Generation of Chinese classical poetry based on pre-trained model

1st * Ziyao Wang  
dep. School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu, China  
* Corresponding author: 2126546982@qq.com

2nd Lujin Guan  
College Of Life Science Technology Huazhong University  
line Of Science And Technology  
Wuhan, Hubei, China  
e-mail address: guanlujin@gmail.com

3rd Guanyu Liu  
School of Computer and Cyberspace Security  
Communication University of China  
Beijing, China  
e-mail address: 932662798@qq.com

4th Jun Ma  
dep. School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu, China  
e-mail address: majun@lzu.edu.cn

Abstract—Chinese traditional poetry is an important art form, and it is a meaningful task to use artificial intelligence algorithm to generate poetry. In the past, researchers have proposed different generation algorithms, but these algorithms have some limitations. This paper mainly tries to use BART and other pre-trained models, proposes FS2TEXT and RR2TEXT to generate metrical poetry text and even specific style poetry text, and solves the problem that the user’s writing intention gradually reduces the relevance of the generated poetry text. Finally, in order to test the effect of the algorithm, this paper imitates Turing test and finds more than 600 testers to distinguish the works generated by the algorithm from the works of real human poets, and the results show that they can’t tell whether the work is from an algorithm or a human. This shows that the algorithm proposed in this paper works well.

Keywords: Deep Learning; Text Generation; Pre-trained model; BART; Classical Chinese poetry; Turing test

I. INTRODUCTION

Chinese classical poetry occupies an important position in Chinese classical literature and has a far-reaching influence. It can be told that Chinese classical poetry is the most important artistic form in East Asian Cultural Sphere. Metrical poetry is a very important part of Chinese classical poetry, which is also called quatrains [1]. It can be divided into five words and seven words, no matter which one should follow the structural, tonality and semantic requirements of ancient poetry. Word to word and line to line all meet specific phonological patterns. For example: The last character of the second, fourth and (possibly) first lines must rhyme, and the third line has no restrictions. In addition, poetry must follow a prescribed tone pattern: Tones or tones of the tones.

Nowadays, with the progress of technology, it is possible to use machines to generate quatrains [2]. The research on automatically generated poetry began in the 1960s and has been a focus in recent decades. In the field of generating ancient poems, after experiencing the use of templates, genetic algorithms, based on other text generation methods. With the focus on the intersection of deep learning and natural language process, neural networks have been widely used in poetry creation [3]. In this paper, we created the completest Chinese classical poetry data set so far by web crawler and inputting data from ancient books manually. Taking advantage of the data set, we trained a specific pre-trained model of Chinese classical poetry named BART-poem, which is different from both modern Chinese and ancient Chinese. We believe that the fundamental works can make contribution to the development in the field. Above that, we created FS2TEXT model, where after inputting genre, first sentence, keywords and subject words we can get an ancient poem which people hardly distinguished from true poems in the test we sent. In addition, we amended the model so that it can create sub-rhyme poems.

The main contributions of this paper are:

1. Constructed the most complete data set of Chinese classical poetry
2. Trained a Pre-trained model suitable for the field of Chinese classical poetry
3. Proposed two different poetry generation algorithms to adapt to different scenes.
4. Solved the problem that the user’s writing intention gradually reduces the relevance of the generated poetry text.

II. METHODS

A. BART - poem

Although there are pre-trained models for modern or ancient Chinese, the language of classical poetry is quite different from that of modern Chinese, and when it occurs to the combination of words and the logic of affixes, there are also some differences with ancient Chinese in pragmatics and grammar [4]. Therefore, it is necessary to train a pre-training model suitable for poetry task. By evaluating the effects of different models, the author selected BART model from BERT, Roberta, T5 and BART model for training. The size of this model is about 1.5 GB, and the main parameters are shown in Table 1:

| Parameter | Value |
|-----------|-------|
| Model     | BART  |
| Size      | 1.5 GB |
| Parameters| ...   |

Table 1.
The author uses UER.py to train the model, firstly, Chinese characters that appear more than or equal to 100 times in the poetry corpus are added to the thesaurus, set the sequence length to 64 and specify the data processor as BART mode to preprocess the data. Then set the batch size to 64, span max length to 3, and train 60000 steps. Finally, the accuracy of the model is stable at 0.91, and the loss is about 0.50. We named this model BART-poem[5].

B. Data Processing

(1) Theme Words Extraction

Firstly, the poetry text is divided into words by THULAC[13] developed by Tsinghua University, and then the words in the stop-word list are removed, and the remaining words are extracted by TF-IDF algorithm [6]. TF-IDF (term Frequency-Inverse Doku-ment frequency) is a commonly used weighting technology for information retrieval and data mining [7]. TF is Term re-quency, and IDF is Inverse Document Frequency [8-11].

The author uses this algorithm to extract subject headings, and the number of text subject headings extracted from each poem is 1/12 of the length of the poem text [12].

(2) Key Character Extraction

The author notes the concept of the "poetic eye" in classical poetry, where a word in a poem is central to the context, the meaning of other words around this word, then this word is the poem eye.

The author chooses to use the word vectors trained by Shen Li et al. using word2vec algorithm on the corpus of the four databases. After removing the suspended words of the poem text, the remaining words are transformed into word vectors for representation. By calculating the center points of these word vectors and finding the closest Euclidean distance from the center point, the characters corresponding to these vectors are the keywords of the poem. The number of keywords extracted from each poem is 1/10 of the length of the poem text [13].

(3) Genre Judgment

Metrical poetry as the object of the author's study is divided into four types, which are different in sentence length and sentence number.

Five words quatrains: Each sentence is made up of five words, and each quatrain contains four sentences.

Seven words quatrains: Each sentence is made up of seven words, and each quatrain contains four sentences.

Five words lvshi: Each sentence is made up of five words, and each lvshi contains eight sentences.

Seven words lvshi: Each sentence is made up of seven words, and each lvshi contains eight sentences.

Many previous poetry generation models regard the task of poetry generation as "mapping from title to text", however, there is sometimes no clear correspondence between title and poem content, such as the famous "Falling Flower poem", there are at least a thousand poems titled "Falling Flowers," all with the same title but different content, therefore, there is a lack of proper mapping between the title and the text of the poem. So the author decided to change the main mapping relationship into "first sentence to whole poem", which is almost a one-to-one mapping. In order to solve the problem that the following sentences are gradually less relevant to the user's writing intention, the author uses subject headings and keywords to control the generation process of poetry and the generation of specific style works.

(2) FS2TEXT Model Structure

The author performed the seq2seq task on the basis of the already trained BART-poem model. The input is "the first sentence of each poem &the random number subject words of this poem &the keywords of the random number of this poem&the poem genre", the output is the full text of the poem. Figure 1 shows FS2TEXT.

Figure 1. FS2TEXT

(3) Generation of Specific Style Poetry

If you want to generate poetry in a particular style, you need to construct a dataset that contains only poetry in a particular style, extract all the subject words and keywords, and on the basis of the trained FS2TEXT model, use the subject words and keywords in the dataset containing only specific style poetry, input this into the model along with the genre and the first sentence to get a poem of a particular style.

D. RR2TEXT

(1) Overview

Second rhyme is a kind of creation behavior that the poems are created by using the same genre and rhyme as the target work, which is very popular in poetry creation. Du Fu's "Eight Pieces of Autumn Prosperity", Wang Shizhen's "Four Chapters of Autumn and Willow", Huang Jingren's "Sixteen Pieces of Qi
Huai” and other poems have been repeatedly repeated by later generations. We can abstract the act of sub-rhyme as creating a work that is similar in style to the original one, with the same rhyme and genre. The first sentence of a rhyme work should not be the same as the original work, so the author uses rhymes to generate a rhyme poem work, and uses the key words and subject words of the original poem to control that the rhyme work and the original work have a similar style.

(2) RR2TEXT Model Structure

The author performed the seq2seq task on the basis of the already trained BART-poem model. The input is “rhyme of each poem random number of this poem: The amount of keywords in this poem is a random number of keywords,” and the output is the full text of the poem. Since subrhyme works are generally similar to the original works in style and content, we can control the generated works by entering the same subject words and keywords as the original works, however, most subrhyme works are not particularly similar to the original works, therefore, when performing the generation task, the author will only input part of the subject words and keywords of the original work to ensure that the generated work will not be highly similar to the original work. Figure 2 shows RR2TEXT.

III. RESULTS AND ANALYSIS

A. Data Set

The authors have constructed the most comprehensive poetry dataset ever made publicly. The poetry works of each era are put into a csv file, a total of about 1.2 million pieces, divided into four fields: "title", "dynasty", "author" and "content" for storage.

The open source project collects about 800,000 ancient poems from the pre-Qin period to the modern era. The data of ancient poetry is stored in multiple csv files according to dynasties, including four fields: "Title", "Dynasty", "author" and "content". Some rare characters in the corpus of poetry cannot be displayed, so "to replace" is used.

The project lacks many works of Ming and Qing. Through web crawlers, the author collected many works of poets in Jin, Yuan, Ming and Qing dynasties from various poetry websites.

Due to the scarcity of some poetry materials, there are no open digital resources on the Internet. The author manually inputted the PDF files of ancient books and turned them into convenient text materials.

B. FS2TEXT Training

On the dataset constructed by the author, the BART-poem model is used to fine-tune the seq2seq task. Set the sequence length to 64 and specify the data-processor as BART mode to preprocess the data. Then set batch size to 64 and train to loss of about 2.60. At this time, the model has been able to generate poetry of reasonable quality by manual judgment, and the FS2TEXT model is obtained after stopping training. At this time, specify the genre, input the first sentence and a certain number of keywords and subject words, you can generate the corresponding poetry. The user’s intention can influence the generated text through keywords and subject headings. Table 2 shows the FS2TEXT results.

| Input                  | Output                  |
|------------------------|-------------------------|
| Genre                  | 七言绝句                |
| Theme words            | 青鸟                  |
| Key Character          | 烟二百一山               |
| First sentence         | 杨柳花飞芜草青         |

The results generated by the model conform to the level and oblique pattern and format of the specified genre, and it can be seen that the subject words and keywords affect the global text.

C. FS2TEXT Training

The author selected some works of Wang Yanhong and Sun Yuanxiang as the data set of colorful poems, extracted the subject words and keywords in the data set of colorful poems, and specified in the trained FS2TEXT model that only subject terms and keywords in this dataset can be used and be generated in the style of erotic poetry works. Table 3 shows the FS2TEXT-amorous result.

| Input                  | Output                  |
|------------------------|-------------------------|
| Genre                  | 七言律诗                |
| Theme words            | 相见时难别亦难，临歧无奈暂盘桓 |
| Key Character          | 舟沿碧草同千里，人隔青天共一峦 |
| First sentence         | 从今珍重琼瑶字，莫作鸳鸯万缕看 |

D. RR2TEXT Training

Fine-tune the BART-POEM model to perform seq2seq task, set the sequence length to 64, and specify the data processor as bart mode to preprocess data. Then set the batch size to 64 and train to a loss of about 2.80. At this time, the model has been able to generate poetry of reasonable quality by manual judgment, and the RR2TEXT model is obtained by stopping training. Users input the works they want to rhyme into the program, after processing, the program analyzes the genre, rhyme, theme words, keywords, etc., and then inputs them into the model. The sub rhyme works are obtained, as shown in Table 4.
E. Questionnaire Survey (Turing Test)

In order to verify the quality of the poetry produced by the model, the authors created a set of questions that mimicked the Turing Test. The process is as follows: Select the history of literature, the most important poets (representing the higher level of human beings); Select obscure works from these poets (so that no one has read them and the results are not affected); According to the genre, 4 poems were selected from each genre, and 16 poems were selected in total; After the selection, input the first sentence to generate works by FS2TEXT model, put the works generated by FS2TEXT model and the original work together to choose one, and let people judge which one is created by AI. The test exploded in a small area and was picked up by official accounts such as "Young Chinese Stories Micro Magazine". Finally, 616 valid responses were received, and professional researchers from prestigious universities such as Peking University and Fudan University participated in the test. There are also members of major poetry societies and other metrical poetry writers. Generally speaking, the subjects' attainments in poetry are obviously higher than ordinary people.

According to the questionnaire, the correct rate of most questions fluctuated around 50%, and only three questions were lower than 40% or higher than 60%. Overall, the testers made 9,856 choices and got them right 4,960 times, which is about 50.33% of the total and very close to 50%, therefore, we made 9,856 choices and got them right 4,960 times, which is about 50.33% of the total and very close to 50%, therefore, we draw the conclusion that even for more professional people with high attainments in poetry, it is difficult to distinguish the works of AI [14-15].

IV. CONCLUSION

Chinese classical poetry is dominated by metrical poetry (i.e. modern poetry), which is stereotyped in the early Tang Dynasty, and the existing computational power and algorithms only support explicit rules and the generation of metrical verse text with short text length, so the author takes the generation of metrical verse text as the research object. This paper mainly tries BART and other pre-training models, and proposes FS2TEXT and RR2TEXT to generate metrical poetry texts and even poetry texts of specific styles, such as "Jiangxi School Style", "colourful style", etc., it also solves the problem that the correlation between the user's writing intention and the generated poetry text is decreasing gradually. In order to benefit the huge number of poetry lovers, the author's model has reached cooperation with relevant poetry websites for everyone to use, hoping that the work of this model can bring people creative inspiration.

References

[1] Lai, J. . (2020). Wellspring of inspiration: the mandarin union version and modern chinese poetry in the early twentieth century. Breast Cancer Online, 30(1), 163-177.
[2] Nenu, T. . (2022). Douglashofstadter's gdelian philosophy of mind. Journal of Artificial Intelligence and Consciousness, 09(02), 241-266.
[3] Ma, M. Q. . (2021). Guest editor's introduction – poetry elsewhere, elsewhere poetry: a poetics of the interstitial in contemporary american poetry. CounterText, 7(3), 289-293.
[4] Zhang, F. . (2022). Artistic features of lu you's cicada chanting poetry. Arts Studies and Criticism, 3(1), 14-16.
[5] Boldireff, A. , & Bober, C. . (2022). Reconsidering poetry's difficulty and value in english as a second language and english as a foreign language education: a systematic review. Pedagogical Linguistics, 3(1), 57-83.
[6] Shiba, K. . (2021). Izumi shikibu nikki in manga: recreating classical poetry, letters, and shared sensibilities. University Library System, University of Pittsburgh4(1)12-15.
[7] Shahiditabar, M. , & Pourghasemian, H. . (2021). Conceptual metaphors of ayrlq (separation) in azerbaijani turkish poetry. Azerbajian Shahid Madani University1(1)21-25.
[8] James, P. . (2022). Thinking with birds: avian song and psychology in old english poetry. ISLE: Interdisciplinary Studies in Literature and Environment4(3), 3.
[9] Daryanavard, Z. , Balavi, R. , & Khezri, A. . (2021). Semantic functions of repetitive synthesis in adnan al-sayegh's poetry.9 (100)36-41.
[10] Morales-Rodas, L. Y. , Galimberti, J. V. , GM Cerda-Solís, & Riera-Guachichultlea, E. J. . (2021). Blackout poetry: re-envisioning writing strategy in english learning. Imprenta y Casa Editora "Coni"4(3)7-12.
[11] Ajmi, H. , Ibrahim, M. , & Ismail, Y. B. . (2021). Colors' symbolism in the modern omani poetry: an analytical descriptive study. 4(18)5-12.
[12] Niall, B. . (2021). Ornithological competence and literary biodiversity in spanish american poetry. ISLE: Interdisciplinary Studies in Literature and Environment7(3), 3-10.
[13] Mazur, E. . (2021). Poetry after 1989 and its contexts in polish language education. Uniwersytetu Marii Curie-Sklodowskiej w Lubliniec7(6)4-6.
[14] Carpenter, J. R. . (2022). Nested folders: on birds in digital poetry. ISLE: Interdisciplinary Studies in Literature and Environment9(3), 3-5.
[15] Inomnazarov, M. , & Jumaeva, K. . (2021). Developmental stages of irfani poetry. International Journal of Multicultural and Multireligious Understanding (IMMU)8(8)6-11.

| TABLE IV. RR2TEXT RESULTS |
|---------------------------|
| **Input**  | **Output** |
| 独起凭栏对晓风 | 日没荒墟生晓风 |
| 溪流春水小桥东 | 溪流流碧山东 |
| 始知昨夜红楼梦 | 不知渔父相扶醉 |
| 身在桃花万树中 | 独立苍茫烟雨中 |