Chinese Poetry Generation with Topic Extraction based on Machine Learning

Tianle Wang
Shanghai Jiao Tong University, Shanghai, China
wtl666wtl@sjtu.edu.cn

Abstract

As a literary from unique to the Chinese, the traditional Chinese poetry displays the unique cultural deposits of China with their special formats and rhymes. In this paper, I propose a simple but useful poetry generating method which first gain user's writing image of each poem by input or automatically generated, then the LSTM model is used to generate the poetry based the outline. In practice, it is found that this method makes the generated poems more fluent and it fits the user's intentions.

1 Introduction

For thousands of years, a large part of the history of Chinese literature has been the history of poetry. There are as many poets as stars in the night sky, so our country is called the country of poetry. So there is no denying that traditional Chines poetry is vital heritage in the glorious ancient Chinese civilization. With assorted types, the traditional poetry can be divided into Tang Poetry, Song Ci, and others, from the perspective of era; or "ancient poetry" and "morden-style poetry" from the perspective of rhymes. Among them, Tang Poetry is undoubtedly the most characteristic model in our country's poetry. Hence the generation of my ancient poems is mainly based on the study of seven-character Tang poems.

Artistic creation is considered to be the most capable job of the most spiritual genius, so it has always been a challenging problem in the field of natural language processing to use machines to imitate humans to write poetry. In recent years, the research of automatic poetry generation has received great attention. These methods usually require the input of the first sentence or word, or select a row from the poetry data set to generate the first row based on several keywords entered by the user, and generate the other rows based on the first row and the previous row. The user's writing intention can only affect the first line, and the remaining lines may not be related to the theme of the poem, which may cause semantic inconsistency when generating the poem.

In this paper, my method not only hopes to generate verses simply by doing similar translations, but also hopes to give each verse its own image. Therefore, by analyzing the verses before learning, select their own keywords for each verse, and re-splicing the verses, and then put them into the LSTM model for learning. When generating poems, users can not only input the first sentence to compose the poem, but also give the user's writing intention, which needs to be a set of keywords. We can use the keyword as the outline of each poem through the generation method similar to the generation of acrostic, and generate the entire poem through keywords. Compared with the previous method, the topic extraction based mechanism has two advantages. Firstly, each line of the generated poem is more closely related to the user's writing intention. Besides, users can generate poems by inputting the first sentence or image. They can also specify the main image of each sentence to specify the perspective, for example, by specifying "grass"(草) and "wind"(風) to achieve different scene changes for each sentence, the generated results are as follows, which well represent the two themes：

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

草>白玉堂前金紫翠，玉釵金殿照青冥。
風>風吹玉樹香光動，露濕紅蕉葉不開。

The rest of this paper is organized as follows. Section 2 describes some previous work on poetry generation. Section 3 describes my topic extraction based poetry generation framework. We introduce the datasets and experimental results in Section 4. Section 5 concludes the paper and puts forward the next step of research and improvement ideas.

2 Overview of Related Work

This section will introduce related work from two aspects based on traditional poetry generation methods and based on deep learning technology.

2.1 Traditional Methods

In the early days, the method of randomly connecting words was used to generate poetry by using simple calculation procedures. But the generated poetry is like a mess of random words. These poems are entirely the result of machines randomly pieced together based on basic words. The poems generated by this method hardly consider the content, form and meaning of the poems, and only happen to be able to write good poems. Strictly speaking, what is produced in this way cannot be called poetry.

In order to make up for this shortcoming, some researchers have proposed a template-based poetry generation method, through which the template is used to limit the smoothness of grammar and semantics. This method will define a good template for the verse, the template contains some fixed words or phrases, and the rest are left blank. That is, remove some keywords in an article as a template, and select new words to fill in the limited words. The words that need to be filled are based on part of speech, such as nouns, verbs, etc.

Poems generated based on templates often have better results. However, this type of system largely depends on the quality of the template, and the flexibility of the system is poor. To improve the quality of poetry, a large number of manual carving and portraying templates are required, which is far from the goal of automatic poetry generation.

In addition to template generation of poems, some researchers have proposed a pattern-based poem generation system, which generates poems that meet the requirements of grammar and rhythm under a set model. For example, the number of sentences, the number of words in each sentence, the proportion of different parts of speech words, etc. will be planned in advance. When generating poems, some uncomplicated algorithms will be used to continuously select the words that best meet the search criteria and fill in the specified positions one by one in the lexicon. Compared with the template generation mode, it has better flexibility.

2.2 Methods Based on Machine Learning

Facing the various problems of traditional methods, RNNPG(1), a Chinese poetry generation model based on RNN neural network is proposed. This method first uses the keywords given by the user to search for expanded meaning groups in the corpus and generate all candidate verses under the related constraints, then use the language model to score them, and select the verse with the highest score as the first sentence. Then the next sentence is generated from this, and then the first two sentences are input to generate the next sentence, and a complete poem is generated in this way.

Compared with traditional methods, the RNNPG model can automatically learn from the corpus without the need for human design evaluation functions, and has better results. However, the user's intention can only be expressed in the first sentence, and the subsequent content generated by the model is difficult to retain the user's central thoughts in the first sentence, which will cause the problem of subject shifting.

The ANMT(2) model is a neural network translation model based on the attention mechanism, it was used in poetry generation shortly after it was proposed. The system introduces the translation model into the problem of poetry generation, and "translates" the last sentence of the poem to the next sentence through the translation model. The process of generating is that the user first provides the first sentence, then the second sentence is generated from the first sentence, and the first and second sentences generate the third sentence, and this process is repeated until the complete poem is generated.

Compared with the ordinary RNN model based on the attention mechanism, the advantage of LSTM is that the problem of gradient disappearance is obviously reduced. The addition of the attention mechanism allows the model to load more historical information, and also makes the connection between sentences closer. However, this model has the same user intention as the previous model, which can only stay in the first sentence and lacks penetrating power. It is difficult to avoid subject drift even if loading more historical information.

So I try to make some improvements on this basis. The main model is still used, but for model training, we not only train the simple text mapping of the first few sentences to generate the next few sentences, but also add a keyword theme to each sentence and train the theme and poetry together. This model not only retains the capabilities of the previous model, but can also solve the problem of topical bias, that is, users can specify the keywords they want to express in each sentence, so as to ensure that each sentence of poetry conforms to the user's ideas. At the same time, conversely, sudden changes in poetry can also be achieved, which usually occur uncontrolled in the above models, achieving sorrows and joys, and simple changes in perspective.

3 Approaches

In this section, I will first briefly introduce my entire method in **Overview**, and some details of the specific method will be introduced in detail in the following.

3.1 Overview

As mentioned in the previous article, the existing machine learning models are often very easy to be disturbed. In the whole process, the machine is only performing mechanical translation, and does not recognize what the written poetry is depicting. As a result, the written poems appear to be connected and smooth, but the meaning changes or shifts. Hence I consider incorporating the way that humans often use when writing poetry and composition, that is, before the pen is written, we tend to make a simple idea of what we want to write. This idea is usually not very precise, and even changes constantly with the pen. , But there is always a theme, or keywords, throughout the whole process, and all text is developed around it. For example, in Zhizhang He's well-known poem *the willow*, it is undoubtedly the willow tree that runs through the whole text.

With this as the direction, it is conceivable that if you first determine the key words for each poem when creating a poem, and expand around it, then the poem will become clear. So I considered disassembling the entire poetry training into two parts. The first part is to perform some processing on the original poetry, specifically to extract keywords from each sentence of each ancient poem, and then mark the keywords and insert them into each sentence. And recombine the modified verses. The second part is to put the modified ancient poems into the existing LSTM model for training. The model only needs a few simple adaptive modifications to work, and the training result will be It will appear in the form of "keywords> ancient poems".

After this model is trained, its functions are diverse. As mentioned above, it has the function of the previous model, that is, it can write poems with the first sentence or several characters, or generate acrostic. In addition, users can input the subject they want to describe in the form of keywords, one or several. The system will refer to the generation of acrostic, fix each subject to the position of the beginning keyword of each sentence, and then start to generate sentence by sentence. Each sentence is generated based on the sentences that have been generated in the previous sentences and the keywords specified by the user. so as to ensure that the given image is portrayed into the ancient verse, the generated ancient verse The number is related to the number of keywords entered. This system also supports the generation of ancient poems that the user specifies several themes and the first sentence or several words, making the generation more flexible.

Table 1: High-frequency Chinese characters

Chinese characters	**Selected times**
風	5884
山	5053
花	4090
日	3604
雲	2809
年	2091
春	1961
水	1516
知	1460
月	1223

3.2 Topic Extraction

We define keywords as some words or phrases that are closely related to the content of the target article. Keywords can often summarize the subject content of a certain aspect of the target article, and are indivisible unit words, generally composed of nouns, such as "tree", "moon", "mountain" and so on. In my poetry writing system, our goal is to extract the topic of existing verses and fix them at the beginning of the sentence in the form of keywords.

Taking into account the uniqueness of Chinese characters, that is, they belong to the Sino-Tibetan language family, which is quite different from Indo-European languages. Specifically, the basic unit of Chinese is "character", which is then formed into words, while Indo-European languages are only dominated by words. Such writing rules make Chinese unable to be accurately divided by the displayed separator like Indo-European languages. Therefore, the keyword acquisition algorithms commonly used in English have relatively general effects in processing Chinese sentences. At the same time, since the object of our processing is ancient Chinese, it is usually more concise and there may be word order structures that are not commonly used in modern Chinese, which makes it more difficult to segment words. And because the number of words in a sentence is very small, there is usually a certain poetic grammar to require that there is no word composed of too many characters, so usually extracting a single character is enough to summarize the topic of the sentence.

Due to the above reasons, I choose here to determine whether it is a keyword based on the frequency of each single Chinese character in all ancient verses. This is a very simple but effective way. In fact, it is often The images that appeared in all poems were immediately divided, such as "moon"(月), "flower"(花) and so on. However, it was also discovered that some Chinese characters that could not highlight the theme often appeared, such as "no"/"not"(不/無), "what"(何). These unmeaning characters will be discarded after simple manual verification. The system will ensure that in the end only clear nouns and verbs are left to express the main idea. The frequently selected characters(selected more than 1000 times) can be seen in Table 1. These high-frequency characters are images that appear frequently and are easy to grasp. There are about 6,000 seven-character ancient poems in the sample, so in fact they account for almost most of the keywords.

This method is practical and fast. Two examples of the reorganized verses are as follows(he selected words are highlighted):

水>錦**水**東流繞錦城，星橋北挂象天星。 春>華陽**春**樹號新豐，行入新都若舊宮。
山>四海此中朝聖主，峨眉**山**下列仙庭。 花>柳色未饒秦地綠，**花**光不滅上陽紅。

3.3 Poem Generation

In the stage of poetry generation, poetry is generated line by line. Each line is generated by taking the specified keyword and all preceding text as input. This process can be considered as a sequence-to-sequence mapping problem. I simply choose LSTM to train the model, which is an improved network specifically aimed at the problem of RNN gradient disappearance, and its structure is more suitable for dealing with forecasting time series problems with long-term dependence. It is very easy to use the LSTM model to train the model described above, so I won't go into details.

Table 2: Comparison of Poems Generated in Two Ways

Basic method	**Topic extraction based method**
春來無伴閑遊少，	春來無伴閑遊少，
何處春風伴客來。	風雨蕭條一夜愁。
一日不知何處去，	山下不知天子處，
一枝花落舊山頭。	水邊長在白頭翁。

Table 3: Final Results

	Basic method is better	**Topic extraction is better**	**Almost the same**
A	1	3	1
B	2	3	0
C	1	1	3
D	1	2	2

4 Experiments

4.1 Dataset

In this article, we focus on the generation of ancient poetry with four or more lines and 7 characters in each line of the same length. This kind of poetry has no specific requirements for rhythm. The poems is mainly derived from Tang poems and some early seven-character poems. Although the metric requirements for "modern-style poetry" began in the Tang Dynasty, after all, the poetry of the Tang Dynasty is still blooming. Many famous poets do not seek metric or even write four-character poems or long and short sentences according to the general verse. The data is mainly collected on the following websites, after a simple selection, a total of 6569 poems：

`https://github.com/chinese-poetry/chinese-poetry/`

4.2 Evaluation

4.2.1 Baselines

Here I set the method of LSTM model training directly on ancient poems without extracting keywords as baseline and employed the same pre-processing method for two methods.

4.2.2 Results

First of all, due to the new ability to write poems on keywords, the new model has undoubtedly been more applicable than the basic method. So now consider evaluating the pros and cons of the poems they made.

Since the basic method does not have the ability to input keywords to generate ancient poems, in order to ensure fairness, I choose some of the opening sentences of the existing uncommon ancient poems as the common beginning of the two, avoiding the ability to determine the theme of the new model when inputting. Then let the two models generate poems. For example, Table 2 is the result of two models based on the first sentence of a poem by Juyi Bai.

Since the poems made by the two have no fixed requirements on the metric, the main method for the evaluation of the ancient poems of the two is artificial, but considering that the evaluation of them varies from person to person. Therefore, I anonymously handed a total of 5 ancient poems generated by each of the two models to 4 people of different language proficiency for evaluation (a Chinese teacher A, an old literature lover B, a sophomore C and a high school student D), and did my best to investigate as reasonably as possible.

The final results are shown in Table 3. It can be seen that the new method is generally better than the old method. In fact, the gap between the levels of the two can be seen in Table 2. In the basic method, there is an unreasonable way of writing the same Chinese characters in one sentence. It can be seen that although the effect of extracting themes has been weakened in this very fair comparison, it is still very effective, which has improved the overall level of poetry.

5 Conclusion and Future Work

This article proposes a simple ancient poetry generation system, which can successfully support poetry composition based on keywords. The model uses the topic extraction module to plan the content of the poems in advance, and effectively restricts each poem in the generation process through keywords, which strengthens the connection between the poems and greatly reduces the common theme drift phenomenon in the poem generation process. And proved that this method outperforms ordinary methods.

Although the current model is performing well, there are still many areas that need improvement. For example, it still does not support the user to input a paragraph and use it as the content to compose poems, and the extraction of keywords is relatively naive, only supporting the extraction of single words, and does not really identify the subject of the sentence. In addition, poetry creation is an extremely complicated process. The confrontation of verses, the degree of rhyme of rhyming feet, and the consistency of words are all factors that affect excellent poetry. A lot of basic knowledge is needed to guide the generation of poetry.

All in all, automatic poetry writing by machines is a forward-looking research, and it is also a very challenging problem that awaits more attention from researchers.

Acknowledgments and Disclosure of Funding

I would like to thank Teacher Zhang and all assistants for their guidance and help, and my classmate Chuhao Xu for his insightful discussion and suggestions.

References

[1] Zhang, X. & Lapata, M. . (2014). Chinese Poetry Generation with Recurrent Neural Networks. Conference on Empirical Methods in Natural Language Processing.

[2] Bahdanau, D. , Cho, K. , & Bengio, Y. . (2014). Neural machine translation by jointly learning to align and translate. Computer Science.

[3] Wang, Zhe & He, Wei & Wu, Hua & Wu, Haiyang & Li, Wei & Wang, Haifeng & Chen, Enhong. (2016). Chinese Poetry Generation with Planning based Neural Network.