribution of POS n-grams11 is a useful feature to represent the style of a piece of text
- This implies that, as with lexical diversity, the models have no difficulty fitting the statistical distribution of human syntax.
- However, we note that as k increases, lexical diversity reaches human level sooner than syntactic diversity.



(a) Distinct-1 (ratio of unique unigrams in the story to total number of generated unigrams in the story).

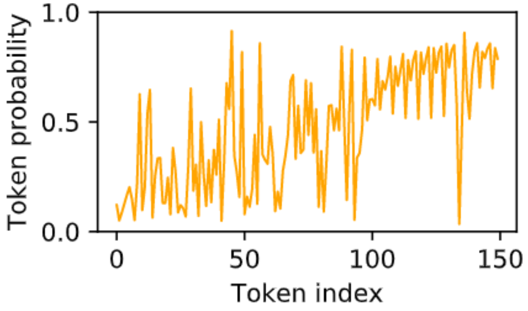
our results show that syntactic under-diversity is primarily caused by low k, not insufficient training data.

Experiment-The element of surprise

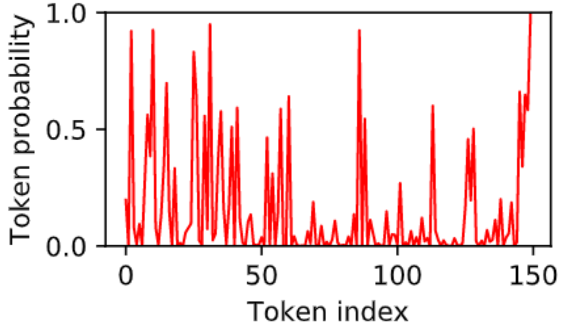
Do Massively Pretrained Language Models Make Better Storytellers? (CoNLL 2019)

Model confidence over time

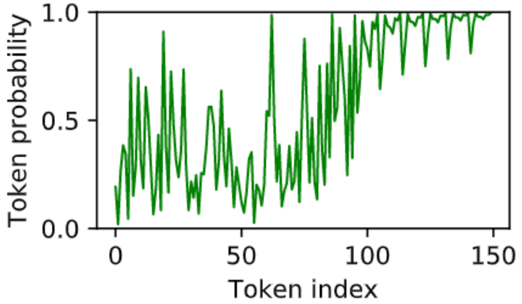
- Several researchers have observed that model overconfidence (the model placing high probability on a small range of tokens) can cause poor quality generation.
- both models fall into self-reinforcing repetitive loops with rising confidence.
- when generating with top-k sampling, the probabilities increase more rapidly, and the increase is even more rapid for smaller k.
- like repetition, model over-confidence is unlikely to be solved by more training data, and is largely governed by choice of k



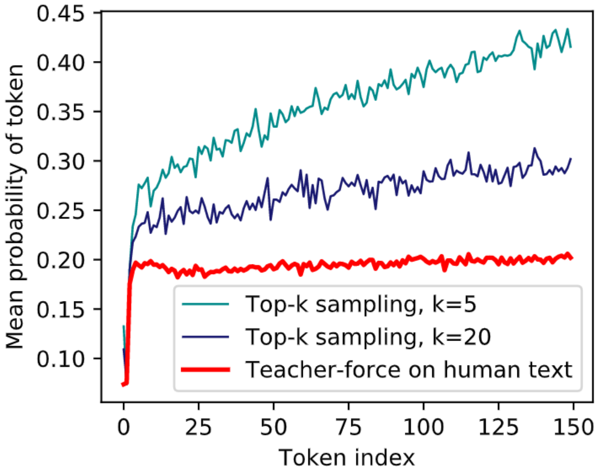
(a) **Fusion Model** ($k = 2$): *I had*



(b) **Human Text**: *“Looks like the*



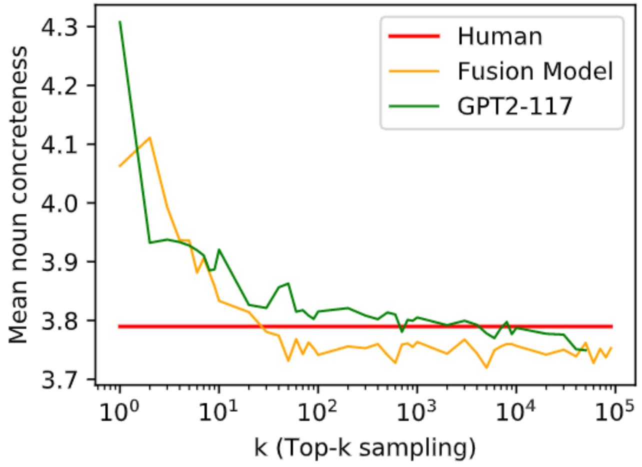
(c) **GPT2-117** ($k = 2$): *I’ve always*



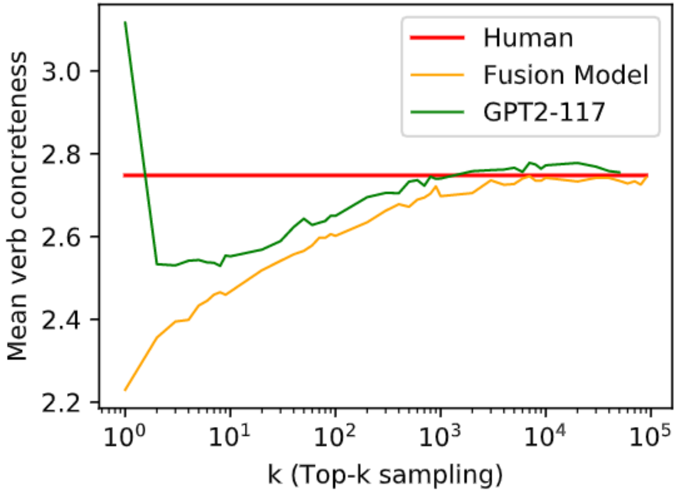
Experiment-Concreteness

Do Massively Pretrained Language Models Make Better Storytellers? (CoNLL 2019)

- the concreteness of a word as 'the degree to which the concept denoted by a word refers to a perceptible entity'.
- Brysbaert et al. provide human concreteness ratings for 40,000 common English lemmas rated on a scale from 1 to 5. We use these ratings to measure the mean concreteness of the nouns and verbs in the story text.
- for the same k, GPT2-117 tends to generate more concrete words than the Fusion Model, and that for both models, concreteness converges to approximately human levels as k increases.
- for small k, both models produce stories that, compared to human-written stories, have too many physical objects (as opposed to abstract nouns), and too few physical actions (as opposed to abstract verbs).



(a) Mean concreteness rating (1-5) of nouns in the story.



(b) Mean concreteness rating (1-5) of verbs in the story.

Summary

Do Massively Pretrained Language Models Make Better Storytellers? (CoNLL 2019)

The effect of massive pretraining

- GPT2-117 is a better story generation model than the Fusion Model in several specific ways: it conditions much more strongly on the provided context, is more sensitive to correct ordering of events, and generates text that is more contentful (using more rare words, concrete words, and named entities).

The effect of *k*

- The negative characteristics of low k output (genericness, repetition, oversimplicity) are by now familiar to researchers.
- As k increases to vocabulary size, we find that the model-generated text closely fits the human text on most of the metrics we measured. However, it is clear by inspection that the high-k model-generated text lacks many crucial aspects such as commonsense reasoning
- true progress in open-ended Natural Language Generation will come from attempting to address these high k problems.

Limitations of this study

- This study uses only the smallest version of GPT2.
- This study did not include human evaluation.

Counterfactual Story Reasoning and Generation

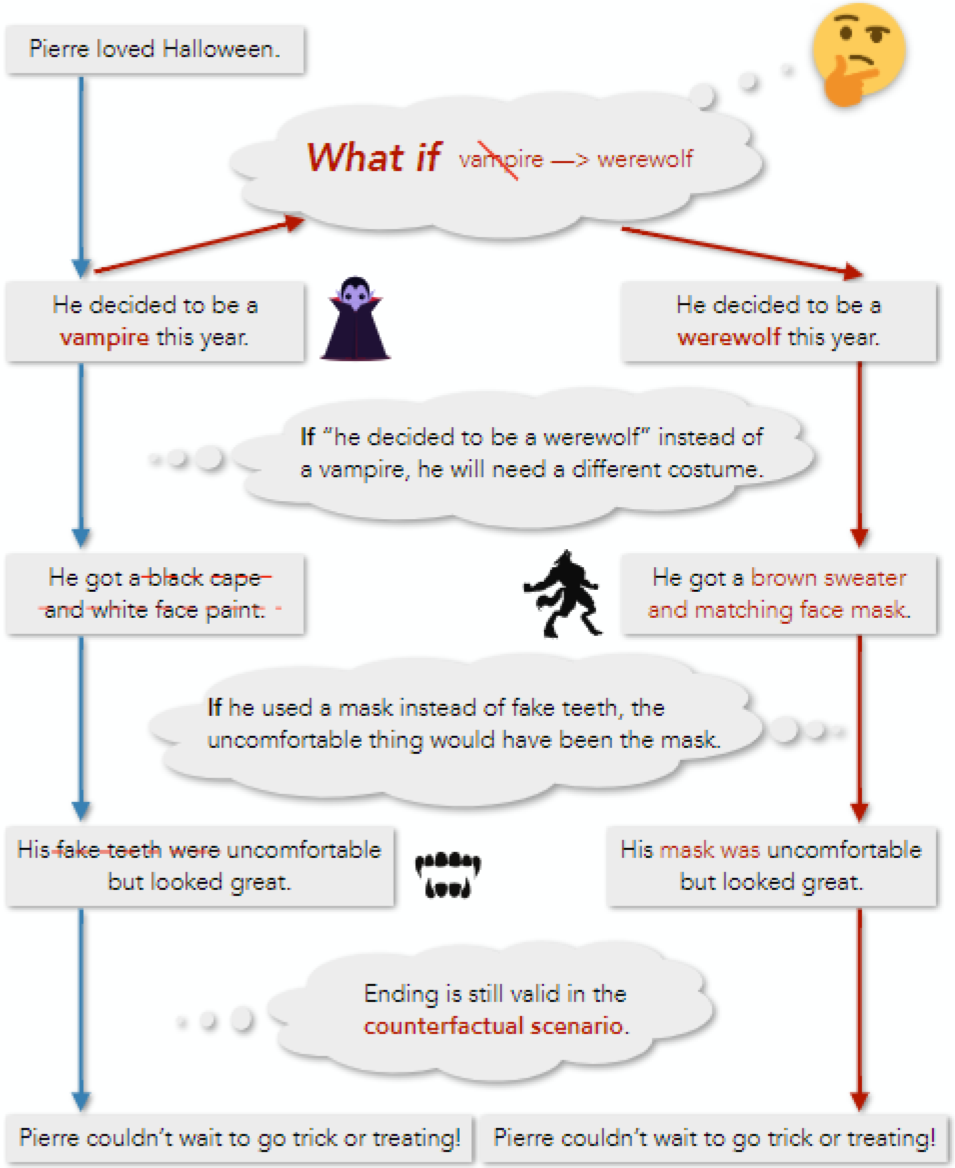
Lianhui Qin [♠◇] **Antoine Bosselut** [♠◇] **Ari Holtzman** [♠◇] **Chandra Bhagavatula** [◇]
Elizabeth Clark [♠] **Yejin Choi** [♠◇]

[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[◇]Allen Institute for Artificial Intelligence
{lianhuiq,antoineb,ahai,eaclark7,yejin}@cs.washington.edu
chandrab@allenai.org

Introduction

Counterfactual Story Reasoning and Generation (EMNLP 2019)

- Counterfactual reasoning requires predicting how alternative events, contrary to what actually happened, might have resulted in different outcomes.
- Counterfactual Story Rewriting: given an original story and an intervening counterfactual event, the task is to minimally revise the story to make it compatible with the given counterfactual event.



Pierre loved Halloween.

What if ~~vampire~~ —> werewolf

He decided to be a vampire this year.

He decided to be a werewolf this year.

If "he decided to be a werewolf" instead of a vampire, he will need a different costume.

He got ~~a black cape and white face paint.~~

He got a brown sweater and matching face mask.

If he used a mask instead of fake teeth, the uncomfortable thing would have been the mask.

His ~~fake teeth were~~ uncomfortable but looked great.

His mask was uncomfortable but looked great.

Ending is still valid in the counterfactual scenario.

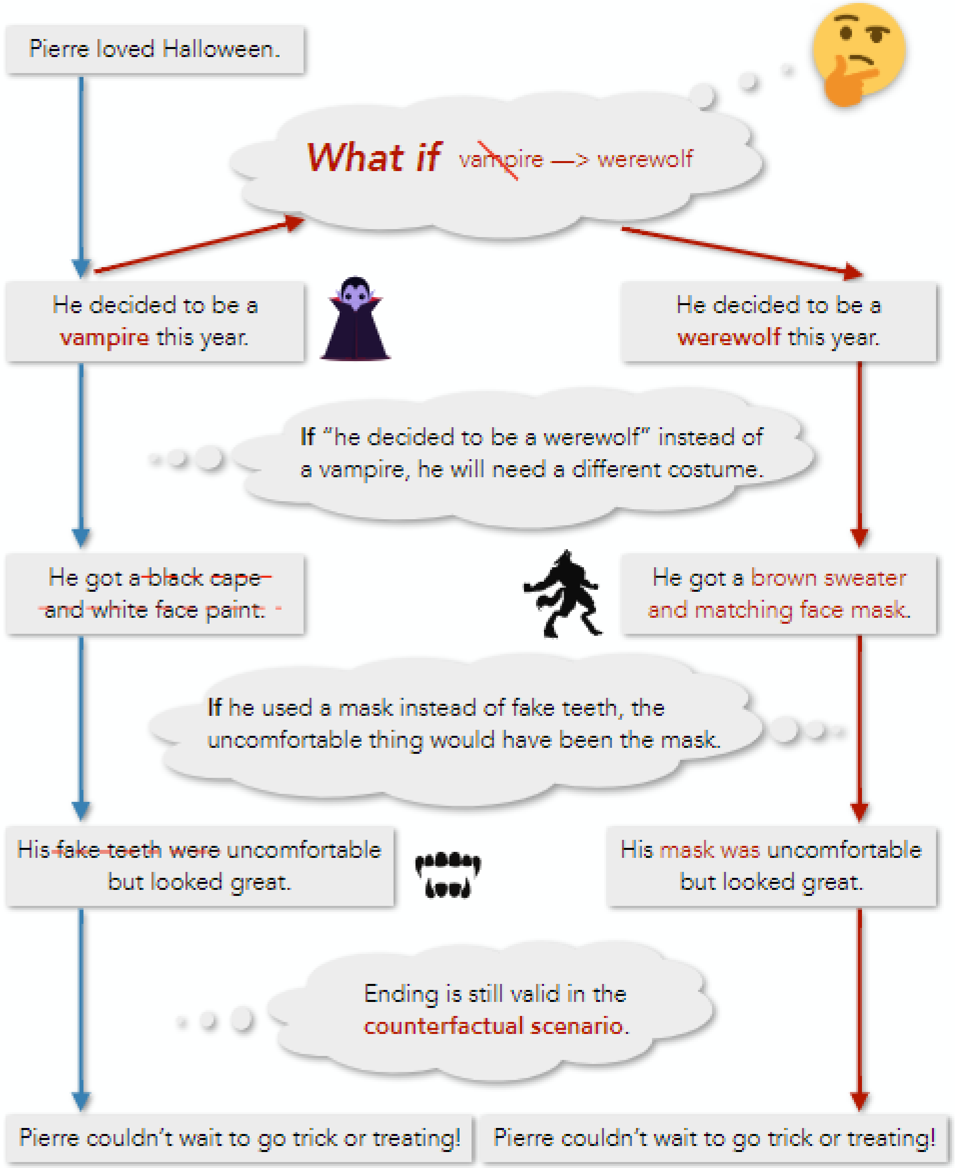
Pierre couldn't wait to go trick or treating!

Pierre couldn't wait to go trick or treating!

Introduction

Counterfactual Story Reasoning and Generation (EMNLP 2019)

- TIMETRAVEL, a new dataset of 29,849 counterfactual rewritings, each with the original story, a counterfactual event, and human-generated revision of the original story compatible with the counterfactual event.
- We evaluate the counterfactual rewriting capacities of several competitive baselines based on pretrained language models



Pierre loved Halloween.

What if ~~vampire~~ —> werewolf

He decided to be a vampire this year.

He decided to be a werewolf this year.

If "he decided to be a werewolf" instead of a vampire, he will need a different costume.

He got ~~a black cape and white face paint.~~

He got a brown sweater and matching face mask.

If he used a mask instead of fake teeth, the uncomfortable thing would have been the mask.

His ~~fake teeth were~~ uncomfortable but looked great.

His mask was uncomfortable but looked great.

Ending is still valid in the counterfactual scenario.

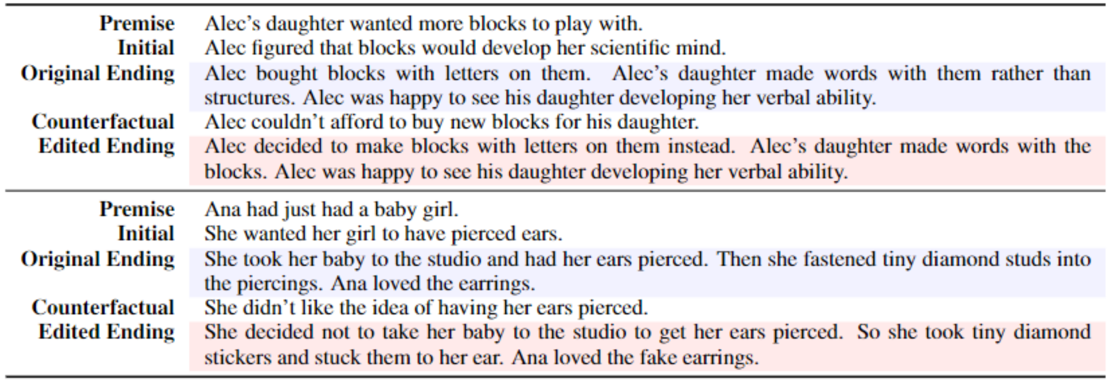
Pierre couldn't wait to go trick or treating!

Pierre couldn't wait to go trick or treating!

Data Collection

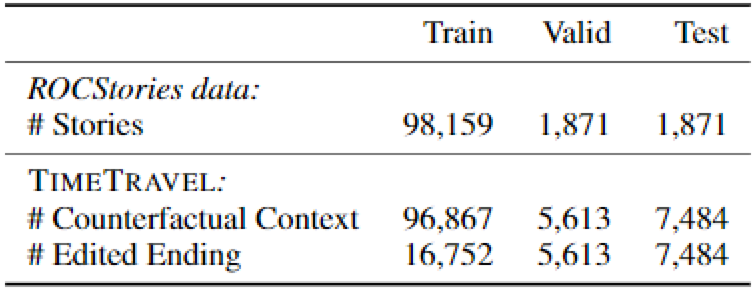
Counterfactual Story Reasoning and Generation (EMNLP 2019)

- Our dataset is built on top of the ROCStories corpus, which contains 98,159 five-sentences stories in the training set, along with 3,742 stories in the evaluation sets.
- **Counterfactual Event Collection**. We present workers with an original five-sentence story S =(s1; s2; : : : ; s5) and ask them to produce a counterfactual event s2` based on s2.
- **Continuation Rewriting.** Once a counterfactual sentence s2` is provided, we present it to a new set of workers with the original story S. Now that s2` invalidates the original storyline, workers are instructed to make minimal edits to s3:5, such that the narrative is coherent again.



Premise	Alec's daughter wanted more blocks to play with.
Initial	Alec figured that blocks would develop her scientific mind.
Original Ending	Alec bought blocks with letters on them. Alec's daughter made words with them rather than structures. Alec was happy to see his daughter developing her verbal ability.
Counterfactual	Alec couldn't afford to buy new blocks for his daughter.
Edited Ending	Alec decided to make blocks with letters on them instead. Alec's daughter made words with the blocks. Alec was happy to see his daughter developing her verbal ability.
Premise	Ana had just had a baby girl.
Initial	She wanted her girl to have pierced ears.
Original Ending	She took her baby to the studio and had her ears pierced. Then she fastened tiny diamond studs into the piercings. Ana loved the earrings.
Counterfactual	She didn't like the idea of having her ears pierced.
Edited Ending	She decided not to take her baby to the studio to get her ears pierced. So she took tiny diamond stickers and stuck them to her ear. Ana loved the fake earrings.

Table 1: Examples from TIMETRAVEL

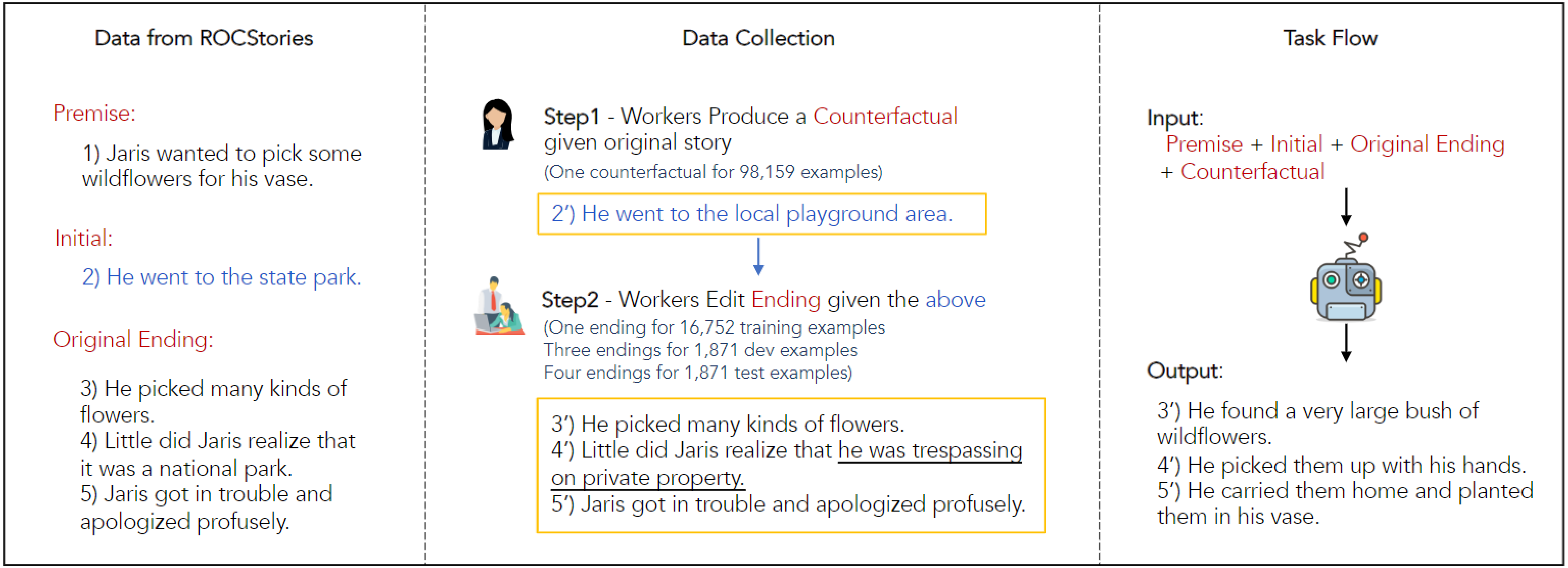


	Train	Valid	Test
ROCStories data:			
# Stories	98,159	1,871	1,871
TIMETRAVEL:			
# Counterfactual Context	96,867	5,613	7,484
# Edited Ending	16,752	5,613	7,484

Table 2: Dataset statistics

Data Collection

Counterfactual Story Reasoning and Generation (EMNLP 2019)



Data from ROCStories

Premise:

1) Jaris wanted to pick some wildflowers for his vase.

Initial:

2) He went to the state park.

Original Ending:

3) He picked many kinds of flowers.
4) Little did Jaris realize that it was a national park.
5) Jaris got in trouble and apologized profusely.

Data Collection

Step1 - Workers Produce a Counterfactual given original story
(One counterfactual for 98,159 examples)

2') He went to the local playground area.

Step2 - Workers Edit Ending given the above
(One ending for 16,752 training examples
Three endings for 1,871 dev examples
Four endings for 1,871 test examples)

3') He picked many kinds of flowers.
4') Little did Jaris realize that he was trespassing on private property.
5') Jaris got in trouble and apologized profusely.

Task Flow

Input:
Premise + Initial + Original Ending + Counterfactual

Output:

3') He found a very large bush of wildflowers.
4') He picked them up with his hands.
5') He carried them home and planted them in his vase.

Experiment

Counterfactual Story Reasoning and Generation (EMNLP 2019)

Unsupervised Training

- **Zero-shot** (ZS) In our simplest setting, we evaluate the counterfactual reasoning abilities already learned by these models due to pretraining on large corpora.
- **Fine-tuning** (FT) In this setting, the model is further fine-tuned to maximize the loglikelihood of the stories in the ROCStories corpus.

$$\mathcal{L}^{ft}(\boldsymbol{\theta}) = \log p_{\theta}(S),$$

- **Fine-tuning + Counterfactual** (FT + CF) The above training loss, however, does not make use of the additional 81,407 counterfactual training sentences for fine-tuning.

$$\mathcal{L}^{cf}(\boldsymbol{\theta}) = \log p_{\theta}(\boldsymbol{s}_2'|\boldsymbol{s}_1),$$

$$\mathcal{L}^{ft+cf}(\boldsymbol{\theta}) = \mathcal{L}^{ft} + \mathcal{L}^{cf},$$

- **Reconstruction + Counterfactual** (RC + CF) One issue with the above training procedures is that models are not explicitly trained to retain as much text of the original outcome x3:5 as possible (i.e., minimum edits).

$$\mathcal{L}^{cf}(\boldsymbol{\theta}) = \log p_{\theta}(\boldsymbol{s}_2'|\boldsymbol{s}_1),$$

$$\mathcal{L}^{rc}(\boldsymbol{\theta}) = \log p_{\theta}(\boldsymbol{s}_{3:5}|S, [s], \boldsymbol{s}_1, [mask]),$$

Experiment

Counterfactual Story Reasoning and Generation (EMNLP 2019)

$$\mathcal{L}^s(\boldsymbol{\theta}) = \log p_\theta(\boldsymbol{s}'_{3:5} | S, [s], \boldsymbol{s}_1, \boldsymbol{s}'_2).$$

Supervised Training (Sup)

Our dataset also provides 16,752 training instances that include human annotated rewritten endings for supervised learning.

Rewritten Sentence Scoring

(1) Does the rewritten ending keep in mind details of the original premise sentence?

(2) Is the plot of the rewritten ending relevant to the plot of the original ending?

(3) Does the rewritten ending respect the changes induced by the counterfactual sentence?

Experiment

Counterfactual Story Reasoning and Generation (EMNLP 2019)

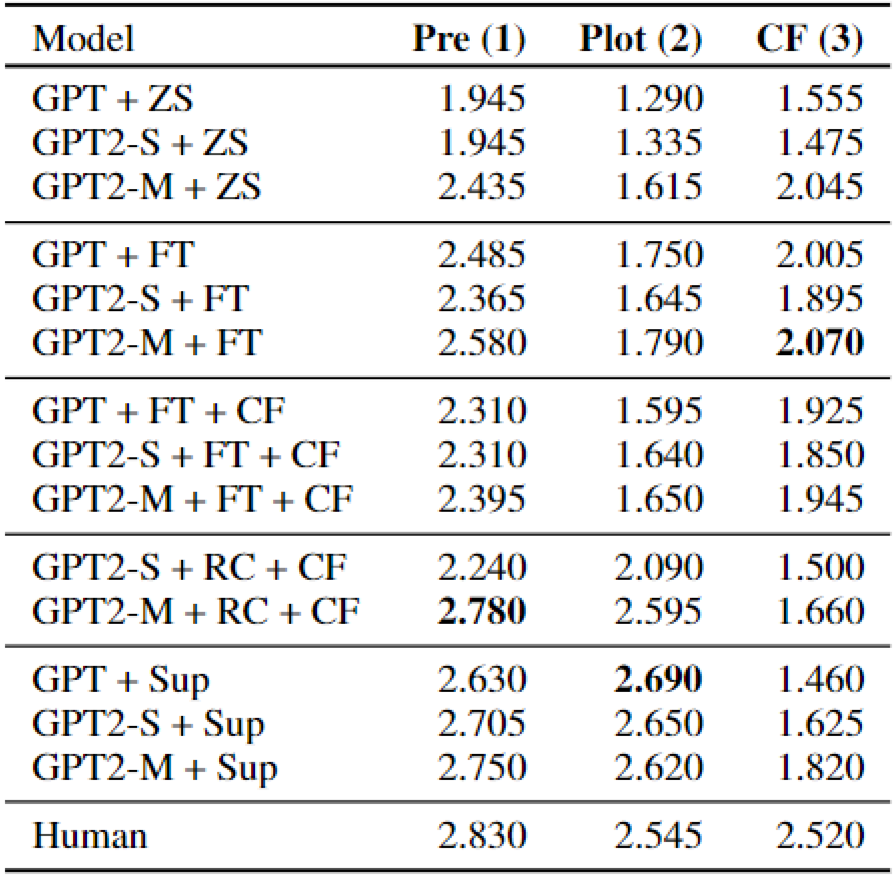
- Model Size and Pretraining Data

We observe that models with more parameters are better at the counterfactual rewriting task than smaller models.

- Domain Adaptation

Fine-tuning on the ROCStories data (FT) is always helpful for increasing performance on counterfactual relevance (CF (3) in Table 4)

Interestingly, however, fine-tuning with the larger set of counterfactuals (CF loss) does not seem to help in rewriting endings that relate to the counterfactuals well.



Model	**Pre (1)**	**Plot (2)**	**CF (3)**
GPT + ZS	1.945	1.290	1.555
GPT2-S + ZS	1.945	1.335	1.475
GPT2-M + ZS	2.435	1.615	2.045
GPT + FT	2.485	1.750	2.005
GPT2-S + FT	2.365	1.645	1.895
GPT2-M + FT	2.580	1.790	**2.070**
GPT + FT + CF	2.310	1.595	1.925
GPT2-S + FT + CF	2.310	1.640	1.850
GPT2-M + FT + CF	2.395	1.650	1.945
GPT2-S + RC + CF	2.240	2.090	1.500
GPT2-M + RC + CF	**2.780**	2.595	1.660
GPT + Sup	2.630	**2.690**	1.460
GPT2-S + Sup	2.705	2.650	1.625
GPT2-M + Sup	2.750	2.620	1.820
Human	2.830	2.545	2.520

Table 4: Likert scale scores for different models. The top performing model for each question is **bolded**.

Experiment

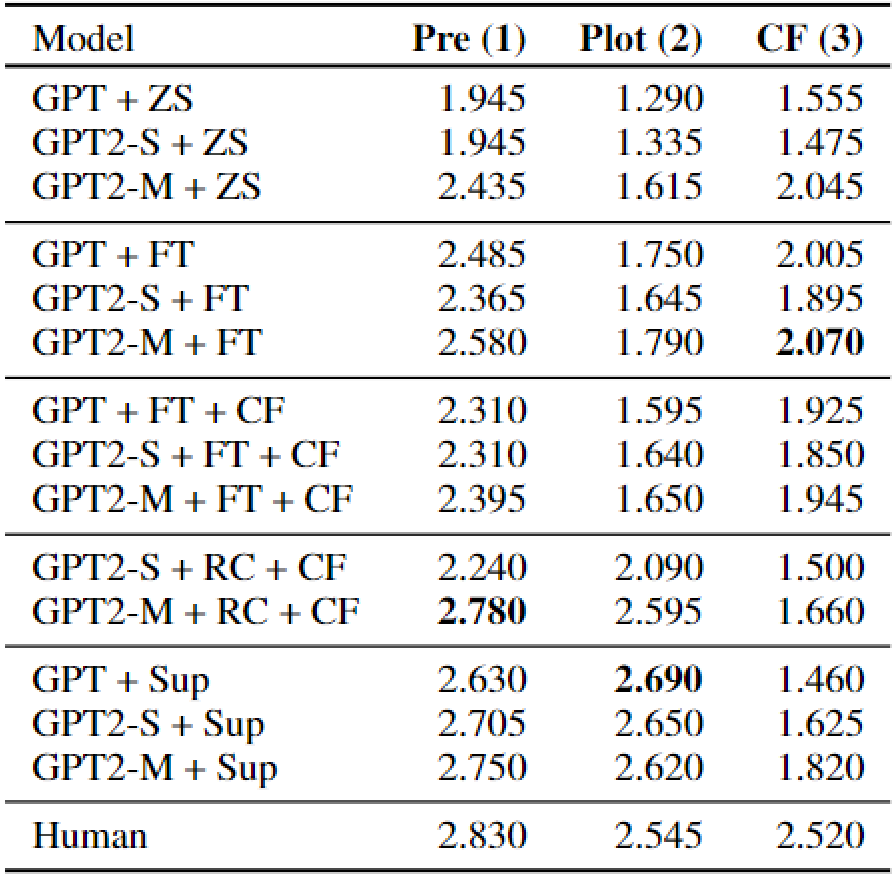
Counterfactual Story Reasoning and Generation (EMNLP 2019)

- Supervised vs. Unsupervised Learning

A surprising observation is that using the dataset of labeled rewritten endings for training does not seem to help the language models learn to rewrite endings better.

The supervised models are generally able to adhere to the plot better than unsupervised methods

Their new endings do not score well on question (3), indicating that they may be copying the original ending or learning to paraphrase the original story ending without acknowledging the counterfactual sentence.

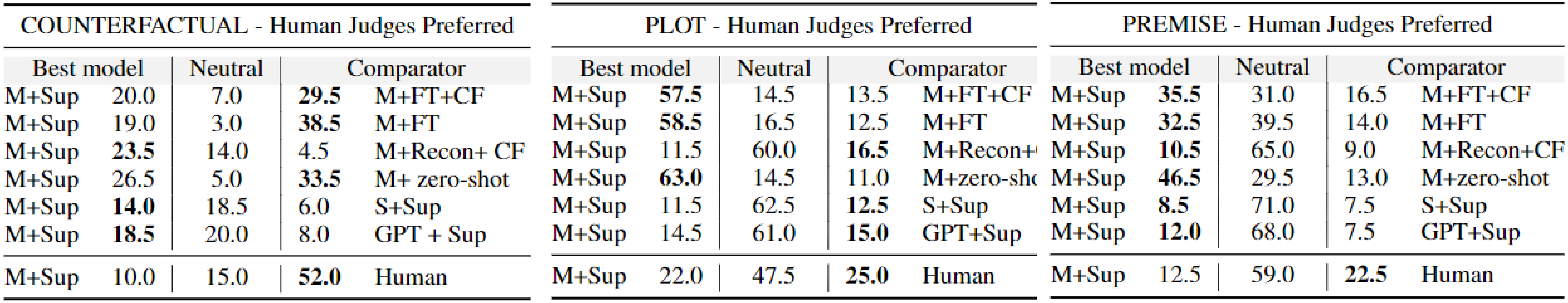


Model	Pre (1)	Plot (2)	CF (3)
GPT + ZS	1.945	1.290	1.555
GPT2-S + ZS	1.945	1.335	1.475
GPT2-M + ZS	2.435	1.615	2.045
GPT + FT	2.485	1.750	2.005
GPT2-S + FT	2.365	1.645	1.895
GPT2-M + FT	2.580	1.790	**2.070**
GPT + FT + CF	2.310	1.595	1.925
GPT2-S + FT + CF	2.310	1.640	1.850
GPT2-M + FT + CF	2.395	1.650	1.945
GPT2-S + RC + CF	2.240	2.090	1.500
GPT2-M + RC + CF	**2.780**	2.595	1.660
GPT + Sup	2.630	**2.690**	1.460
GPT2-S + Sup	2.705	2.650	1.625
GPT2-M + Sup	2.750	2.620	1.820
Human	2.830	2.545	2.520

Table 4: Likert scale scores for different models. The top performing model for each question is **bolded**.

Experiment-Pairwise Model Preference

Counterfactual Story Reasoning and Generation (EMNLP 2019)



COUNTERFACTUAL - Human Judges Preferred				
Best model		Neutral	Comparator	
M+Sup	20.0	7.0	**29.5**	M+FT+CF
M+Sup	19.0	3.0	**38.5**	M+FT
M+Sup	**23.5**	14.0	4.5	M+Recon+ CF
M+Sup	26.5	5.0	**33.5**	M+ zero-shot
M+Sup	**14.0**	18.5	6.0	S+Sup
M+Sup	**18.5**	20.0	8.0	GPT + Sup
M+Sup	10.0	15.0	**52.0**	Human

PLOT - Human Judges Preferred				
Best model		Neutral	Comparator	
M+Sup	**57.5**	14.5	13.5	M+FT+CF
M+Sup	**58.5**	16.5	12.5	M+FT
M+Sup	11.5	60.0	**16.5**	M+Recon+
M+Sup	**63.0**	14.5	11.0	M+zero-shc
M+Sup	11.5	62.5	**12.5**	S+Sup
M+Sup	14.5	61.0	**15.0**	GPT+Sup
M+Sup	22.0	47.5	**25.0**	Human

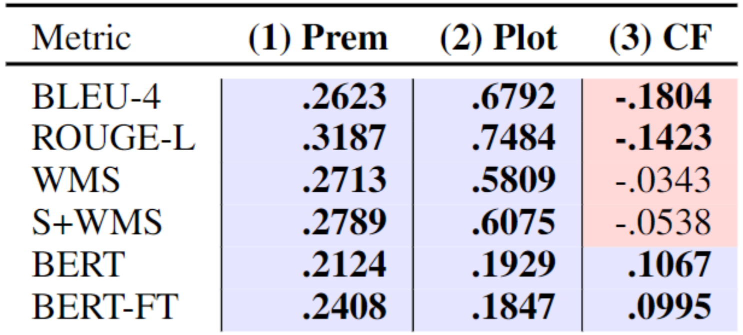
PREMISE - Human Judges Preferred				
Best model		Neutral	Comparator	
M+Sup	**35.5**	31.0	16.5	M+FT+CF
M+Sup	**32.5**	39.5	14.0	M+FT
M+Sup	**10.5**	65.0	9.0	M+Recon+CF
M+Sup	**46.5**	29.5	13.0	M+zero-shot
M+Sup	**8.5**	71.0	7.5	S+Sup
M+Sup	**12.0**	68.0	7.5	GPT+Sup
M+Sup	12.5	59.0	**22.5**	Human

- The best model outperforms the comparison baselines in terms of consistency with premise, while being less consistently better with regards to the other two questions.
- Interestingly, a model that performs better on one of the evaluated dimensions often performs worse for another question, indicating plenty of room for future work in counterfactual reasoning for story rewriting

Experiment-Human Correlation with Metrics

Counterfactual Story Reasoning and Generation (EMNLP 2019)

- we compute the Pearson Correlation between automatic scores and human scores for 800 validation set data points, 300 taken from the gold annotations and 100 generated from each of the 5 GPT2-M variants.
- the automatic metrics are decently correlated with human scores for adherence to the premise sentence and plot
- However, these same metrics correlate negatively with question(3) – adherence to the counterfactual sentence.
- Only the BERTScore metrics appear to positively correlate with human scores for counterfactual understanding, making them usable for evaluating generations across properties related to all three questions.
- However, the correlation is weak, and the results in Table 7 indicate that the BERTScore metrics are difficult to distinguish between models.

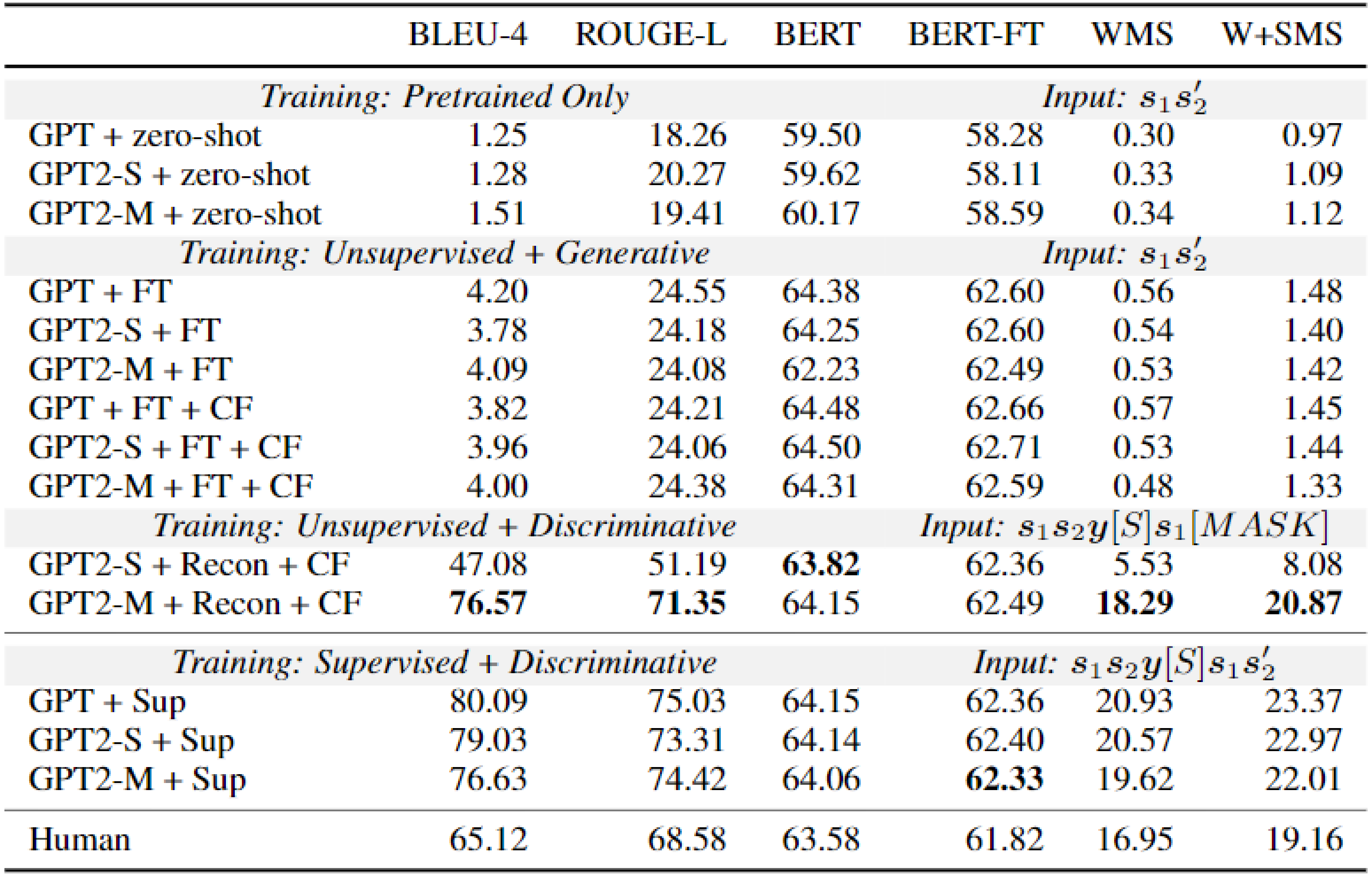


Metric	**(1) Prem**	**(2) Plot**	**(3) CF**
BLEU-4	**.2623**	**.6792**	**-.1804**
ROUGE-L	**.3187**	**.7484**	**-.1423**
WMS	**.2713**	**.5809**	-.0343
S+WMS	**.2789**	**.6075**	-.0538
BERT	**.2124**	**.1929**	**.1067**
BERT-FT	**.2408**	**.1847**	**.0995**

Table 6: Pearson correlation between automatic metrics and human scores. **Bolded** numbers are statistically significant at $p < 0.05$.

Experiment-Human Correlation with Metrics

Counterfactual Story Reasoning and Generation (EMNLP 2019)



	BLEU-4	ROUGE-L	BERT	BERT-FT	WMS	W+SMS
Training: Pretrained Only				*Input: $s_1 s_2'$*		
GPT + zero-shot	1.25	18.26	59.50	58.28	0.30	0.97
GPT2-S + zero-shot	1.28	20.27	59.62	58.11	0.33	1.09
GPT2-M + zero-shot	1.51	19.41	60.17	58.59	0.34	1.12
Training: Unsupervised + Generative				*Input: $s_1 s_2'$*		
GPT + FT	4.20	24.55	64.38	62.60	0.56	1.48
GPT2-S + FT	3.78	24.18	64.25	62.60	0.54	1.40
GPT2-M + FT	4.09	24.08	62.23	62.49	0.53	1.42
GPT + FT + CF	3.82	24.21	64.48	62.66	0.57	1.45
GPT2-S + FT + CF	3.96	24.06	64.50	62.71	0.53	1.44
GPT2-M + FT + CF	4.00	24.38	64.31	62.59	0.48	1.33
Training: Unsupervised + Discriminative				*Input: $s_1 s_2 y [S] s_1 [MASK]$*		
GPT2-S + Recon + CF	47.08	51.19	**63.82**	62.36	5.53	8.08
GPT2-M + Recon + CF	**76.57**	**71.35**	64.15	62.49	**18.29**	**20.87**
Training: Supervised + Discriminative				*Input: $s_1 s_2 y [S] s_1 s_2'$*		
GPT + Sup	80.09	75.03	64.15	62.36	20.93	23.37
GPT2-S + Sup	79.03	73.31	64.14	62.40	20.57	22.97
GPT2-M + Sup	76.63	74.42	64.06	**62.33**	19.62	22.01
Human	65.12	68.58	63.58	61.82	16.95	19.16

Summary

Counterfactual Story Reasoning and Generation (EMNLP 2019)

- We introduced a new task of Counterfactual Story Rewriting that challenges current language understanding and generation systems with counterfactual reasoning.
- Our new dataset, TIMETRAVEL, provides nearly 30k counterfactual revisions to simple commonsense stories together with over 100k counterfactual sentences.
- We establish baseline performances of state-ofthe-art neural language models with over 14 model variants with zero-shot, unsupervised and supervised settings.
- **Strength**: a new task and a new dataset.
- **Weakness**: Some examples of rewriting failures and more in-depth analysis to show what kind of reasoning is required.

Thanks