Designing an Automatic Story Evaluation Metric

Karen Ge [2] Hannah Huddleston [1] William Shabecoff [1]

[1]Computer Science, Stanford [2]SymSys, Stanford

Problem Introduction

A rise in open-domain natural language generation applications like dialog systems and story generators creates the need for automatic story evaluation metrics. Evaluation metrics rate whether a machine-generated story is plausible or implausible. Reference-based evaluation approaches like BLEU and ROUGE do not perform well here because of the open-ended nature of story generation. We aim to create a continuous story metric for use in improving story generation models through backpropogation.

Background

Previous work has shown that running a beam search in GPT-2 that optimizes for baseline metrics such as UNION results in coherent text but not very story-like text [2]. This suggests that these previous metrics are somehow flawed in the way they detect how a story is constructed. This study aims to capture the fundamental high-level features of a story, like plot and emotional states, which can then be used to create a score of "storyness."

Methods

COMET Commonsense Transformers

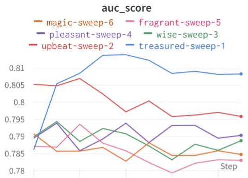
We use **Commonsense Transformers** to automatically generate knowledge graphs about our story at each sentence [1]. For each sentence we parse out the simple knowledge graph, [subject, relation, object] which we then use to query COMET over the relations "Has Prerequisite" and "Causes" in order to generate inferences about the state of the story. A generation looks like:

Source Sentence "It was my final performance in marching band."

Has Prerequisite "rehearse", "musical instrument", "band"
Causes "cheer", "applause", "crowd"

Multilabel emotions model

We were also able to train our submodule which detects whether one of the eight following emotions is entailed by a sentence: {anger, anticipation, disgust, fear, joy, sadness, surprise, trust}. We ran 6 sweeps to tune hyperparameters.



Data Augmentation

The StoryCloze dataset is rather small (4k examples). The ROCStories dataset contains 90k similar commonsense stories but lacks incoherent test cases. We generated new adversarial negative examples with two methods. 1) four sentences of a ROCStory and then swapped the ending sentence with an ending sentence of a different ROCStory. 2) we passed the first four sentences of a ROCStory as context to GPT-2 and had GPT-2 generate the ending sentence.

Reordering Negative Example "Jenny has a drinking problem. Jenny got arrested for public intoxication. Jenny was in jail for three days. Jenny could not write a check for rent from jail. Finally she got the cord free."

GPT-2 Generated Negative Example "David noticed he had put on a lot of weight recently. He examined his habits to try and figure out the reason. He realized he'd been eating too much fast food lately. He stopped going to burger places and started a vegetarian diet. He also began to lose the braids in his head thanks to his behavior."

Model Architecture

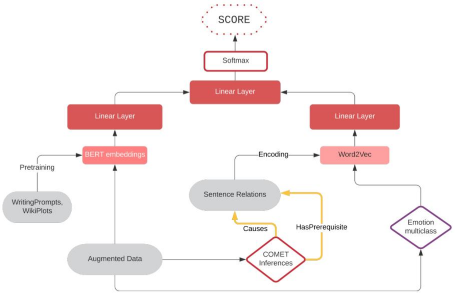


Figure 2. Our full model architecture

Our model relies on three key components: BERT encodings of the original story sentences, outputs from the Causes and HasPrerequisite relations from COMET, and embeddings of emotional states of each sentence.

Experiments and Analysis

We have run experiments on a simplified model architecture with the following results.

BERT version	Data-subset	Accuracy	F1 Score
BERT-base-uncased	StoryCloze	0.82	0.76/0.82
BERT-base-uncased	ROC/GPT-2	0.96	0.96/0.95
RoBERTa-base-pretrained	StoryCloze	0.61	0.46/0.70
RoBERTa-base-pretrained	ROC/GPT-2	0.83	0.84/0.82

Table 1. Results of our model's accuracy

Pretraining on WritingPrompts and WikiPlots seemed to worsen the accuracy of BERT. It is possible that training on these datasets is counter-productive to creating useful BERT embeddings.

Further Work

The use of a pre-trained BERT model with linear classification layer was unsurprisingly quite effective for the task of classifying commonsense stories with our model's 82% accuracy on StoryCloze far surpassing the Mostafazadeh et Al's original StoryCloze paper's 58.5% accuracy with Deep Structured Semantic Model.

For further work we want to run experiments to determine the extent to which our additional architecture will improve performance. This entails encoding our COMET inferences and extracted emotions with word2vec and passing these encodings through a linear layer, then taking the output from this linear layer and the linear layer over BERT through another linear layer before taking the SoftMax of this result and returning the generated probability that the story is coherent as our score.

Additional further work includes broadening the scope of the classifier by using stories where the logical inconsistency exists outside of the last sentence and using datasets like WikiPlots and WritingPrompts which contain text with a different style from the ROCStory and StoryCloze datasets.

References

[1] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi.
COMET: commonsense transformers for automatic knowledge graph construction.
CoRR, abs/1906.05317, 2019.

[2] Jian Guan and Minlie Huang.
UNION: an unreferenced metric for evaluating open-ended story generation.
CoRR, abs/2009.07602, 2020.

github.com/karenge/story-metrics.git

CS224N Winter 2022 Poster Session

{kge, hannhudd, wis23}@stanford.edu