Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4658–4663
Marseille, 11–16 May 2020

Generating Major Types of Chinese Classical Poetry in a Uniformed Framework

Jinyi Hu, Maosong Sun*

Department of Computer Science and Technology, Tsinghua University, Beijing, China
Institute for Artificial Intelligence, Tsinghua University, Beijing, China
State Key Lab on Intelligent Technology and Systems, Tsinghua University, Beijing, China
hujy17@mails.tsinghua.edu.cn, sms@tsinghua.edu.cn

Abstract

Poetry generation is an interesting research topic in the field of text generation. As one of the most valuable literary and cultural heritages of China, Chinese classical poetry is very familiar and loved by Chinese people from generation to generation. It has many particular characteristics in its language structure, ranging from form, sound to meaning, thus is regarded as an ideal testing task for text generation. In this paper, we propose a GPT-2 based uniformed framework for generating major types of Chinese classical poems. We define a unified format for formulating all types of training samples by integrating form information, then present a simple form-stressed weighting method in GPT-2 to strengthen the control to the form of generated poems, with special emphasis on those forms with longer body length. Preliminary experimental results show this enhanced model can generate Chinese classical poems of major types with high quality in both form and content, validating the effectiveness of the proposed strategy. The model has been incorporated into Jiuge, the most influential Chinese classical poetry generation system developed by Tsinghua University (Guo et al., 2019).

Keywords: poetry generation, GPT-2, form control

1. Introduction

Chinese poetry is a rich treasure in Chinese traditional culture. For thousands of years, poetry is always considered as the crystallization of human wisdom and erudition by Chinese people and deeply influences the Chinese history from the mental and cultural perspective.

In general, a Chinese classical poem is a perfect combination of three aspects, i.e., form, sound, and meaning. Firstly, it must strictly obey a particular form which specifies the number of lines (i.e., sentences) in the poem and the number of characters in each line. Secondly, it must strictly obey a particular sound pattern which specifies the sound requirement for each character in every position of the poem. Lastly, it must be meaningful, i.e., with grammatical and semantic well-formedness for each line and, with thematic coherence and integrity throughout the poem. These three points form the universal principles for human poets to create Chinese classical poems.

Chinese Classical poetry can be classified into two primary categories, SHI and CI. According to the statistical data from CCPC1.0, a Chinese Classical Poetry Corpus consisting of 834,902 poems in total (We believe it is almost a full collection of Chinese Classical poems). 92.87% poems in CCPC1.0 fall into the category of SHI and 7.13% fall into the category of CI. SHI and CI can be further divided into many different types in terms of their forms. We briefly introduce the related background knowledge as follows.

1.1. SHI

The majority of SHI has a fixed number of lines and a fixed and identical number of characters for all lines. Two major forms of SHI are *Jueju* and *Lvshi* with four lines and eight lines accordingly. *Jueju* and *Lvshi* are further divided into *Wuyan Jueju* and *Qiyan Jueju* as well as *Wuyan Lvshi* and *Qiyan Lvshi* where *Wuyan* means five characters each line and *Qiyan* means seven characters. Figure 1 is a famous classical poem of *Wuyan Jueju*. In addition, *Lvshi* has a strict requirement for the two-sentence pairs composed of <the third line, the fourth line> and <the fifth line, the sixth line>: they must satisfy the requirement of *Duizhang*, this is, a strict parallel matching for both part of speech and sense of every character in two lines. This obviously increases the difficulty of poem composition.

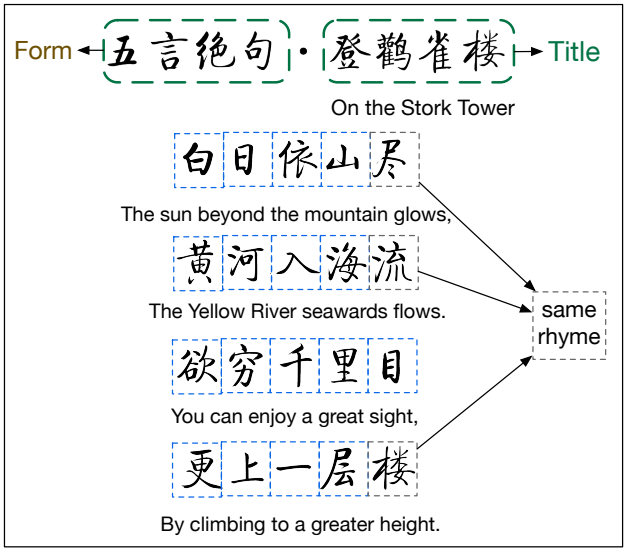


Figure 1: An example of SHI with *Wuyan Jueju* as its form. The array of small boxes, usually each surrounds a Chinese character, illustrates the form requirement in the number of lines and the number of characters per line for a poem. A full correspondence between character and box, no more and no less, indicates this basic form requirement is satisfied by the given poem.

* Corresponding author

According to CCPC1.0, *Wuyan Jueju*, *Qiyan Jueju*, *Wuyan Lvshi*, and *Qiyan Lvshi* constitute 67.96% of SHI, with 4.26%, 22.57%, 15.99%, and 25.14% respectively.

1.2. CI

CI is another primary type of Chinese poetry. In contrast to SHI, CI has nearly one thousand forms. Each form of CI (it is called Cipai scholarly) is defined by a fixed number of lines for the poem and, a fixed number of characters for a particular line which usually varies for different lines. The above settings for different Cipai are very distinct, for instance, the Cipai of *Busuanzi* contains 8 lines and 44 characters, as shown in Figure 2, whereas the Cipai of *Manjianghong* contains 22 lines and 94 characters. The high diversity regarding the forms of CI further significantly increases the difficulty of poem composition.

We observe the statistical distribution of all the forms (Cipai) of CI over CCPC1.0. It roughly follows Zipf's law (Zipf, 1949). There exists a long tail in the distribution where a lot of Cipai only has a few instances which are far less enough for a computational model (algorithm) to learn its forms. So we choose the top frequent 121 forms of CI, constituting 80% of CCPC1.0, as the focus for CI in this research.

As can be seen from the above analysis, the greatest challenge for machine generation of Chinese classical poems lies in how to make machine capable of following the universal principles underlying the writing of Chinese classical poems. The to-date research cannot deal with this challenge well. Most of the work so far mainly targeted at automatic generation of *Jueju* (including *Wuyan Jueju* and *Qiyan Jueju*), for an obvious reason that it is much easier for an algorithm to handle the requirements of form, thematic coherence and integrity in the scenario of four lines than that in the scenario of *Lvshi* with eight lines, let alone much more complicated scenarios, i.e., CI, are taken into account. In fact, the research on the automatic generation of CI is just at the very beginning stage.

In this paper, we propose a uniformed computational framework that tries to generate major types of Chinese classical poems with two major forms of SHI, *Jueju*, and *Lvshi*, as well as 121 major forms (Cipai) of CI using a single model. Preliminary experimental results validate the effectiveness of the proposed framework. The implemented model has been incorporated into Jiuge (Guo et al., 2019), the most influential Chinese classical poetry generation system developed by Tsinghua University (refer to `http://jiuge.thunlp.cn/`).

2. Related Work

With the development of deep learning, the mainstream of poem generation research has been shifted from traditional statistical models to neural network methods in recent years. Most existing works are based on the Encoder-Decoder architecture (Sutskever et al., 2014). In Chinese classical poetry generation, Yan et al. (2013) proposed a model using the Encoder-Decoder architecture and Wang et al. (2016) further used attention-based sequence-to-sequence model.

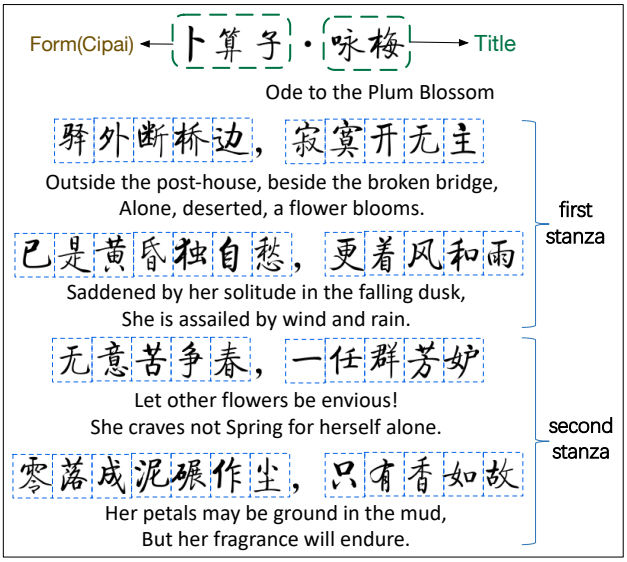


Figure 2: An example of CI with the form(Cipai) *Busuanzi*. In contrast to the case of SHI in Figure 1, the array of small boxes here shows the predefined number of characters per line of CI tends to be variable.

The key factor in designing the model architecture is how to treat the generated context so far in the process of generating a poem. The input to the encoder could be as short as a single poetic line or all the previously generated lines (whole history). Theoretically, considering the whole history is more appropriate for keeping the thematic coherence and integrity of the generated poem than considering the short history, at the expense that may hurt the fluency of the generated sentences due to the data sparseness problem possibly caused by the more sophisticated model.

Thus we have two basic ways to figure out the history. One is to consider the whole history. Zhang and Lapata (2014) first introduced the neural network method into poetry generation by proposing the so-called incremental Recurrent Neural Network, where every sentence (line) is embedded into a sentence vector by a Convolutional Sentence Model and then all are packed into a history vector. Yi et al. (2018b) presented a working memory mechanism in LSTM, designing three kinds of memory to address the whole history. Another is to select part of history. Yi et al. (2018a) observed that considering the full context may not lead to good performance in LSTM, and proposed salient clue mechanism where only salient characters in partial history are under consideration.

The Transformer (Vaswani et al., 2017) architecture and other models based on this, including GPT (Radford et al., 2018), Bert (Devlin et al., 2019), show much better results in various NLP tasks. Transformer utilizes the self-attention mechanism in which any pair of tokens in the sequence can attend to each other, making it possible to generate much longer SHI or CI while keeping the coherence throughout the poem.

Liao et al. (2019) applied GPT to Chinese classical po-

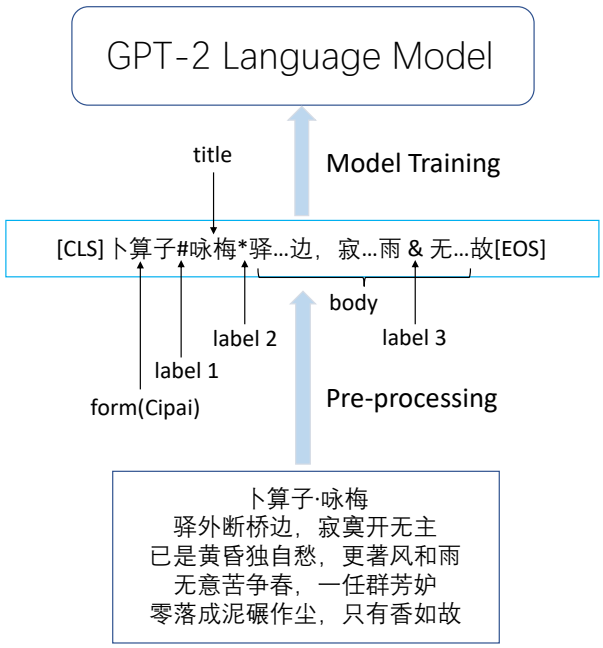


Figure 3: Format pre-processing of poem samples for training.

etry generation. They pre-trained the model on a Chinese news corpus with 235M sentences and then fine-tuning the model on Chinese poem corpus with 250,000 *Jueju* and *Lvshi*, 20,000 CIs, 700,000 pairs of couplets. A key point is they defined a unified format to formulate different types of training samples, as [*form*, *identifier 1*, *theme*, *identifier 2*, *body*], where "*body*" accommodates the full content of an SHI, CI, or couplet in corresponding "*form*" with "*theme*" as its title. Experiments demonstrated GPT-based poem generation gained promising performance, meanwhile still faced some limitations, for instance, only 70% of the generated CIs for the Cipai *Shuidiaogetou*, a sort of CI with quite long body, are correct in form.

Regarding this, we think the work of Liao et al. (2019) could be improved in the following three respects. First, there is a large improving room for better fitting the form requirement of CI in the process of generation, especially for those with relatively long body length. Second, their formulation format for training samples can be supplemented, for example, the stanza structure of CI is missing. Third, using contemporary Chinese news corpus to pre-train the model may not be necessary, owing to distinctive differences in both meaning and form between contemporary Chinese and Chinese classical poetry language.

For the above considerations, we give up the pre-training on the news corpus and add a separation label to indicate the stanza structure of CI. Then we make use of GPT-2 to train the model. Furthermore, we propose a form-stressed weighting method in GPT-2 to strengthen the control in particular to the form of CI.

3. Model

3.1. Pre-processing

We present a unified format for formulating all types of training samples of SHI and CI by extending the format given in Liao et al. (2019). First, we change various punctuations between lines into the comma ',', serving as a uniform separation label between two lines. Second, we utilize three separation labels, $[label_1]$ and $[label_2]$ to separate between form, title, and body of the poem respectively, and $[label_3]$ to separate two stanzas of CI if needed. Third, we enclose $[EOS]$ at the end of the body. Thus, the format for SHI is as follows:

$$[CLS]form[label_1]title[label_2]body[EOS]$$
$$body : line_1, line_2, ..., line_n$$

where n is the number of lines in the poem.

The format of CI will be enriched with $[label_3]$ if it has two stanzas in the body:

$$[CLS]form[label_1]title[label_2]body[EOS]$$
$$body : stanza_1[label_3]stanza_2$$
$$stanza_1 : line_1, line_2, ..., line_m$$
$$stanza_2 : line_{m+1}, line_{m+2}, ..., line_n$$

Here, $[label_1]$, $[label_2]$ and $[label_3]$ are set as '#', '$*$' and '&'.

After pre-processing, all the formatted poem samples will be sent to the poetry generation model for training, as illustrated in Figure 3.

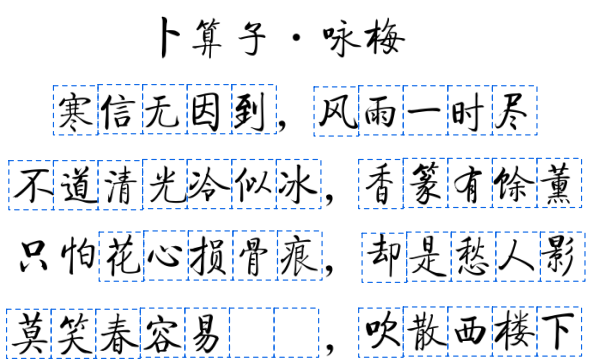
3.2. Basic Model

We leverage the Transformer-based GPT-2, which is often used to train a robust language model, as the basic model of poetry generation. Compared to previous neural network-based language models such as RNN and LSTM, it is reported that GPT-2 exhibits good performance in the quality of generated texts given quite a long history (Radford et al., 2019). To weaken the so-called degeneration problem in generation and increase the diversity of generated texts, we use the top-k stochastic sampling strategy (Fan et al., 2018) (k is set as 15 in our experiment) to choose the next tokens to generate. In addition, our poetry generation model takes the Chinese character rather than the word as a basic linguistic unit, so word segmentation is not needed.

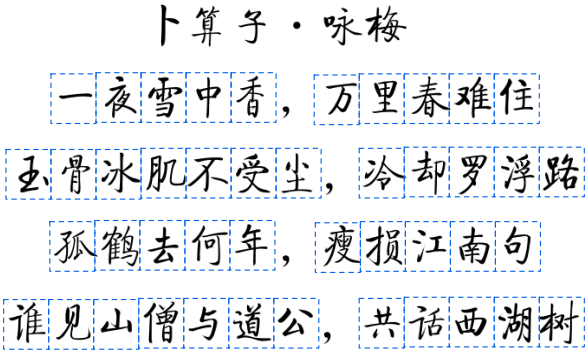
With this naive GPT-2 model, we see from the experimental results that the generated poems appear pretty good in both meaning and sound(including rhyme), though if being observed carefully, there still exist some in-depth problems in sentence fluency and thematic coherence of the whole poem which are uneasy to solve. As for form, the model can perform well in generating *Jueju* and *Lvshi* of SHI whereas rather poorly in generating various Cipai of CI, with quite high form errors. Figure 4(a) is an example of a generated CI by this model, under Cipai of *Busuanzi*, where two characters are mistakenly missing which obviously violates the form requirement.

3.3. Enhanced Model

In the basic model, the loss function for training with respect to the ith token in the text is conventionally defined



(a) A generated poem by the basic model: two obvious errors in form.



(b) A generated poem by the enhanced model, with the same inputting title (or theme) under the same Cipai as in (a): full correctness in form.e

Figure 4: Comparison of two generated poems by the basic model and the enhanced model.

as the cross-entropy:

$$Loss(x,i) = -\log \frac{\exp x[i]}{\sum_j \exp x[j]} = -x[i] + \log \sum_j \exp\left(x[j]\right)$$

where $x[i]$ is the vector of ith token, j is over all possible token types.

To address the form problem, we simply add a weighting factor into the loss function with particular stress on the aforementioned three types of form-related tokens, i.e., the line separation label ',', the stanza separation label '&', and $[EOF]$, as in:

$$Loss(x,i) = weight[i]\big(-x[i] + \log \sum_j \exp\left(x[j]\right)\big)$$

where $weight[i]$ is set as 1 for any Chinese character, 2 for ',' and '&', and 3 for $[EOF]$.

This simple method (we thus call it the form-stressed weighting method) enhances the model's capability to form control quite significantly. Figure 4(b) shows an example that contrasts the case in Figure 4(a).

4. Experiment

4.1. Experiment Setup

We implement the GPT-2 model based on the transformers library (Wolf et al., 2019). The model configuration is 8 attention heads per layer, 8 layers, 512 embedding dimensions, and 1024 feed-forward layer dimensions. We employ the OpenAIAdam optimizer and train the model with 400,000 steps in total on 4 NVIDIA 1080Ti GPUs. The characters with frequency less than 3 in CCPC1.0 are treated as UNK and a vocabulary with 11259 tokens (characters) is finally built up.

4.2. Performance Comparison of the Two Models in Form

For *Jueju* and *Lvshi* of SHI, because of their simplicity in form, the two models hardly make form errors. We generate 500 poems for each type using the two models accordingly. All of these poems are in the right form. This demonstrates that both models are all very powerful in generating *Jueju* and *Lvshi* with almost perfect performance in form.

For CI, we select 6 Cipais, with the body length varying from 33 to 114 characters and with relatively sufficient training samples in CPCC, as our observation target. We generate 300 poems with the two models accordingly. Table 1 summarizes the correct rates of the two models under these 6 Cipais (a generated poem is considered to be correct in form if and only if its form fully matches the expected form). As can be seen, a tendency is the longer the body of CI, the worse the performance of the two models in form and, the more significant the gain in the form correct rate for the enhanced model (an extreme is in the case of *Qinyuanchun* where the correct rate is raised from 12.0% to 55.0%).

4.3. Effect of the Stanza Separation

The preliminary observation on the generated poems suggests that the inclusion of the stanza separation into the unified format of training samples is beneficial in some degree for meeting the form requirement. For instance, we input the same title to the enhanced model and to a model trained under the same condition except without the stanza separation, asking them to generate a number of CIs with Cipai of *Busuanzi*, a task similar to that in Figure 4. We find that about 20% of CIs generated by the latter suffer from some errors in form, as illustrated in Figure 5, meanwhile all the CIs generated by the former ideally match the expected form.

4.4. Case Observation

According to our observation, the enhanced model is likely to generate poems with both high quality and diversity. We present two examples generated by the model and give some comments on the meaning of each poem.

七律·远望
江上微茫一叶舟，天涯芳草满汀洲
数声渔唱隔船过，几点人家落帆游
春色不从莺语到，夕阳空度客心愁
何时重向长桥饮，同泛溪光共白头

Cipai	Length of Body	Number of Training Samples	Correct Rate in Form of Basic model	Correct Rate in Form of Enhanced model
Rumengling	33	682	86.0%	90.0%
Jianzimulanhua	44	866	87.3%	95.7%
Qingpingyue	46	1236	84.0%	96.0%
Dielianhua	60	1578	89.7%	91.3%
Manjianghong	93	1398	42.1%	83.3%
Qinyuanchun	114	1061	12.0%	55.0%

Table 1: Comparison between two models on the control to the form of CI.

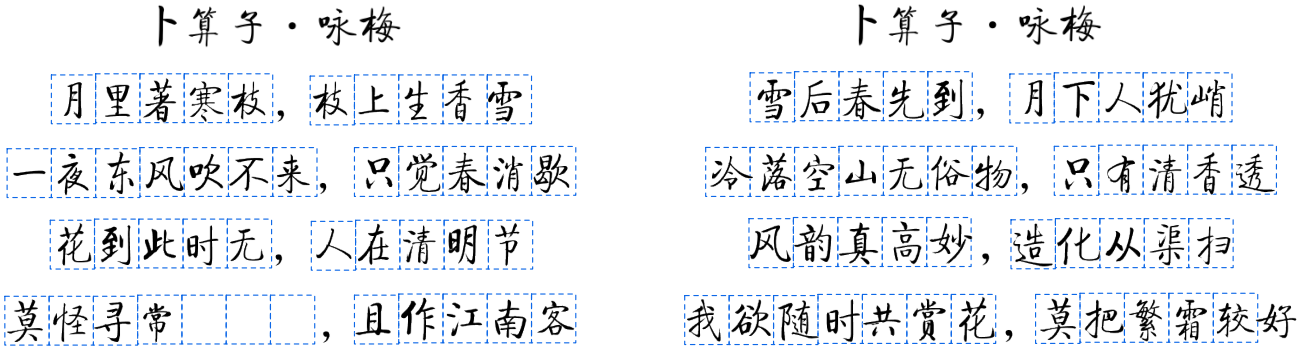


Figure 5: Two example poems generated by the model without considering the stanza separation. Both have errors in form. Refer to Figure 4(b) for comparison.

The example above is a *Qiyan Lvshi*. The title of this poem means "look far around". In this poem, the first four lines depict a view seen from the river bank-misty and rolling waters, a drifting boat, lush vanillas, melodies from passing boats and cottages on the bank, creating a tranquil and halcyon atmosphere. However, the poet is still overcome by solitude and nostalgia because of the lonely trip, which is vividly revealed in the second four sentences. The poem adopts a typical semantic structure of *Qiyan Lvshi* with its first-half delineating a view and then conveying the poet's feeling in the second-half (the contrast between the view and the feeling is one of the appreciated artistic methods in Chinese classical poems). In addition, for *Lvshi*, the pairs of <the third line, the fourth line> and <the fifth line, the sixth line> must satisfy the requirement of *Duizhang*, a correspondence in both part-of-speech(POS) and word sense between two parallel lines. This point is perfectly reflected in the generated poem, as shown in Table 2.

满江红·塞外
风急秋空，天欲暮，黄云飞处。
人不见，沙堤野戍，乱鸦啼苦。
万里胡笳吹雁断，三更羌笛愁如许。
甚关河、征妇泪痕多，无行路。
青狼火，荒烟树。
白露草，残阳度。
但寒山远近，故乡千古。
一角斜晖归梦绕，满江红叶西陵去。
待明年，又到汉家城，重回顾。

The example above is a CI in the form of *Manjianghong* and the title means "beyond the Great Wall". It vividly depicts a typical view of the Northwestern China howling wind, clouds of dust, crying crows and lugubrious sound of flutes. The poem is saturated with nostalgia, solitude and desolate feelings of life, which is not only embodied in the bleak scenery but also overtly revealed in the last three sentences. The combination of visual and audio feelings and of reality and imagination is tactfully employed in the poem and makes it even more impressive and resonating.

5. Conclusion and Future Works

In this paper, we propose a GPT-2 based uniformed framework for generating major types of Chinese classical poems, including SHI and CI. To this end, we at first define a unified format for formulating all types of training samples by integrating more detailed form information, then present a simple form-stressed weighting method in GPT-2 to strengthen the control to the form of CI. Preliminary experiments validate the effectiveness of our method. Nevertheless, we also find that enabling GPT-2 to have a strong capability in form manipulation for the generated texts remains a difficult challenge, particularly for those forms with longer body length and fewer training samples. We plan to figure out a more sophisticated way to make the model better learn the form structure and hope to enrich the general GPT-2 from this special perspective.

	数	声	渔歌	隔	船	过
POS	NUMBERAL	N	N	V	N	V
Word	several	sound	fishing song	next-door	boat	pass
Meaning	The next-door boat is passing by, with several sounds of a fishing song					
	几	点	人家	落	帆	游
POS	NUMERAL	N	N	V	N	V
Word	a few	point	home	fall	sail	move
Meaning	A few of far-away boats which carry on the whole family, still wander with the falling sails, looking like small pieces of points					

Table 2: Illustration of *Duizhang*.

6. Acknowledgements

We would like to thank Zhipeng Guo, Xiaoyuan Yi, Xinran Gu and anonymous reviewers for their insightful comments. This work is supported by the project Text Analysis and Studies on Chinese Classical Literary Canons with Big Data Technology under grant number 18ZDA238 from the Major Program of the National Social Science Fund of China. Hu is also supported by the Initiative Scientific Research Program and Academic Training Program of the Department of Computer Science and Technology, Tsinghua University.

7. References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Guo, Z., Yi, X., Sun, M., Li, W., Yang, C., Liang, J., Chen, H., Zhang, Y., and Li, R. (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.

Liao, Y., Wang, Y., Liu, Q., and Jiang, X. (2019). Gpt-based generation for classical chinese poetry. *arXiv preprint arXiv:1907.00151*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, D. Z., He, W., Wu, H., Wu, H., Li, W., Wang, H., and Chen, E. (2016). Chinese poetry generation with planning based neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1051–1060.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yan, R., Jiang, H., Lapata, M., Lin, S.-D., Lv, X., and Li, X. (2013). I, poet: automatic chinese poetry composition through a generative summarization framework under constrained optimization. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Yi, X., Li, R., and Sun, M. (2018a). Chinese poetry generation with a salient-clue mechanism. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 241–250.

Yi, X., Sun, M., Li, R., and Yang, Z. (2018b). Chinese poetry generation with a working memory model. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4553–4559.

Zhang, X. and Lapata, M. (2014). Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, (Mass.): Addison-Wesley.