Collaborative Storytelling with Large-scale Neural Language Models

Eric Nichols, Leo Gao, Randy Gomez

MIG '20: Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games

Aim

A novel game of collaborative storytelling where a human player and an artificial intelligence agent construct a story together.

Related Research

Story Generation

Fan et al. construct a corpus of stories and propose a hierarchical story generation model.
Yao et al. approach the task by first generating a plot outline and then filling in the language.
Gupta et al. generate story endings by incorporating keywords and context into a sequence-to-sequence model.
Luo et al. incorporate sentiment analysis into story ending generation.
See et al. conduct an in-depth analysis of the storytelling capabilities of large-scale neural language models.

Related Research

Interactive Language Generation

AI Dungeon (https://play.aidungeon.com/) is a text adventure game.

=> Their study is similar to this game, but their task is not restricted on the first-person adventure, and they ranked model output.

[Cho and May 2020] build an improvisational theater chatbot.

=> Their task is not improv, and they do not use the model's output as-is.

Related Research

Language Models

A language model is a mathematical model that assigns likelihoods to sequences of words where sequences that are more likely in a target language are given higher scores.

$$P(x) = \prod_{i=0}^{n} P(x_i \mid x_{<i})$$

They used GPT-2.

Approach

Generator model: a large-scale neural language model tuned on storytelling data to generate story continuation candidates.

+

Ranking model: trained on human storyteller preferences to score them and select the highest quality continuation.

Approach

Generation

They used the publicly-available pretrained 774M parameter GPT-2-large model tuned on their WritingPrompts dataset.

The main solutions for the output may be ill-formed or lacking in logical coherence are the following:

using larger models

using different sampling methods

using various methods of traversing the search space of possible sentences

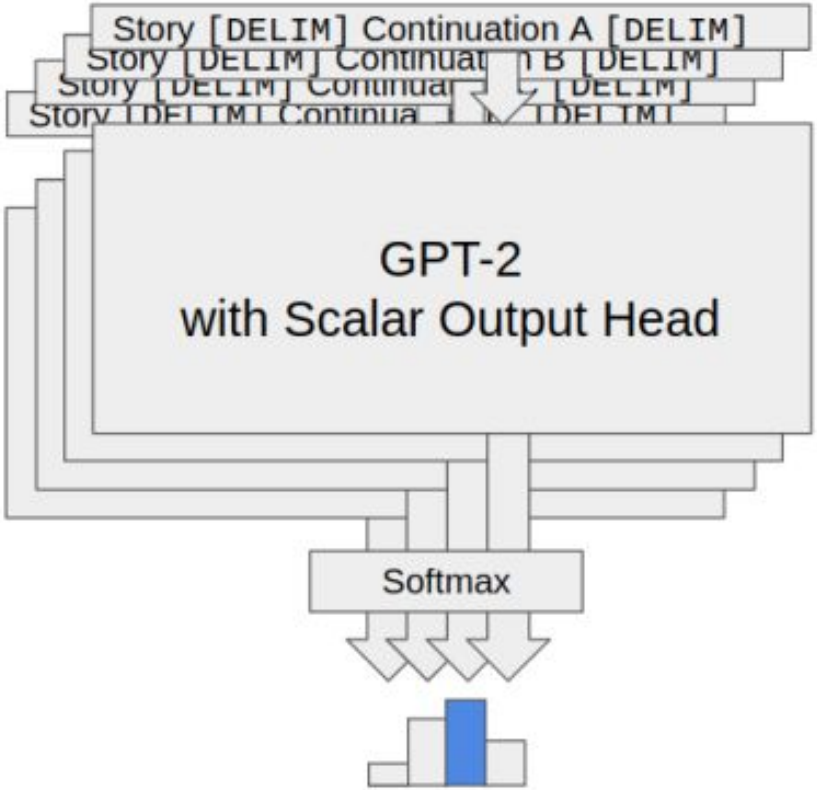
Approach

Sampling

They used nucleus sampling with $p = 0.9$, which takes the smallest set of the most probable tokens whose cumulative probability exceeds the threshold p.

Approach

Ranking



Approach

Datasets

Writing Prompts Dataset:

- collected from r/WritingPrompts subreddit
- used for training the Generator model

Collaborative Storytelling Dataset:

- collected using Mechanical Turk
- used for training the Ranker model

Approach

Story Continuation Sampling and Ranking

Generator model => Cleanliness heuristics => Ranker model

rejected objects:

less than 60% alphabetic characters

unbalanced quotations, select profanity

words like “chapter” that are not typically part of the story

Approach

Training

The Generator model is trained with a maximum likelihood estimation loss function using Adafactor with a learning rate of 5e-5 on a weighted mixture of the WritingPrompts and BookCorpus.

The Ranking model is trained using Adam with a maximum learning rate of 1e-5 on the BookCorpus and Writing Prompts dataset and Collaborative Storytelling dataset. The entire model is trained.

Evaluation

Measuring story continuation ranking accuracy and story continuation acceptability

Adapting the Acute-eval chatbot evaluation metric to collaborative storytelling evaluation

Evaluated systems:

untuned (pretraind GPT2)

tuned (GPT-2 tuned on storytelling data)

tuned+ranker (GPT-2 tuned on storytelling data with a single story continuation selected by the Ranker model)

Evaluation

Story Continuation Prediction Accuracy

System	Dataset	Accuracy	
tuned+ranked	validation	22.9%	(229 / 1000)
tuned+ranked	test	23.3%	(233 / 1000)
random baseline	-	10.0%	

Table 2: Accuracy of the tuned+ranked model at predicting the story continuation that was selected by the Mechanical Turker who constructed the story. Note that a random baseline would pick the correct continuation 1 out of 10 times.

Evaluation

Story Continuation Accetability

System	Acceptability	
untuned	33.9%	(305 / 900)
tuned	39.8%	(358 / 900)
tuned+ranker	**62%**	(62 / 100)

Table 3: Mean acceptability of story continuations in the test set. To evaluate untuned **and** tuned**, acceptability is calculated over all 9 continuations from each system, while** tuned+ranked **uses the Ranker to consider only the best one.**

Evaluation

Human Annotator Story Preferences

Characteristic	Question
Engagingness	Who would you prefer to collaborate with for a long story?
Interestingness	If you had to say one of these storytellers is interesting and one is boring, who would you say is more interesting?
Humanness	Which storyteller sounds more human?
Story Preference	Which of these stories do you like better?

Table 4: Questions asked to human evaluators of collaborative storytelling systems. Characteristics and questions are based on the *PersonaChat* evaluation metric of [Li et al. 2019], with minor changes to wording to reflect the task's storytelling nature.

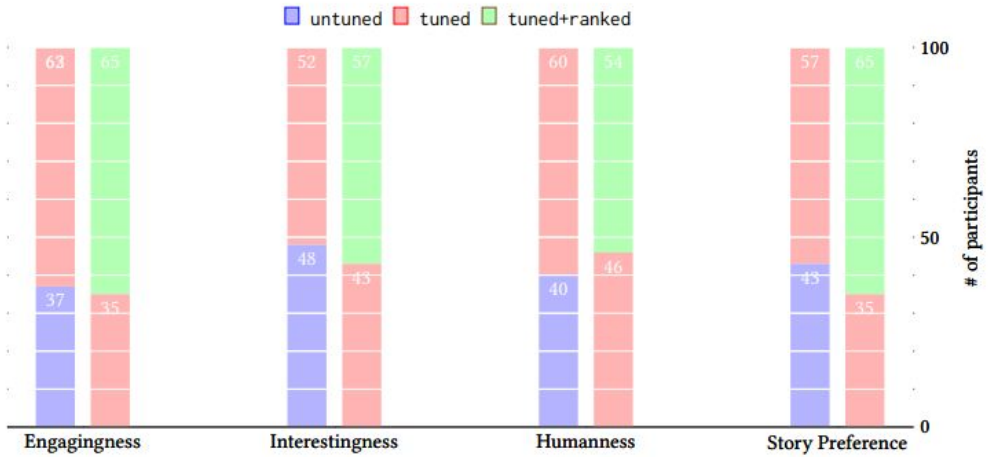


Figure 5: Human evaluation of collaborative storytelling systems. We compare the pairs (`untuned`, `tuned`) and (`tuned`, `tuned+ranking`). Each bar graph shows a comparison of two different systems generating stories through self chat. A larger portion of the bar indicates that system was preferred by evaluators.

Discussion

Advantages:

The system can produce well-formed story contributions.

Limitations:

The system makes incoherent output.

Requiring human demands can be difficult.

Self-chat evaluation is not suitable for the system.

Conclusion

The collaborative storytelling system can select story continuations that are aligned with human preference.

The results show that both the sampling and ranking approaches contribute to generating high-quality story continuations.

Thank you for your attention!