Classical Chinese poetry generator with RNN

Tianhui Zhang

ABSTRACT

In this paper, a recursive neural network based on topic model and deep learning is proposed to make the next sentence of the poem realize the relational mapping. Firstly, the sample corpus is obtained, and the Chinese word segmentation technology and topic model method are used to build the word collection and realize the topic clustering of words. Then, the first sentence is generated by selecting the words according to the structure of the first sentence through the first sentence language model. On the basis of obtaining the first sentence, the context model is used to compress the sentence vector, and the compressed vector is fed to the recursive neural network for training. Finally, a trained neural network is used to generate poems automatically. Through machine learning of large-scale poetry data, the characteristics of classic Chinese poetry creation are integrated into the statistical probability model, thus realizing the auxiliary creation of classic Chinese poetry, providing help for the majority of classic Chinese poetry lovers, and having positive significance for the inheritance and development of classic Chinese poetry literature.

Keywords: RNN, NLP, Classic Chinese poetry

1. INTRODUCTION

The emergence of natural language processing transcends the language barrier and increases the possibilities of language research. For example, the creation of classical Chinese poetry. Classical Chinese poetry is traditional Chinese poetry written in Classical Chinese and typified by certain traditional forms, or modes; traditional genres; and connections with particular historical periods, such as the poetry of the Tang Dynasty. The existence of classical Chinese poetry is documented at least as early as the publication of the Classic of Poetry (Shijing). Various combinations of forms and genres have developed over the ages. Many or most of these poetic forms were developed by the end of the Tang Dynasty, in 907 CE.[1] It is not easy to create poetry because of its special aesthetic needs and rhyme and allusion requirements.

1.1 The basics of NLP

Natural Language Processing (NLP) is a branch of human intelligence and linguistics. As a broad subject, NLP involves many research directions such as machine translation syntactic analysis and risk retrieval.

1.1.1 The word vector

Natural language processing mainly studies language information (words, sentences, texts, etc.). It is a kind of abstract entity of dissociative cognition produced in the process of human cognition, while speech and image belong to the bottom layer's raw input signal. Voice, image data expression does not require special coding, and there are the sequence and relevance, the approximate number will be regarded as the characteristics of the approximation. As the image is made up of pixels, language is made up of words or words, language can be transformed into words or words expressed collections.

However, the numeric size of the word is very difficult to represent the meaning of words. At first, one-hot encoding was used for convenience. This word representation method is very simple and easy to implement, which solves the problem that classifiers are difficult to process attribute (Cate-gorical) data. Its disadvantages are also obvious: it is too redundant and fails to reflect the relation-

ship between words. You can see that the representations of these 10 words are all orthogonal to each other, that is, any two words are not identical to each other, and any two words are the same distance from each other. At the same time, the dimension of one-hot vector will increase sharply with the increase of crop word number. Although one-hot encoding scheme performs well in traditional tasks, dimensional disasters often occur when applied in deep learning due to the high dimension of words, so word vector representation is generally adopted in deep learning.

Word Vector, also known as Word Embedding, has no strict unified definition. Conceptually, it refers to embedding an off-dimensional space (tens of thousands of words, hundreds of thousands of words) whose dimension is the number of all words into a continuous most space (generally 128 or 256 dimensions) with a much lower dimension, and each word or phrase is mapped to a vector in the real number field.[2]

1.1.2 RNN

RNN, in short, is a Recurrent Neural Network, which is almost an essential tool in deep learning to solve NLP problems. Assuming that we now have a word vector representation for each word, how do we get the meaning of the sentences these words form? We can't analyze a single word because each word depends on the previous one, and we can't get the information of the sentence by looking at a single word. RNN can solve this problem well. It calculates the new state each time by combining the hidden state of the previous word with the current word cabinet.

An RNN architecture called LSTM is used in deep learning Structure. LSTM (Long Short Term Memory Networks) has the structure of is shown below:

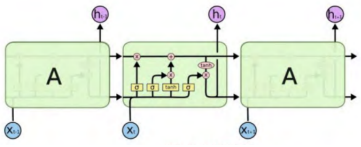

Fig. 1. LSTM structure

1.2 Characteristics of Classical Chinese poetry

Classical Chinese poetry forms are poetry forms or modes which typify the traditional Chinese poems written in Literary Chinese or Classical Chinese. Classical Chinese poetry has various characteristic forms, some attested to as early as the publication of the Classic of Poetry, dating from a traditionally, and roughly, estimated time of around 10th–7th century BC. The term "forms" refers to various formal and technical aspects applied to poems: this includes such poetic characteristics as meter (such as, line length and number of lines), rhythm (for example, presence of caesuras, end-stopping, and tone contour), and other considerations such as vocabulary and style. These forms and modes are generally, but not invariably, independent of the Classical Chinese poetry genres. Many or most of these were developed by the time of the Tang Dynasty, and the use and development of Classical Chinese poetry and genres actively continued up until the May Fourth Movement, and still continues even today in the 21st century.[3]

1.2.1 Meter

For the purpose of metrically scanning Classical Chinese verse, the basic unit corresponds to a single character, or what is considered one syllable: an optional consonant or glide (or in some versions of reconstructed Old or Middle Chinese a consonantal cluster), an obligatory vowel or vowel cluster (with or without glides), and an optional final consonant. Thus a seven-character line is identical with a seven-syllable line; and, barring the presence of compound words, which were rare in Classical Chinese compared to Modern Chinese (and even people's names would often be abbreviated to one character), then the line would also be a seven words itself. Classical Chinese tends toward a one-to-one correspondence between word, syllable, and a written character. Counting the number of syllables (which could be read as varying lengths, according to the context), together with the caesuras, or pauses within the

line, and a stop, or long pause at the end of the line, generally established the meter. The characters (or syllables) between the caesuras or end stops can be considered to be a metric foot. The caesuras tended to both be fixed depending upon the formal rules for that type of poem and to match the natural rhythm of speech based upon units of meaning spanning the characters.

1.2.2 Line length

Line length could be fixed or variable, and was based on the number of syllables/characters. In more formal poetry it tended to be fixed, and varied according to specific forms. Lines were generally combined into couplets. Lines tended to be end-stopped; and, line couplets almost always. Line length is the fundamental metrical criterion in classifying Classical Chinese poetry forms. Once the line length is determined, then the most likely division(s) of the line by caesuras is also known, since they are as a rule fixed in certain positions. Thus, specifying the line-length of a Chinese poem is equivalent to specifying both the type of feet and the number of feet per line in poetry using quantitative meter.

1.2.3 Rhythm

Rhythm was mostly a matter of tonal variation, line length, caesuras within lines, and end stopping. Variations of rhythm were subtly played off in between the various lines within a poem.

1.2.4 Rhyme

Rhyme, or rime, was important in some forms of poetry. However, it was often based on a formal and traditional schema, such as is in a Rime table or rime dictionary, and not necessarily upon actual vernacular speech. The Pingshui Yun system was the standard for poetry rhyme from Yuan to Qing Dynasty, even though it was very different from actual contemporary pronunciations. Also, generally level tones only rhymed with level tones, and non-level tones with non-level tones. The original rhymes of a poem can be difficult to detect, especially in Modern Chinese, such as Mandarin Chinese and Cantonese pronunciations (including syllable finals and tone) tend to be quite different from in the older, historical types of Chinese language, although perhaps to a lesser extent in Cantonese: either way, Classical Chinese is no longer a spoken language, and pronunciation was subject to major historical variation, as attested through linguistic studies.

2. IMPLEMENTATION

2.1 Pre-process data

The data used in this experiment are more than 50,000 original tang poems collected by Chinese poetry lovers on GitHub. On this basis, I did two things.

Traditional Chinese to simplified Chinese

The original data is in traditional Chinese. Although the poems are more rhyming and ambiguous, it may still be a little more complex for who are used to simplified Chinese.

Truncate and complement all the data into a sample length

It is not easy to combine the length of thousands of different poems into a batch, so they need to be processed into a sample length.

2.2 Training

Training parameters are as follows:

```
num_epoch = 40
batch_size = 128
lr = 1e-3
weight_decay = 1e-4
max_gen_len = 200
max_len = 125
embedding_dim = 300
hidden_dim = 256
```

.

After four days of training of more than 200 Epochs, the loss values of each epoch have gradually converged as shown in the figure.

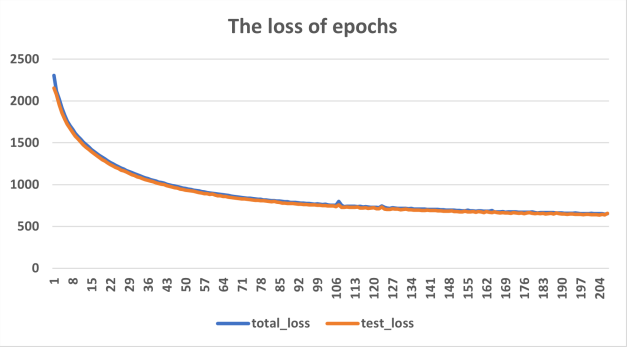

Fig. 2. Loss value of epochs

Many of the resulting poems are of high quality, and some have even learned simple pairs and rhymes. For example: 菩提本无树，不得身闲行。岂唯天地志，众界本焉知。大道无明月，尘中游圣明。万年俱是梦，万事应相闻。What's interesting is that if the poem is long enough, you can see that the mood of the poem that's being generated changes over time.

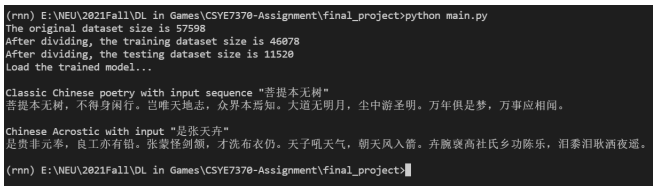

Fig. 3. Sample output of prediction

3. CONCLUTION

On the whole, the procedually-generated poetry works well, and the combination of words is quite artistic, but the poetry lacks a consistent theme, making it difficult for readers to get a main idea from a poem. This is because as poems grow in length, even THE LSTM inevitably forgets tens of words of previous input. Another prominent problem is that repetitive words often appear in generated poems, which should be avoided in traditional poetry creation, but often appear in procedually-generated poems.

REFERENCES

[1] Wikipedia contributors, "Classical chinese poetry — Wikipedia, the free encyclopedia," 2021. [Online; accessed 19-December-2021].

[2] Y. Chen, *PyTorch: Getting Started with Deep Learning*. 2018.

[3] Wikipedia contributors, "Classical chinese poetry forms — Wikipedia, the free encyclopedia," 2021. [Online; accessed 19-December-2021].