Generating one-liners and six-word stories using Natural Language Processing

Sriharsha Srungaram
srisrung@iu.edu

Rithwik Sarma
rpanyam@iu.edu

Abstract

Text generation has the potential to put AI's creative abilities to the test. In this project, we gathered data from Reddit's r/SixWordStories and r/OneLiners subreddits and separately generated six-word and one-line stories and using a character LSTM and a fine-tuned GPT-2 model. When we compared the stories generated by GPT-2 to those generated by the LSTM in terms of readability, cohesiveness, humanness, and overall quality, we discovered that the GPT-2 stories scored higher in all categories than the LSTM stories, with more consistent plot development. However, the GPT-2 stories, on the other hand, varied in length more often than LSTM for six-word stories. While the LSTM model managed to produce readable and syntactically correct sentences, they are often not sensible. The GPT-2 however, could capture the wordplay and sentence structures found in human-written stories from the two subreddits.

1 Introduction

Over the past few decades, the application of deep learning models in the field of Natural Language Processing has been constantly increasing. From text classification to language modelling to question answering, the applications of deep learning in NLP are vast. In this project, we try to explore the text generation aspect and try to generate both six word stories and one liners by training a character level LSTM model and later fine tuning the GPT-2 model on both datasets. Ernest Hemingway, the famous American novelist first rose to fame by his famous six word flash fiction story:

"For sale: baby shoes, never worn." [8]

The popular r/SixWordStories subreddit features thousands of such six word sentences which reflect a short story. These stories tend to be dark and funny. Similary the subreddit r/OneLiners contains thousands of texts which depict a story expect for the six word limit.
We transform each text to contain a start and an end token, perform some pre-processing and later feed it to both the models.
Like every text generation task, our challenges included generating g grammatically correct text that is coherent and human-understandable. However, on top of that we dealt with additional level of complexity in the form of generating text which depicts a story which is funny/sad/dark but also has a word limit in one case and limiting to a sentence in the other.

2 Related Work

LSTMs, and other Seq2Seq models were used by Keppler and Chen [1] to generate the second sentence of a two-sentence horror story when provided with an input sentence.
Another form which imposes strict form requirements upon the text generated, is poetry generation which has been explored [6]. Transformer models, which are based solely on attention mechanisms

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

[4] have previously been used for language modeling tasks. One model that uses generative pre-training followed by discriminative fine-tuning and also can easily adapt to different tasks is OpenAI's GPT model [3]. Its successor, GPT-2, generally follows the architecture of GPT, but also moves layer normalization to the input of each sub-block, increases the vocab size, and increases the context size, resulting in a model with more parameters [2].
Another paper by Ben Swanson, Elif Yamangil, and Eugene Charniak explores Natural Language Generation with Vocabulary Constraints. They study two constraints concerning the words that are allowed in a sentence. The first sets a fixed vocabulary such that only sentences where all words are in-vocab are allowed. The second demands not only that all words are in-vocab, but also requires the inclusion of a specific word somewhere in the sentence.

3 Dataset

The datasets have been generated from the following two subreddit's:

r/OneLiners [9]

r/SixWordStories [10]

We initially explored Reddit's own API in order to scrape the data for the project. Although it is easy to use it only allows the users to pull a limited number of submissions (hot, top, new submissions) from a subreddit. Hence, to download all the submissions to a subreddit, we needed to look at other options. We later encountered the Pushshift API [5] which fit the bill perfectly.
Pushshift is a service that ingests new comments and submissions from Reddit, stores them in a database, and makes them available to be queried via an API endpoint. We used the PRAW [7] (the Python Reddit API Wrapper) to call the Pushshift API to query data from both the subreddits.
For both the subreddit's, the text we required were the titles of the submissions and hence used the $object\_type = `title'$ to collect the titles of the submissions. After scraping the entire subreddit, we ended up with 24689 One Line stories and 70467 six word stories.

A couple of examples from the six word story subreddit:

And here we go again, Monday.
Conflict - was more familiar than love.

A couple of examples from the one liners subreddit:

Wouldn't a library filled with all non-fiction books be called a truth-brary?
I hate it when people mispel words

3.1 LSTM Pre-processing

For the LSTM model, we used one-hot encoding meaning that each character is converted into an integer (via our created dictionary) and then converted into a column vector where only it's corresponding integer index will have the value of 1 and the rest of the vector will be filled with 0's.Later, to train on this data, we also created mini-batches. The batches here are multiple sequences of some desired number of sequence steps.

3.2 GPT-2 Pre-processing

A start (<S>) and an end (<E>) token were appended at the start and the end of each sentence respectively.

4 Methods

Modelling was done in the same way for both the datasets. No explicit constraint was set on the word limit for generating six-word stories. The idea was to see how well these models inferred this constraint from the dataset.

4.1 LSTM

A baseline character level LSTM was implemented that takes in an input sequence of characters and returns a probability distribution over all the characters in the vocabulary. This essentially gives the probability of each character that could follow the input sequence. Inputs were padded to the same length so they could be batched for training. A total of 4 LSTM layers were used each with 512 hidden units.

4.2 GPT-2

GPT2 is a transformer model that is similar to the original GPT model created by OPENAI.
We fine-tuned the pretrained GPT-2 (medium) model. We used a learning rate scheduler with exponential decay after every 500 steps with a decay rate of 0.7, and an initial learning rate of 0.001 We used a pre-trained tokenizer from gpt2-medium, and special BOS, EOS and PAD tokens were provided with the number of tokens limited to 60
The model was set to train for 25 epochs since training it for very high epochs resulted in the model reproducing exact sentences from the training dataset. However, due to the early stopping criterion, the model only trained for 6 epochs.
Beginning of sentence (BOS) token was provided for the model to produce output examples. It then generated words until it produced the end of sentence (EOS) token.

5 Results and Discussion

5.1 LSTM output

The LSTM model's loss came down to 1.185 after 100 epochs for the one-liners dataset.
Similarly, the loss was 1.225 for six-word story dataset by the end of the training.

5.1.1 One Liners

Some hand-picked outputs for one-liners include:

i am not sexing my feet for any charge.
i wouldnt say i was too much, im the best of them.
a paraceite walks into a bra.
I have a sex issues, its to go on my body.
my friend said, "I can die, but I'm getting mirrored."

As we can see our character LSTM model does a good job of learning the English language with only a few input training sentences. It does a good job at starting and ending an sentence and also tries to model that sentence in the format of a story. Nevertheless, it makes some spelling mistakes and fails to produce meaningful results. Fine tuning a bigger model like GPT-2 should get rid of these issues and also help produce better outputs.

5.1.2 Six-Word Stories

Some hand-picked outputs for six-word stories include:

trump isn't a child she said
six words that's what i suck
today i woke up the storm
i'm not alone in my soul
she said i don't know me
im not starting to be alone

Almost all the stories generated by the character LSTM conform to the six word limit. Moreover, the stories are always syntactically correct and make semantic sense most of the times. While these outputs are not as good as the stories found in the r/sixwordstories subreddit, most of them capture the tone and structure of the sentences common on the subreddit.

Note that the outputs for this model do not have any punctuation or capitalization because of the pre-processing steps applied to the data before training.

5.2 GPT-2 output

The LSTM model's loss came down to 1.185 after 100 epochs for the one-liners dataset.
Similarly, the loss was 1.225 for six-word story dataset.

5.2.1 One Liners

Some hand-picked outputs for one-liners include:

I've always been in the dark, watching my dark side.
What do you call an impotence? A strong woman
Id like to make a joke about how bad of an electrician Id like to be, but that would have to be a bit powery.
I'm writing a joke about construction... Its coming soon
I am the worst kind of liar....I never lie!

The fine tuned GPT-2 model outputs one liners which are syntactically and grammatically correct. Moreover, it also comes up with One Liners which are both dark and funny. It also does really well to capture the tone of the sentence as we can see that some of the hand picked examples above have a sarcastic and dark tone to them.

5.2.2 Six-Word Stories

Some hand-picked outputs for six-word stories include:

She fell from grace onto stone.
I need a gun, not money.
The world is not worth living.
His name still haunts me everyday
He loved her; she loved others.

The GPT-2 generates Six-word stories that are always syntactically correct and make semantic sense. The results do a great job to model the tone of a story as well.

However, some of the results did not conform to the six word limit. They deviate to 7 words most of the time, and rarely to 5 words.

This is just a six word story.
No. The gun jammed, the family panicked.
Luggage, never unloaded, never opened.

5.3 Quantitative Evaluation

Looking at the quality of the output generated by the models, specially from GPT-2, we were concerned that the model might be generating sentences that are replicas of the training data. We had already taken care to prevent this from happening by keeping the number of epochs very low.

In order to check this, we calculated what fraction of generated output is copied from the training data, for GPT-2 trained on both the datasets (one-liners and six-word stories), and less than 1 percent of the generated lines were found in the training dataset when 200 lines each were randomly sampled from generated output for both the datasets.
To make the comparison more fair, both source data and generated data were processed to remove all special characters and letters were lower-cased in order to prevent cases where if GPT-2 copies the sentence from training data and adds an exclamation point at the end to be counted as unique lines.
Further, for six-word story generation, since the models were not provided with an explicit constraint on the number of words, we attempted to measure the fraction of outputs that conform to the 6 word constraint.
Of the 200 stories generated by GPT-2, 131 (65.5%) had exactly 6 words, 42 (21%) had 7 words, 26 (13%) had 5 words. Only one story deviated by more than one word from the six word limit. While for LSTM, the numbers are even better. Of 200 stories, 186 (93%) had 6 words. Only 8 (4%) stories had 7 words, and 2 (1%) stories had 5 words. There were 4 stories with a deviation of more than one word.

5.4 Qualitative Evaluation

For qualitative Evaluation, we sent a survey to gain insights from 12 human evaluators. The survey contained 18 stories/one-liners in total out of which 3 were human generated one liners, 3 were human generated six-word stories, 6 were generated by LSTM (3 each for six-word stories and one liners) and 6 were generated by GPT-2 (3 each for six-word stories and one liners). We chose the LSTM and GPT-2 outputs for the survey at random but made sure to only select the outputs which were of six words only in regards to six word stories so as to not make it obvious which were computer-generated. We also did not select the outputs which matched the training set examples.

For each of the 18 sentences in the survey, we asked the 11 evaluators to evaluate the sentences on scales from 1 to 10

- Readability: How readable a given sentence is in terms of being grammatically correct? 1 being least readable to 10 being easily readable
- Cohesion: Does the sentence talk about the same topic throughout? 10 corresponds to very cohesive.
- Humanness: How likely it was that the story was written by a human. 10 corresponds to being most likely
- Overall Quality: How much would the evaluator like to rate that sentence? 10 being the best possible score they could give

The results for one liners were as follows:

Model	Avg. Readability	Avg. Cohesiveness	Avg. Humanness	Avg. Overall Quality
Human	9.03	8.52	9.26	8.11
LSTM	3.86	3.41	3.11	3.54
GPT-2	8.32	7.76	7.93	7.87

The results for six-word stories were as follows:

Model	Avg. Readability	Avg. Cohesiveness	Avg. Humanness	Avg. Overall Quality
Human	9.12	8.88	8.86	8.89
LSTM	4.96	4.11	3.81	3.87
GPT-2	8.52	8.17	7.63	7.34

As we can see for both, the One Liners and Six-Word stories, the human generated outputs scored the highest followed by GPT-2 outputs followed by the LSTM outputs. The scores for the GPT-2

model are good indicating that the model does a good job of mimicking the style of the training sequences. However, our human evaluators were still able to differentiate between the computer and human generated outputs as for both the cases, the average humanness was 8.86 and 9.26 which is quite high. The character LSTM model scores below 4 in the average overall quality indicating that the output was of poor standard. This was expected as we saw that the output from the LSTM had spelling mistakes, grammatical errors and lacked semantic sense sometimes.

6 Conclusion

Text generation is a difficult work in and of itself, but adding a six-word constraint makes it even more difficult. It is possible to make stories that conform to the word constraint and are human-readable by using an LSTM to generate six-word stories. The punctuation in these stories for both cases, however, is sometimes uneven, and while their meaning can be construed as vaguely metaphorical, they rarely imply any events or acts in the way that a storey would.
GPT-2-generated stories, on the other hand, are more consistently readable, cohesive, and human-like than LSTM-generated stories. GPT-2's outputs, on the other hand, have greater inconsistencies in tale length, and many of them are direct copies of the instances in the training dataset.
Overall, while neither the LSTM nor the GPT-2 are capable of capturing the full breadth of storytelling that human authors can in only six words, both are capable of producing readable and coherent text; the GPT-2, in particular, performed exceptionally well on all metrics when human-evaluated, producing stories with true plot progression or implication.

7 Future Work

The evaluation of generated stories is an area where more research is needed. The creation of an automated quantitative metric, such as one that calculates the percentage of stories with exactly six words or assigns a cohesion or readability score to a story automatically, could allow for a standardised, more objective way to evaluate model outputs.

References

[1] Adam Keppler, Jennie Chen. HorrifAI: Using AI to Generate Two-Sentence Horror.

[2] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. Language Models are Unsupervised Multitask Learners.

[3] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. 2018.

[4] Ashish Vaswani et al. Attention Is All You Need. 2017. 31st Conference on Neural Information Processing Systems.

[5] Jason Baumgartner. Reddit statistics - pushshift.io. [Online; accessed 29-October-2020].

[6] Marjan Ghazvininejad, Xing Shi, Yejin Choi, Kevin Knight. Generating Topical Poetry. 2016. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.

[7] PRAW: The Python Reddit API Wrapper.

[8] Wikipedia. "For sale: baby shoes, never worn."

[9] Reddit: OneLiners.

[10] Reddit: SixWordStories.