SI630 Final Project: Poetry Generation

Fangzhe Li
University of Michigan / 536 S. Forest Ave
School of Information / Ann Arbor, MI
fangzhel@umich.edu

1 Introduction

Classical Chinese poems excel themselves by their conciseness and elegance. It is always an interesting task to create poem. The problem to solve is that given some hints as the first line, the output is supposed to be a quatrain with the same amount of Chinese characters in each line which fits the rhythm. Many different methods have been applied for this problem, which can be works can be divided into three main categories: (1) template-based method (2) statistical machine translation model (He et al., 2012) (3) deep learning method(Li et al., 2018a) and (Zhang et al., 2017). But there is no paper published to use GPT model to generate poems till now. In this paper, we fine-tuned a pre-trained Chinese GPT-2 model on 10,0000 four-line poems and evaluate the result by human evaluation and BLEU-1. The result shows that even though the consistency and innovation of the average level of these poems are still have some gap from the average of human poets, this model can generate some really good poem by chance. More work can be done to have an auto criteria for good poems so that we can make use of the efficiency of poetry generation by model to get a lot of good poems. In homework 1, we have done a binary-classification problem in English. Through this project, the difficulties and challenges of doing nature language processing on a different language as mentioned in the first lecture of our class is clearly showed. Except for the reason that we want our "poet" to surprise us with beautiful rhythms, the other natural language generation tasks can also be benefited from the research on poem generation. (Li et al., 2018b)

2 Task Definition and Data

2.1 Task Definition

Given some collections of classical Chinese poem as the corpus, a model that given some keywords as the first few characters of the poem from users as input and outputs a quatrain with either 5 or 7 Chinese characters in each line which focuses on the topic of keywords and fits the rhythm is wanted. In another way, let $X = (X_1, X_2, \ldots, X_n)$ be a sequence of keywords, where X_j represents jth word in X, the output will be $Y = (Y_1, Y_2, \ldots, Y_{28})$, where y_j represents jth character in Y.

2.2 Data

I found the data on the public repo in GitHub called THU Chinese Classical Poetry Corpus (THU-CCPC). (Guo et al., 2019) https://github.com/chinese-poetry/chinese-poetry/tree/master/json There are 109,727 quatrains stored in json. Each data contains dynasty, author, content, title and keywords. In this model, only contents are used to fine-tuned model and BLEU-1 is calculated using the first line of quatrains as input. There is another poem dataset in https://github.com/Werneror/Poetry, which had more poems. Finally, we use THU-CCPC as our training data and have 109,727 quatrains.

3 Methodology

3.1 Preprocess

Here only paragraphs in the dataset is used. More information like authors and titles can be used for the future improved version. To mark the starting and endpoint of poems, we add <start> and <end> at the beginning and

end of the text. We change "|" in dataset to "，" and "。" separately to indicate different lines in poems. So now every input will be 26 or 34 characters in training dataset. For the test dataset, we get the number of characters for this poem to see if it is 5 characters or 7 characters a line. The first line of the poem are extracted for as the prompt for later generation. Also, we keep the last three lines for origin poems to evaluate BLEU later.

3.2 Model

I used uer/gpt2-chinese-poem in Hugging Face as my pre-trained model. Although this model is called poem, it is actually trained on multiple types of ancient Chinese sources including essays, poems, cis, etc. So when try using it directly, sometimes we can get some patterns that are definitely from Chinese ancient poem. But it is still a great pre-trained model for this task because even though poems have specific pattern compared to essays and cis, they are all sources of ancient Chinese, which worked much better compared to model trained based on model Chinese. We trained this pre-trained model for 6 epochs with our data and used a fixed learning rate of 2e-4. We used TextGenerationPipeline to generate our result after training the model.

4 Related Work

Generally, according to the methodology, works can be divided into three main categories: (1) template-based method (2) statistical machine translation model (He et al., 2012) (3) deep learning method(Li et al., 2018a) and (Zhang et al., 2017). In (He et al., 2012), the authors use the keywords as the input and generate the poem sentence by sentence through a phrase-based SMT model. In this way, each sentence takes all the previous sentences into consideration to ensure coherence between lines. Also, the authors make a comparison between BLEU and human evaluation to show BLEU metric is a good way to evaluate poem generation models. In (Li et al., 2018a), the authors use CAVE to generate novelty and discriminator to ensure coherence. They combines CAVE with adversarial training. In (Zhang et al., 2017), a memory-augmented neural model is used to solve the problem that the model only generate poems based on general rule and has very few innovations. In (Ghazvininejad et al., 2016), the authors use an interesting preprocess before using encoder-decoder model, they asked some input as the keywords of the poem, they make full use of word2vec to find words related to keywords and choose the words fit the rhyme. Put these words to the encoder and let decoder generate the poem. I think I can try this kind of preprocess to fix the output in some spot to make a better performance on rhyme. In (Yan, 2016), the author uses a line-by-line generation encoder-decoder system, the interesting part in this paper is that the author tries to mimic the process of human polishing the poem by use the input of the first round (from writing intention) and the output poems from the the first round as the input of next round to mimic polishing. For LSTM model, I am not quite sure if I can do the similar process to mimic polishing but it is an interesting thought.

5 Evaluation and Results

Since it is very subjective to judge whether a poem is a good way, it is very hard to find a good criteria to assess the poems generated. After reading the literature in the related works part, we find there are two main ways to evaluate the machine-generated poems: human expert evaluation and BLEU (Papineni et al., 2002), which is a metric used widely in machine translation.

5.1 Human expert evaluation

In most of paper related to poem generation, human expert evaluation is used as the main criteria for evaluation because the poem is relatively complicated form of literature and it is very hard to quantitatively evaluate. Human evaluation can be done by setting several group of standard: Rhythm, Consistency, Fluency and Innovation. Another task can be done by human evaluation is that let experts tell if one poem is written by human poets or generated by our model, we can get the accuracy as the criteria. I expect to have the accuracy of distinguishing below 0.7.In this evaluation, four experts who have got or will get at least bachelor's degree in Chinese language and literature were invited to do two tasks

mentioned above independently.

5.1.1 Human score

In this task, the setting is that two poems are randomly selected from the test dataset. Two poems were generated by model from the first sentence of two poems separately, then we labeled the poem A and the poem generated from poem A as poem 1 and 2. Poem B and the poem generated from poem B as poem 3 and 4. The participants were not told which one is written by human and which one is generated by the model. They evaluated these four poems by giving points from 1 to 5 to each perspectives of these four: Rhythm, Consistency, Fluency and Innovation. 1 stands for worst and 5 stands for best, 3 is the average level in their mind for ancient Chinese poems. The means of four scores they gave to two poems written by human is 3, 3.25, 3, 3.5. Except for the fact Innovation is kind of overestimated, the other three dimensions are quite close to average level of poems. But also, the sample of 2 poems are too few but due to evaluation of a poem is about 5 minutes for an expert, the author can't afford asking people to evaluate a lot of poems, which may be improved in the future research.

For Rhythm, we can see from 1 that there is no big difference between poems written by human and generated by the model. Even though mean for Rhythm is 2.625, it means our participants have higher standard to poems on Rhythm, but it doesn't mean model-generated poems are bad.

For Consistency, we can see from 2 that there is an obvious difference between poems written by human and generated by the model. Poems written by human are much better than poems generated by the model. After taking a close look at, we can see that in some cases, the consistency between different lines are not kept well. An example given by a participant is that in poem 2, the first line mentioned the hotness of summer, but the third line described the scenery of fall, which is inconsistent in the poem.

For Fluency, we can see from 3 that there is no big difference between poems written by human and generated by the model, which means our model really learn the way the human poets used to make their lines.

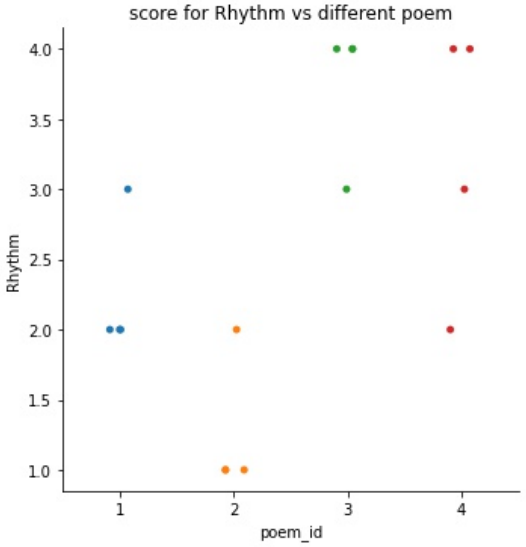

Figure 1: This picture shows the rhythm scores on different poems

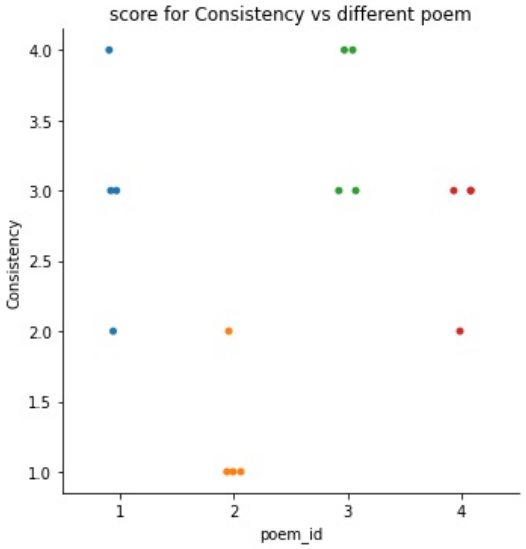

Figure 2: This picture shows the Consistency scores on different poems

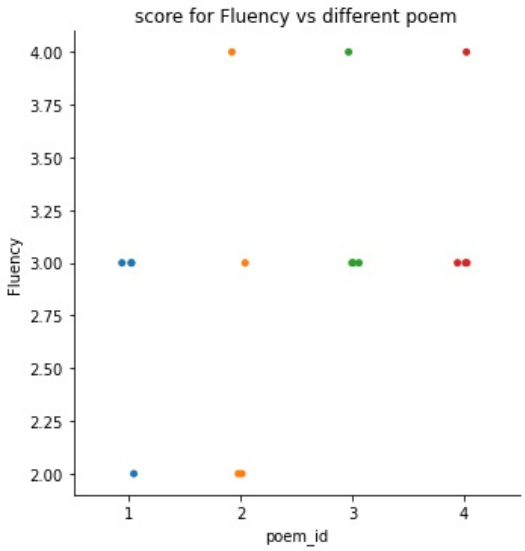

Figure 3: This picture shows the Fluency scores on different poems

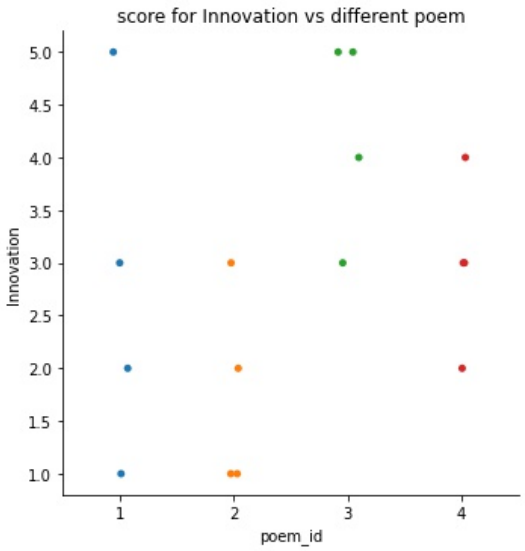

Figure 4: This picture shows the Innovation scores on different poems

For Innovation, we can see from 4 that the scores for poems written by human are slightly higher than the ones for poems generated by the model. However, it doesn't mean poems generated by model is not innovative. Since this is a very small sample size, we will show in next section that poems generated by model can also be quite innovative.

5.1.2 Confusion Test

In this task, four participants were given 10 poems random selected from the test dataset. They were told 5 of them are written by human and 5 of them are generated by the model. This task is about random selecting 10 poems from the dataset, since it is a huge dataset, it is very unlikely these poems are famous ones that already learnt by people by heart. Using the first line as the input, use this model to generate another five poems and mix with the 5 human-written poems, let people decide which ones belong to human and which ones belong to the model. They were asked to label each poem whether it was a poem written by human or AI. The accuracy for four participants are 0.8, 0.6, 0.6, 0.8 separately. The baseline for this test is the uni-gram language model and the accuracy for four participants are 1.0, 1.0, 1.0, 1.0. We can refer to following table to see the overall accuracy for our model and baseline.

After taking a close look at our result, we found it is interesting that there is a poem generated by the model (poem 7) was token as a human-written one by three participants. The poem is quite innovative and the style is quite like a human poet. Let's enjoy this poem here:

	Baseline	GPT-2
# of poems	40	40
Mistakes	0	12
Accuracy	100%	70%

春来雪有香，
香暖不胜长。
一任随风去，
多应作雨伤。

A translation for that poem is: The snow is fragrant when spring comes. But the perfume can't last long (like the snow) when warmth comes. The snow will go with the wind (like the perfume), becoming spring rain and weeping. This poem is quite good from all perspective. For rhythm, " 香", " 长" and " 伤" both follow "ang" pronunciation. For consistency, this poem focus on the snowing in the spring to express how perishable and beautiful the spring snow is. The use of language is quite fluent and has the style of human poet. As for innovation, the author using the analogy between perfume and spring snow and personification for spring snow to convey its love for spring snow. In all, these four characters of the poem make it confused for three participants

	Baseline	GPT-2
BLEU-1	0.0144	0.0683

to put it the label of written by human.

5.2 BLEU

Another quantitative way to evaluate poems is the BLEU. BLEU is initially used in machine translation to evaluate the similarity between human translation and generated machine translation. Here a slightly different way will be used: first keywords is extracted from a human-written poem, then these keywords are put into model as input, we compare the BLEU between the output the the original poem. Here a low BLEU value is expected because there are more ways to write a poem compared to general translation. So BLEU is better used to compare the result of different poem generation model. In (He et al., 2012), the author made an analysis to show BLEU actually a good criteria because it shows some relationship with the result of human evaluation. We invite eighteen volunteers, who are knowledgeable in classical Chinese poetry from reading to writing, to evaluate the results of various methods. In our task, we got three lines generated by our model by using the first line of the poems in test dataset as input. These three generated lines worked as the candidate while the origin last three lines worked as the reference. The mean for BLEU-1 score is in following table. It is seen that our model performs much better than uni-gram language model, which is our baseline.

6 Discussion

From the above evaluation, it is seen that the poems generated by our model is actually quite decent in rhythm and fluency. However, they don't perform as good as human-written poems in innovation and especially consistency. The problem of inconsistency is predictable because of the shortcomings of GPT-2, where the model focus more on the character level fluency but not the more sentence level consistency. Rhythm is something between character level and sentence level, and our model is doing ok on that. For the inconsistency, the problem behind that is the lack of common knowledge like the lotus mostly appear in summer, so it will be improper to have it when describing a fall scenery. These knowledge is what human poet have between lines but GPT-2 doesn't seem to grab that. Another interesting point is that from task 2 in human evaluation, we see there were some good poem emerging and had really confused our experts, which means that our model has the ability the generate really good poem by chance. Even though our ultimate goal for AI poet is that they can generate good poems for any topic and with any hints, it will still be great if our model can generate some good poems and pick it up for us. Only if we have a good classifier for whether it is a good poem, we can let the model keep generating poems with different hints in a high efficiency and use that classifier to find good one there. Compared to human poets, the biggest advantage of AI poets is that AI only take a few seconds to generate one while human poets take days, months or even years to make one. Find a good criteria and train a classifier or use clustering will be a good next step. As for BLEU, even though it can indicate how the model get the basic idea the author want to say, maybe a more semantic related method will be a better metric.

7 Conclusion

Poetry generation is a challenging and interesting topic. In this paper, we fine-tuned a pre-trained Chinese GPT-2 model on 10,0000 four-line poems and evaluate the result by human evaluation and BLEU-1. The result shows that even though the consistency and innovation of the average level of these poems are still have some gap from the average of human poets, this model can generate some really good poem by chance. More work can be done to have an auto criteria for good poems so that we can make use of the efficiency of poetry generation by model to get a lot of good poems.

8 Other Things We Tried

For this project, the baseline I use is a character-level unigram language model revised from the in-class notebook. I tried to change it to smoothed n-gram language model to improve the performance of baseline. But the structure for that unigram one is kind of

different from the way we did in class, and it needs a decent amount of work to make it, so I just stopped there.

9 What You Would Have Done Differently or Next

As discussed above, an auto criteria for good poems is needed so that we can make use of the efficiency of poetry generation by model to get a lot of good poems. One way easy to think is definitely clustering, but it is kind of hard to find a cluster with "good poems". Another way is that using some comments made by literary critics as labeled data to train a classifier. Another thing can be done is to improve the way of generation to force the model put more weight on characters from last few sentences.

10 Acknowledgments

Here I want to thank for the help of David, who spent his time after lectures discussing with and guiding me on the structure of this project. Also I want to thank Jiaoyang for his introduction for some pre-trained models that may help. Finally I want to thank my high school classmates hsy, hya, wqy and their friends xyx, who worked as participants for human evaluation even though they were busy for their thesis defenses. And hope they can continue their journey of master's degree on Chinese literature and language well.

References

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1183–1191, Austin, Texas. Association for Computational Linguistics.

Zhipeng Guo, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, Jiannan Liang, Huimin Chen, Yuhui Zhang, and Ruoyu Li. 2019. Jiuge: A human-machine collaborative Chinese classical poetry generation system. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 25–30, Florence, Italy.

Jing He, Ming Zhou, and Long Jiang. 2012. Generating chinese classical poems with statistical machine translation models. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, AAAI'12, page 1650–1656. AAAI Press.

Juntao Li, Yan Song, Haisong Zhang, Dongmin Chen, Shuming Shi, Dongyan Zhao, and Rui Yan. 2018a. Generating classical chinese poems via conditional variational autoencoder and adversarial training. pages 3890–3900.

Juntao Li, Yan Song, Haisong Zhang, Dongmin Chen, Shuming Shi, Dongyan Zhao, and Rui Yan. 2018b. Generating classical Chinese poems via conditional variational autoencoder and adversarial training. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3890–3900, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Rui Yan. 2016. I, poet: Automatic poetry composition through recurrent neural networks with iterative polishing schema. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, page 2238–2244. AAAI Press.

Jiyuan Zhang, Yang Feng, Dong Wang, Yang Wang, Andrew Abel, Shiyue Zhang, and Andi Zhang. 2017. Flexible and creative Chinese poetry generation using neural memory. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1364–1373, Vancouver, Canada. Association for Computational Linguistics.