Generating Haikus with Natural Language Processing

Stanford CS224N Custom Project

Emily Bunnapradist
Department of Mathematics
Stanford University
embunna@stanford.edu

James Zheng
Department of Computer Science
Stanford University
jameszhe@stanford.edu

Raymond Suo
Department of Computer Science
Stanford University
raysuo@stanford.edu

Abstract

Haikus are a traditional Japanese poetry form, characterized by three short lines, and often about nature[1]. There have been several attempts, including those powered by the GPT-3 model, to generate a Haiku given a topic word; however, these poems, while comprehensible and logical, are often not artistic, and can be easily distinguished from the poetry of an actual human poet.[2] We aim to create a transformer language model that takes in a list of topic words and generates a comprehensible, relevant, and artistic three-lined haiku utilizing a finetuned version of the GPT-3 Curie[3] model. Our corpus contains approximately 400 English haikus sourced on the Internet. In addition, we utilize data augmentation techniques and pretraining methods in order to assess the most artistic outputs.

1 Key Information to include

- Mentor: Ethan Chi

2 Introduction

Haikus are a traditional form of Japanese literature that emerged around the 17th century as an alternative to the complex existing standard for poetry during that time period[4]. Characteristically, haikus are most well-known for their unique syllabic structure with three lines and seventeen syllables. Haikus stand out amongst other poetry forms due to its flexibility in terms of topic, rhyme, and concise syntax.

Sample haiku:

The Old Pond

An old silent pond
A frog jumps into the pond –
Splash! Silence again.

Matsuo Bashō (1644-1694)

Although haikus are traditionally Japanese, there has been a resurgence of haikus in English-speaking countries beginning in the late 19th century. With an upward trend in the usage of haikus, the art form has continued to become more prevalent throughout the past century. With our language model, our hope is that machine haiku generation will be able to introduce more English speakers to the art of haikus and preserve the art form. Through providing a free and accessible way to generate haikus, we aim to encourage the writing of and appreciation of not only haikus but also poetry as a whole.

In this project, our goal is to implement complex NLP models that generate English haikus utilizing the GPT-3 framework from OpenAI[3]. In particular, we hope to analyze the difference in the perplexity of outcomes given the presence of pretraining on general poetry and the usage of data augmentation. Specifically, we worked to build off of a dataset we developed of approximately 400 haikus, and used various data augmentation techniques to expand the training data available. Additionally, we utilized a dataset with approximately 8000 generic poems found on Poetry Foundation[5] for targeted pretraining.

3 Related Work

3.1 Poetry Generation Using Deep Neural Networks

Previous research has been done on Haiku generation using deep neural networks (Wu et al. 2017)[6]. As described in their paper, Wu et al. trained various types of neural networks on Japanese haikus found on the web and those created by human users of Rinna, an emotional chatbot that simulates a female high school senior. The four types of models trained were a vanilla recurrent neural network (RNN), an RNN using LSTMs and GRUs, a recurrent convolutional neural network (RCNN), and sequence generative adversarial networks (SeqGAN).

Other work has been done on generating classical Chinese poems to improve novelty and thematic consistency (Li et al. 2018)[7]. In their paper, they used a conditional variational autoencoder with adversarial training.

3.2 Utilizing GPT-3

A transformer-based model coined *Haikoo*[8] has been shown to be capable of generating haikus (Miceli 2021). *Haikoo* is based on GPT-2, a model released in February 2019 that preceded GPT-3. On top of finetuning GPT-2, *Haikoo* also consists of an additional Plug and Play Language Model (PPLM).

As we have seen, there has been some research done on training models to generate various forms of poetry. For our project, we explored whether deep learning models can also extend to English haikus. We also take advantage of GPT-3, an existing language model that has already been pretrained on 45 TB of text data. To the best of our knowledge, research into poetry generation using deep learning methods have not made use of data augmentation techniques. Thus, we decided to use four EDA (easy data augmentation) techniques[9]: synonym replacement, random insertion, random swap, and random deletion (Wei & Zhou 2019).

4 Approach

4.1 Model Architecture

Our models fine-tune the Curie model in GPT-3, the Generative Pre-trained Transformer 3 authored by OpenAI[3]. GPT-3 utilizes a generative language model pre-trained on approximately 45 TB of text data from online and text sources. The Curie model in particular specializes in analyzing complex text, and has capabilities in tasks such as sentiment classification and summarization. Since it tends to lend its skills well to producing text, we decided to utilize this engine for our poetry generation.

We wanted to compare various inputs and data collection methods to see how we could produce more realistic and artistic haikus. In particular, we utilized the following two methods: data augmentation and pretraining on regular poetry.

4.1.1 Method 1: Finetuning

Our first approach to improve the output of GPT-3 is to finetune it with additional training data, which will make the model more specialized for haiku generation[10]. To do this, we compiled a collection of 383 English Haikus that we, along with the opinions of two additional students in creative writing, found to be artistic. We then assigned each haiku a topic word that we believed to be an accurate representation of the poem.

4.1.2 Method 2: Data Augmentation

Given the relative scarcity of haiku poetry available in English, we aimed to increase our dataset to see how the artistry and relevancy would improve given a larger array of poetry to finetune on.

We employed four data augmentation methods: inserting a word by contextual word embeddings utilizing BERT, swapping two random words in a line, deleting random word(s) in a sentence, substituting a word utilizing WordNet's synonym database[11].

Augmentation Method	Original Poem	Augmented Poem
Insert	A spring day– A long line of footprints On the sandy beach	A hot spring day afternoon – A long life drawn of animal footprints On the black sandy beach
Swap	The water flowing The rock trips the falling stream The trees grasp the edge	The flowing water rock The the trips falling trees stream The grasp the edge
Delete	Moon shadows rise Fall as cries draw near Fear not, yet wind knows otherwise	Moon shadows rise fear not yet wind knows otherwise
Substitute	Moon shadows rise Fall as cries draw near Fear not, yet wind knows otherwise	Moonlight shadows rise up Fall as outcry draw near Veneration not, yet wind knows differently

Figure 2: Sample data comparing original poems from our dataset to augmented poems.

4.1.3 Method 3: Pretraining on Regular Poetry

In order to increase the artistry of our generated poetry, our team decided to pretrain our GPT-3 model on regular poetry. Given the relative scarcity of existing haiku datasets, we hypothesized that the artfulness of generic poems would allow our model to be better trained towards producing better text. Then, by finetuning our model on haikus, we preserve the structure of short and concise poetry.

5 Experiments

5.1 Data

To train the CURIE model, we constructed two different training datasets. The first is a dataset of 8000 generic poems found on found on poetryfoundation.org's database. These poems were used to pretrain the model described in the methodology above. The second dataset included approximately 383 haikus as training data that we gathered from various sources, including thehaiku-foundation.org, poets.org, and poets.com[12],[13],[14],[15],[16]. After augmentation, the dataset had 1915 haikus. This dataset of haikus was used to finetune the model after pretraining on generic poems.

Sample generic poem from the poetryfoundation.org dataset:

Abandoned Homestead in Watauga

"All that once was is this,
shattered glass, a rot
of tin and wood, the hum
of limp-legged wasps that ascend
like mote swirls in the heatlight.
Out front a cherry tree
buckles in fruit, harvested
by yellow jackets and starlings,
the wind, the rain, and the sun."

Ron Rash (1953-)

5.2 Evaluation method

5.2.1 Quantitative Evaluation: Perplexity

In order to evaluate our model's performance on generating poetry, we measured the perplexity, a common metric used for generative language models, of our haiku outcomes to quantitatively measure the relevance of our computer-generated haikus in comparison to human-generated haikus. Perplexity is represented by the following formula,

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, ..., w_N)}}$$

where each w_i represents a word, and N is the number of words[17].

To test our model using perplexity, we created a testing data set, separate from the training data, that contained 100 haikus written by real humans and their associated topic words. The five topic words we decided to test were tree, spring, water, winter, and moon, and we gathered 20 haikus for each word. We then calculated perplexity using these haikus for the raw GPT-3 model, the finetuned model, the finetuned model with pretraining, the finetuned model with data augmentation, and the finetuned model with pretraining and data augmentation.

Sample dictionary with words mapped to probabilities:

```
{'the': 0.1;
 'two': 0.0125;
   .
   .
   .
   'autumn': 0.004166667;
   'breeze': 0.008333333;
   'trees': 0.0458333333
}
```

Then, utilizing these dictionaries, we have a list of probabilities that we will use to calculate the perplexity of each generated poem. Specifically, we chose to create a target dictionary that is cased on the prompt rather than all haikus in order to better measure the relevancy of a topic-specific poem. We are hoping for a minimized perplexity, as this correlates to a higher probability of haikus in the test set being generated by the model.

5.2.2 Qualitative Evaluation: Artistry, Comprehensibility, Relevance

For the subjective evaluation, we decided to test and compare haikus generated from three sources: (1) the raw GPT-3 model (2) the GPT-3 model after pretraining on generic poetry and finetuning on

augmented haikus (3) haikus written by real-life poets. We chose 87 topic words, mostly relating to nature ("tree," "sky," "wind" etc), and received a haiku from each of the three sources for each of the topic words. Each of these poems were then evaluated using the methodology described above.

In order to evaluate artistry, comprehensibility, and relevance for our generated poems, we also utilized the aforementioned database of 87 generated poems, as well as created a separate database of 87 human-written haikus. We asked two volunteers with a background in English literature to score each poem on a scale of 0-5 in each category.

Examples of Haikus Used for Quantitative and Evaluative Perplexity:

Model	Prompt: 'Water'	Prompt: 'Moon'
Raw GPT-3 Model	The water is cool It soothes and refreshes me I am at peace here	The moon shines so brightly in the sky it's a beautiful sight
Finetuned Model	Plants wither down Water uses up Loving is wasting	The moon above the towers the only guest a June moon
Pretrained Model	By the spring When the sweet water falls I remember my dream	Having viewed the moon I saw farewell to this world With heart bidding
Augmented Model	The water flowing The rock shifting gently As the falls drop	A bubble blown on by a breath full moon
Pretrained + Augmented Model	The water flowing The rock trip out the fall stream The trees grasp	A brilliant full moon! On the surface of the water Shadows

5.3 Experimental details

During the pretraining and finetuning process, we used 4 epochs and a 0.1 learning rate. The pretraining on generic poetry took around 3 hours to complete. The finetuning of the model using the augmented haikus took around 1 hour to complete.

5.4 Results

We calculated perplexities for the raw GPT-3 model and the fine-tuned model on the test set consisting of 87 human-authored haikus:

Model	'Trees'	'Moon'	'Spring'	'Water'	'Winter'
GPT-3 Raw Model	4.117028	4.525016	29.995968	10.364406	33.670600
Finetuned Model	4.669996	6.389427	11.717591	1.686709	2.159090
Finetuned Model w/ Pretraining	3.380148	3.334415	7.680905	5.13376	9.358612
Finetuned Model w/ Augmentation	9.831432	10.047982	22.22556	1.6694699	6.563061
Finetuned Model w/ Pretraining and Augmentation	11.57064	5.483812	5.914111	2.372567	1.0

Therefore, we can calculate the average perplexities to obtain the following table:

Model	Perplexity
GPT-3 Raw Model	16.5346936
Finetuned Model	5.3245626
Finetuned Model w/ Pretraining	5.777568
Finetuned Model w/ Augmentation	10.067501
Finetuned Model w/ Pretraining and Augmentation	5.268226

From this, we see that the perplexity of the all models generated by our team is lower than the raw model without modification. This means that the probability of the test set poems being generated

using the new model is greater than the raw model, suggesting that the fine-tuned model is better generating more artistic, and real (written by actual poets) haikus.

When running the subjective tests on the four models, after averaging the scores from the two volunteers for all 87 generated poems each from the GPT-3 model and finetuned model, as well as 87 human-generated poems, we received the following scores:

Model	Artistry	Comprehensibility	Relevance
GPT-3 Raw Model	2.2	4.5	4.1
Finetuned Model w/ Pretraining and Augmentation	3.9	3.8	3.7
Haikus from real-life poets	4.5	4.3	3.9

6 Analysis

As shown from the perplexity results, all four of of the models that we pretrained/finetuned had significant improvements when compared to the raw GPT-3 model. Because the models were finetuned on haikus specifically, they are then more likely to generate something that mirrors a real-life haiku.

However, the augmentation on the Haikus in the training set actually led to slight increases in perplexity. This can most likely be explained by the affects and augmentation; because we are inserting, swapping, deleting, and substituting new words into existing haiku's, the model is trained on data that may have words and sentence structures that are different from a real-life haiku, thus leading to results that have lower perplexity.

When first pretraining the model on generic poetry before finetuning on Haikus, we also did not see much improvements in perplexity. This can first be explained by the large amount of Haikus that we finetuned the GPT-3 model with. Because there is a large amount of haiku data, the generic poetry was not really needed to see improvements. Additionally, many generic poems have words, phrases, and structures that are not really evident in Haikus, such as rhyming or long lines. In the case that we were not able to finetune on as many Haikus as we did, then the pretraining on generic poetry may have been more helpful in seeing improvements in perplexity.

However, when combing pretraining and augmentation, we actually saw the best results in perplexity. This can be explained by the significant increase in amount of data from the pretraining helping to balance out the potential negative effects of augmentation. Thus, the augmentation helps with overfitting without taking away too much from the artistry that is prevalent in poetry.

For the subjective evaluation, we decided to compare the raw GPT-3 model, haikus written by real-life poets, and our model with the best perplexity: GPT-3 pretrained on generic poetry and finetuned with augmented Haikus. As seen from the results, after finetuning the GPT-3 model, we saw significant increases in artistry, which was the main goal of the project. However, an increase in artistry led to a sacrifice in comprehensibility and relevance to the topic word. This can be explained by the fact that poems that are often seen as more artistic are often more creative and open to interpretation, thus leading to a less straightforward meaning.

However, we were able improve comprehensibility and relevance by pretraining on generic poems and finetuning on more haikus, limiting the amount of over fitting. While our final model has a much improved artistry score, it is still worse in all three categories when compared to real-life haikus.

7 Conclusion

Haiku generation remains an ongoing challenge in the field of NLP and deep learning. From our research, we were able to decrease perplexity of the raw GPT-3 model after pretraining, finetuning and applying data augmentation. However, as shown by the subjective tests, comprehensibility decreased from haikus generated by the raw GPT-3 model. Additionally, there is a level of artistry which is found in haikus authored by human poets that still cannot be captured through AI generation.

Additionally, a limitation of this study is that the method by which we assigned topic words was on the arbitrary side. Thus, the extent to which a haiku reflects the topic word that it was assigned to may be variable, which would affect how relevant the haikus generated by our model would be the given topic word. A way to mitigate this problem is to devise a better method for assigning topic words. One approach would be to gather the opinions of a large number of people, and choose the most common answer. Another method would be to implement an automated process. However, while the latter method might be consistent, it could be lacking in accuracy.

For future directions, we plan on exploring various strategies in an attempt to improve our model for generating English haikus. We hope that these potential strategies will increase comprehensibility while maintaining the improved artistry we have already seen. These strategies include training the model on a larger dataset, including multiple topic words/prompts for each haiku.

We will also extend our model to be able to take a given haiku and a new topic, then modify the original haiku to reflect the new topic. This will require compiling a new dataset for training. We expect each entry of this dataset to consist of an existing haiku (such as those in our current training set) and a new topic word as the input, and a modified haiku that we will compose as the output. In order to test this, we will use a ROUGE evaluation metric to see how closely the two match.

Lastly, we may also explore the possibility of generating haikus that follow the traditional 5-7-5 syllable structure, given a topic. In the field of natural language processing currently, it is difficult to precisely control the exact number of syllables given the nature of the English language. Even so, a model known as Deep Haiku has shown some promise in conforming to the 5-7-5 syllable structure of haikus (`https://github.com/robgon-art/DeepHaiku`). Deep Haiku is based on GPT-J, a 6 billion parameter model released by Eleuther AI. Out of 20 candidate haikus generated, 11 followed the 5-7-5 syllable pattern, much better than the raw GPT-3 model. We are interested in improving on this using GPT-3, and seeing if emphasis on syllabic structure comes at a cost to the artistic nature of the poem.

References

[1] Academy of American Poets. (n.d.). Haiku. Poets.org. Retrieved February 9, 2022, from https://poets.org/glossary/haiku Rafal Jozefowicz, Wojciech Zaremba, Ilya Sutskever. An Empirical Exploration of Recurrent Network Architectures. In ICML 2015.

[2] Wu, Klyen, Ito, Chen. Haiku generation using deep neural networks. In Proceedings of the 23rd Annual Meeting of the Association for Natural Language Processing, 2017.

[3] OpenAI API, https://beta.openai.com/docs/engines/gpt-3.

[4] "Haiku." Encyclopædia Britannica, Encyclopædia Britannica, Inc., https://www.britannica.com/art/haiku.

[6] Wu, Klyen, Ito, Chen. Haiku generation using deep neural networks. In Proceedings of the 23rd Annual Meeting of the Association for Natural Language Processing, 2017.

[7] Yi, X., Sun, M., Li, R., Li, W. (2018). Automatic poetry generation with mutual reinforcement learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3143-3153).

[8] Miceli, G. (2021). Haiku Generation: A Tranformer Based Approach With Lots of Control. https://www.jamez.it/blog/wp-content/uploads/2021/05/Haiku-Generation-A-Transformer-Based-Approach-With-Lots-Of-Control.pdf

[9] Zou, K. Wei, J. (2020) EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. https://www.jamez.it/blog/wp-content/uploads/2021/05/Haiku-Generation-A-Transformer-Based-Approach-With-Lots-Of-Control.pdf

[10] OpenAI API, https://beta.openai.com/docs/guides/fine-tuning.

[11] “What Is WordNet?” Princeton University, The Trustees of Princeton University, https://wordnet.princeton.edu/.

[12] Emma Baldwin, et al. “10 Of the Best Haikus to Read" Poem Analysis, 1 July 2021, https://poemanalysis.com/best-poems/haikus/.

[13] Voutiritsas, Thea. “10 Vivid Haikus to Leave You Breathless.” Read Poetry, 21 July 2021, https://www.readpoetry.com/10-vivid-haikus-to-leave-you-breathless/.

[14] “Haikus about Nature: Text amp; Image Quotes.” QuoteReel, https://quotereel.com/haikus-about-nature/.

[15] “67 Haiku Poems - Types and Examples of Haiku.” Family Friend Poems, https://www.familyfriendpoems.com/poems/other/haiku/.

[16] Madhukalya, Anwesha. “30 Mesmerising Haikus That Perfectly Capture the Essence of Life and Loss.” ScoopWhoop, ScoopWhoop, 16 June 2016, https://www.scoopwhoop.com/inothernews/mesmerising-haikus-on-life-and-loss/.

[17] AI, Surge. “Evaluating Language Models: An Introduction to Perplexity in NLP.” Medium, Medium, 15 Dec. 2021, https://surge-ai.medium.com/evaluating-language-models-an-introduction-to-perplexity-in-nlp-f6019f7fb914.