Sentiment-Controllable Chinese Poetry Generation

Huimin Chen[1,2,3,4*], **Xiaoyuan Yi**[1,2,3,4*], **Maosong Sun**[1,2,3,4†], **Wenhao Li**[1,2,3,4], **Cheng Yang**[1,2,3,4] and **Zhipeng Guo**[1,2,3,4]
[1]Department of Computer Science and Technology, Tsinghua University
[2]Institute for Artificial Intelligence, Tsinghua University
[3]Beijing National Research Center for Information Science and Technology, Tsinghua University
[4]State Key Lab on Intelligent Technology and Systems, Tsinghua University
{chm15, yi-xy16, cheng-ya14, liwh16, gzp17}@mails.tsinghua.edu.cn, sms@tsinghua.edu.cn

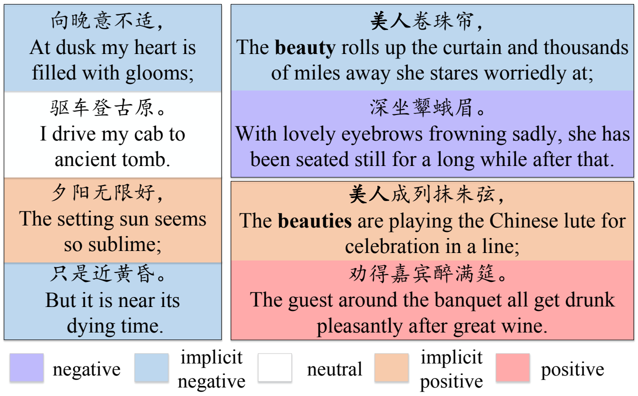
Abstract

Expressing diverse sentiments is one of the main purposes of human poetry creation. Existing Chinese poetry generation models have made great progress in poetry quality, but they all neglected to endow generated poems with specific sentiments. Such defect leads to strong sentiment collapse or bias and thus hurts the diversity and semantics of generated poems. Meanwhile, there are few sentimental Chinese poetry resources for studying. To address this problem, we first collect a manually-labelled sentimental poetry corpus with fine-grained sentiment labels. Then we propose a novel semi-supervised conditional Variational Auto-Encoder model for sentiment-controllable poetry generation. Besides, since poetry is discourse-level text where the polarity and intensity of sentiment could transfer among lines, we incorporate a temporal module to capture sentiment transition patterns among different lines. Experimental results show our model can control the sentiment of not only a whole poem but also each line, and improve the poetry diversity against the state-of-the-art models without losing quality.

1 Introduction

Poetry is an important literary genre which has attracted people and influenced human society with its exquisite expression, rich content and diverse sentiments for thousands of years. Recently, as a classical task in the NLP field, automatic poetry generation has come to the foreground again. Besides the goal towards increasing computer creativity and understanding human writing mechanism, poetry generation is also helpful for applications in areas such as entertainments, advertisement, and education.

For human beings, in addition to recording interesting events and making comments, expressing diverse sentiments is another main purpose of creating poetry (as well as other literary genres) [Morris-Jones, 1962]. For example, expressing the sadness of ageing and the happiness of feasting (Figure 1). Controlling the sentiment of created poems is an essential ability of human poets and also a basic user demand for automatic poetry generation systems.



向晚意不适，
At dusk my heart is filled with glooms;
驱车登古原。
I drive my cab to ancient tomb.
夕阳无限好，
The setting sun seems so sublime;
只是近黄昏。
But it is near its dying time.

美人卷珠帘，
The **beauty** rolls up the curtain and thousands of miles away she stares worriedly at;
深坐颦蛾眉。
With lovely eyebrows frowning sadly, she has been seated still for a long while after that.
美人成列抹朱弦，
The **beauties** are playing the Chinese lute for celebration in a line;
劝得嘉宾醉满筵。
The guest around the banquet all get drunk pleasantly after great wine.

Figure 1: Left: a human-created poem. The whole poem expresses negative sentiment but there is some intensity transition across different lines. Right: two human-created sentences (each with two lines) which contain the same keyword, "beauty", but express different kinds of sentiment.

Though recent neural poetry generation models have achieved significant improvements in different aspects of poetry quality, such as fluency [Zhang and Lapata, 2014] and coherence [Wang *et al.*, 2016b; Yi *et al.*, 2018b], they all neglected to generate sentiment-controllable poetry. Such defect causes a strong sentiment collapse (the generated poems tend to be neutral and meaningless descriptions) or sentiment bias (the generated poems tend to express negative sentiment) which further hurts the semantics and diversity.

To address this problem, we concentrate on automatic sentiment-controllable poetry generation. Due to the lack of off-the-shelf sentimental poetry corpus, we first build a fine-grained manually-labelled sentimental Chinese corpus[1]. Each poem is annotated with sentiments of not only the whole poem but also each line because there are varying granularities (e.g., polarities and intensities) in poetry sentiment expression, and the sentiment of each line could be different under certain holistic sentiment [Janowitz, 1973].

[*] Indicates equal contribution
[†] Corresponding author

[1] Details of the corpus are given in Section 3.

Since labelled data is still a small portion of the whole corpus, we can't take supervised methods designed for related tasks, e.g., poetry style transfer [Wei *et al.*, 2018]. Instead, we propose to adopt a semi-supervised Variational AutoEncoder (VAE) [Kingma *et al.*, 2014], which utilizes the labelled data more efficiently and has been widely used for image generation, to generate sentimental poetry.

Different from previous VAE models which learn a context-based latent variable [Yang *et al.*, 2018b] or suppose the independence of latent variable and the required attributes (e.g., sentiment) [Hu *et al.*, 2017], we make the latent space conditioned on both sentiment and content to capture generalized sentiment-related semantics, because the sentiment is coupled with semantics especially for poetry [Chari, 1976]. Concretely, we extend the semi-supervised version of VAE [Kingma *et al.*, 2014] to the conditional version and deduce a different lower bound for our task to capture generalized sentiment-related semantics more efficiently. Besides, since poetry is a kind of discourse-level text, under the holistic sentiment of a whole poem, the sentiment of each line could have some changes and flexibility, as shown in Figure 1. Therefore, we incorporate a temporal sequence module to learn sentiment transition patterns among different lines. Taking a user keyword as input, our model can generate diverse poems under the control of discourse-level or line-level sentiments. Our model can also predict an appropriate sentiment for the whole poem and infer a sentiment transition pattern across lines when the sentimental labels are not provided.

In summary, the contribution of this paper is four-fold:

(1) To the best of our knowledge, we are the first to endow the poetry generator with the ability to express specific sentiments, which also improves the semantics and diversity of generated poems.

(2) Different from previous works which conduct sentiment transfer only for a single sentence, we utilize a temporal sequence module to control discourse-level sentiment.

(3) We build a fine-grained sentimental Chinese poetry corpus, with sentiment labels for a whole poem and each line.

(4) Experimented on Chinese poetry, our model can control sentiments of a whole poem and lines without losing quality.

2 Related Work

Automatic poetry generation is a classic task in computer writing. The related works in recent decades could be categorized into three main stages. On the first stage, models are based on rules and templates, e.g., [Gervás, 2001], which are the first attempts. On the second stage, statistical machine learning methods point out a new possible direction for this task. Different algorithms are utilized, such as Genetic algorithms [Manurung, 2003; Levy, 2001] and Statistical Machine Translation approaches [He *et al.*, 2012].

In the past several years, researches have stepped into the third stage where powerful neural networks bring new energy to this task. Recurrent Neural Network (RNN) is first used to generate Chinese quatrains [Zhang and Lapata, 2014]. After that, more effective sequence-to-sequence models with attention mechanism [Bahdanau *et al.*, 2015] are also adopted to generate poetry [Wang *et al.*, 2016a]. Aiming at improving different criteria of poetry quality, researchers design various structures. To improve context coherence, Wang et al. [2016b] propose a Planning model, which plans sub-keywords in advance for each line; Yi et al. [2018b] develop a working memory model to maintain the context in dynamic internal memory. Besides, keywords extension [Ghazvininejad *et al.*, 2016] and static external memory [Zhang *et al.*, 2017] are used to purse better meaningfulness.

Beyond these primary criteria, focusing on improving the diversity, a higher-level requirement for generated poetry, Yang et al. [2018a] use mutual information to achieve unsupervised style disentanglement and Yi et al. [2018a] use reinforcement learning to optimize evaluation criteria directly.

Despite notable progress, these models ignore an essential point for poetry creation, the sentiment, which results in the collapse or bias in sentiment expression and hence hurts the diversity and semantics of generated poems. We consider VAE as a feasible method, which has shown great promise in text generation tasks such as dialogue generation [Zhao *et al.*, 2017] and poetry generation [Yang *et al.*, 2018b]. Though inspired by these works, our motivation and proposed models differ from them by a large margin. We are the first effort at sentimental poetry generation with a semi-supervised sentiment-conditioned VAE, which makes latent space conditioned on the sentiments, instead of learning a context-based or attribute-independent latent variable.

Our work is also related to the task of text generation with controllable sentiments [Hu *et al.*, 2017; Cagan *et al.*, 2017; Wang and Wan, 2018]. Different from them, our work focuses on generating sentiment-controllable poems in discourse level and involves a temporal sequence module to capture sentiment transition across lines, while they focus on the controllable generation in sentence level.

3 Fine-grained Sentimental Poetry Corpus

Controlling the sentiments of poems is necessary for automatic poetry generation systems as mentioned in Section 1. However, to the best of our knowledge, there is no off-the-shelf Chinese poetry corpus with sentiment labels, thus we build a manually-labelled *Fine-grained Sentiment Poetry Corpus* including 5,000 Chinese quatrains.

We collect 151,835 unlabelled Chinese quatrains, as quatrain is the dominant genre of Chinese poetry. First, we use a distant supervision method to divide the unlabelled data into 5 classes in terms of the number of sentimental seed words contained in each poem. Then from each class, we select 1,000 poems, with higher priority on those written by famous poets, for manual annotation. As poetry is discourse-level text with fine-grained sentiments as discussed in Section 1, we annotate each poem and each line into 5 classes, namely **negative**, **implicit negative**, **neutral**, **implicit positive** and **positive**. To ensure the quality of labelling, each poem is annotated by at least two annotators, who are members of a poetry association or major in Chinese literature. If the two annotators have disagreements on the poem, it will be assigned to a senior annotator who will decide the final label referring to the two annotations. Statistics of this corpus are reported in Table 1.

Figure 2 shows the normalized label distributions across

Granularity	#Neg.	#Implicit Neg.	#Neutral	#Implicit Pos.	#Pos.
Whole Poem	289	1,467	1,328	1,561	355
Line1	143	1,023	2,337	1,310	187
Line2	268	1,138	1,936	1,423	235
Line3	212	1,107	2,320	1,083	278
Line4	315	1,317	1,650	1,357	361

Table 1: Statistics of labelled sentimental Chinese poems. Neg. is the abbreviation of negative and Pos. is the abbreviation of positive.

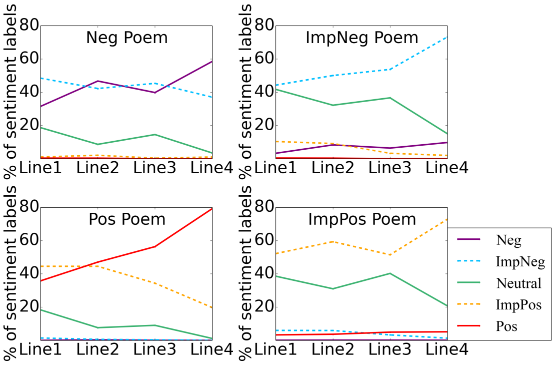


Figure 2: Normalized distributions of sentiment labels across different lines in each holistic sentiment of the whole poem. Neg: negative sentiment; Pos: positive sentiment; Imp: implicit.

different lines. We can find under a certain holistic sentiment of the whole poem, the sentiments across different lines are diverse. Furthermore, the sentiment of the last line is often consistent with the sentiment of the whole poem, while the sentiment of the first line tends to be neutral or implicit. Therefore, it is necessary to model the sentiment transition patterns across different lines.

4 Method

In this section, we introduce our semi-supervised Sentiment-Controllable Poetry Generation model (**SCPG**). Before presenting the details, we first formalize our task.

Define x as a poem with n lines $\{x_1, x_2, ..., x_n\}$ (abbreviated as $x_{1:n}$), w as a keyword which represents the main topic of x, y as the holistic sentiment of x, $\{y_1, y_2, ..., y_n\}$ (abbreviated as $y_{1:n}$) as the sentiments expressed in each line, $p_l(x, w, y, y_{1:n})$ and $p_u(x, w)$ as the empirical distributions over labelled and unlabelled datasets respectively. With the keyword w, we aim to generate poems not only holding the holistic sentiment y for the whole poem x but also expressing the sentiment y_i for each line x_i. In the following parts we will progressively present different settings of our model.

4.1 Holistic Sentiment Control Module

We first introduce the holistic sentiment control module, which adopts a semi-supervised sentiment-conditioned variational autoencoder. As shown in Figure 3 (a), our goal is to learn the conditional joint distribution $p(x, y, z|w)$, where z is the latent variable. We decompose it as $p(x, y, z|w) = p(x|z, w, y)p(z|y, w)p(y|w)$, which describes the generation process: the model infers (if not provided by the user) an appropriate sentiment of the whole poem by the keyword, then samples a z conditioned on the keyword w (required content) and the label y (required sentiment), finally generates the poem x with them. Note that during this process, we don't suppose the independence of z and y as [Kingma *et al.*, 2014], instead, we directly draw z from the sentiment and keyword since the sentiment is coupled with semantics in poetry as discussed in Section 1.

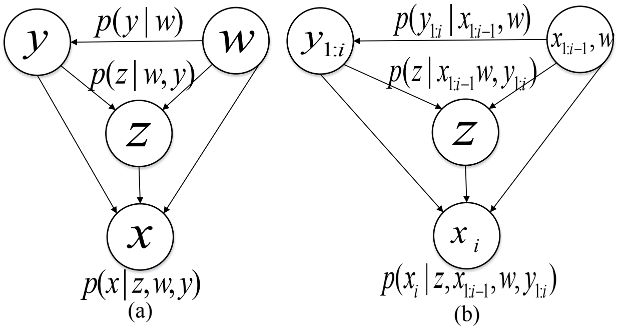


Figure 3: The graphical illustrations of (a) the holistic sentiment control module and (b) the temporal sentiment control module when generating line x_i.

As our model is semi-supervised, we consider two cases.

For the labelled data, inspired by [Kingma *et al.*, 2014], to maximize the distribution $p(x, y|w)$, we involve z and derive the lower bound as:

$$\begin{aligned}\log\, p(x, y|w) \geq\ & \mathbb{E}_{q(z|x,w,y)}[\log\, p(x|z, w, y)] \\ & - KL[q(z|x, w, y)||p(z|w, y)] + \log\, p(y|w) \\ =\ & -\mathcal{L}(x, y, w)\end{aligned} \tag{1}$$

where $q(z|x, w, y)$ and $p(z|w, y)$ are the estimations of the posterior and prior distributions respectively. By optimizing Eq. (1), we also train a classifier $p(y|w)$ to help predict a holistic sentiment if the user doesn't provide any label.

For unlabelled data, by treating the unobserved label y as another latent variable, we have:

$$\begin{aligned}\log p(x|w) =\ & \iint q(y, z|x, w) \log\, p(x|w) dydz \\ \geq\ & \mathbb{E}_{q(y|x,w)}[-\mathcal{L}(x, y, w) - \log\, q(y|x, w)] \\ =\ & -\mathcal{U}(x, w).\end{aligned} \tag{2}$$

Then the total semi-supervised loss is:

$$\mathcal{S}_1 = \mathbb{E}_{p_l(x,w,y)}[\mathcal{L}(x, y, w) - \log\, q(y|x, w)] + \mathbb{E}_{p_u(x,w)}[\mathcal{U}(x, w)], \tag{3}$$

where another classifier $q(y|x, w)$ is simultaneously trained to sample sentiments for unlabelled poems.

Concretely, we get the representation of x by feeding the poem into a bidirectional GRU [Cho *et al.*, 2014]. The classifiers $p(y|w)$ and $q(y|x, w)$ are implemented with MLPs. As previous work [Yang *et al.*, 2018b] did, we assume latent variable z follows the isotropic Gaussian distribution, *i.e.*, $p(z|w, y) \sim \mathcal{N}(\mu_{prior}, \sigma_{prior}\boldsymbol{I})$ and $q(z|x, w, y) \sim \mathcal{N}(\mu_{post}, \sigma_{post}\boldsymbol{I})$. These mean values μ and standard deviations σ are computed by MLPs. For $p(x|z, y, w)$, we use a GRU decoder and set the initial decoder state to $s_0 = f(z, y, w)$, where f is a non-linear layer, to involve the sentiment and keyword. When training, z is sampled from $q(z|x, w, y)$ and when testing z is sampled from $p(z|w, y)$.

4.2 Temporal Sentiment Control Module

Since poetry is a kind of discourse-level text as mentioned in Section 1, under a certain holistic sentiment of a whole poem, the sentiments could vary across different lines. Therefore, we propose a temporal sentiment control module (Figure 3 (b)) to capture such sentiment transition patterns.

In a poem, there are two sequences: the content sequence $x_{1:n}$ and the sentiment sequence $y_{1:n}$. Since the sentiment is coupled with the content (semantics), the sentiment of a line will be influenced by both the content and sentiment of previous lines and vice versa. Therefore, we model these two sequences as an interactive temporal sequence module. In detail, we learn a joint distribution and factorize it as:

$$\log\, p(x_{1:n}, y_{1:n}|w) = \log\, p(x_1, y_1|w) + \sum_{i=2}^{n} \log\, p(x_i, y_i|x_{1:i-1}, y_{1:i-1}, w). \tag{4}$$

For labelled data, when generating line x_i, we follow the holistic module by simply treating $x_{1:i-1}, y_{1:i-1}, w$ as a condition and replacing w in Eq. (1) with it. Then we can get the lower bound $\log\, p(x_i, y_i|c) \geq -\mathcal{L}(x_{1:i}, y_{1:i}, w)$ and directly optimize each factor, where c denotes $x_{1:i-1}, y_{1:i-1}, w$.

For unlabelled data, when generating line x_i, we model the unseen $y_1, ..., y_i$ as latent variables and similar to Eq. (2) we have:

$$\begin{aligned}\log\, p(x_i|x_{1:i-1}, w) &\geq \mathbb{E}_{q(y_{1:i}|x_{1:i},w)}[-\mathcal{L}(x_{1:i}, y_{1:i}, w) \\ &\quad - \log\, q(y_{1:i}|x_{1:i}, w)] \\ &= -\mathcal{U}(x_{1:i}, w).\end{aligned} \tag{5}$$

However, it's too expensive to directly compute the expectation in Eq. (5) because of the exponential number of possible sentiment sequence $y_{1:i}$. Instead, we use the Monte Carlo method to estimate the expectation:

$$-\mathcal{U}(x_{1:i}, w) = \frac{1}{M}\sum_{k=1}^{M}[-\mathcal{L}(x_{1:i}, y_{1:i}, w) - \log\, q(y_{1:i}|x_{1:i}, w)], \tag{6}$$

$$y_{1:i} \sim q(y_{1:i}|x_{1:i}, w), \tag{7}$$

where $q(y_{1:i}|x_{1:i}, w)$ can be factorized as:

$$q(y_1|x_{1:i}, w)q(y_2|x_{1:i}, w, y_1)\ldots q(y_i|x_{1:i}, w, y_{1:i-1}). \tag{8}$$

Since we take a line-to-line generation process, we further assume that future content has no influence on current or past sentiment, *i.e.*, $x_j (j > i)$ is independent with y_i. Then we have the sampling process as $y_1 \sim q(y_1|x_1, w)$, ..., $y_i \sim q(y_i|x_{1:i}, w, y_{1:i-1})$, which shows we need to build a time sequence classifier for predicting each y_i. In detail, we use two RNNs to model the content sequence and the sentiment sequence respectively for predicting y_i:

$$c_i = h_1(c_{i-1}, x_i), c_0 = f_1(w) \tag{9}$$

$$m_i = h_2(m_{i-1}, c_{i-1}, y_{i-1}), m_0 = f_2(w), y_0 = y \tag{10}$$

$$q(y_i|x_{1:i}, w, y_{1:i-1}) = softmax(m_i), \tag{11}$$

where h_1 and h_2 are two RNN cells, f_1 and f_2 are MLPs, y is the holistic sentiment, c_i can be considered as a context vector which contains the semantic information of $x_{1:i}$ and m_i is the hidden state in the RNN chain of sentiments.

Combined with Eq. (4), the total loss of temporal sentiment control module is:

$$\begin{aligned}\mathcal{S}_2 = \mathbb{E}_{p_l(x,w,y,y_{1:n})}&\sum_{i=1}^{n}[\mathcal{L}(x_{1:i}, y_{1:i}, w) - \log\, q(y_i|x_{1:i}, w)] \\ &+ \mathbb{E}_{p_u(x,w)}\sum_{i=1}^{n}\mathcal{U}(x_{1:i}, w).\end{aligned} \tag{12}$$

In the generation process, for one thing, as shown in Figure 3 (b), each line x_i is generated by the decoder $p(x_i|z, x_{1:i-1}, w, y_{1:i})$, that is, each generated line is conditioned on previous content sequence and sentiment sequence. For another, each y_i (if not specified by the user) is predicted by a classifier $p(y_i|x_{1:i-1}, w, y_{1:i-1})$[2]. The content sequence and the sentiment sequence are interactively generated, therefore we name this configuration as the temporal sentiment control module.

4.3 Training

When only using the temporal sentiment control module, we also need to train a classifier $p(y|w)$ to predict a holistic label y which is necessary for Eq. (10). Simply we can add a loss $-\log p(y|w)$ to the loss of the temporal module, but we find it brings much noise for the RNN chain of sentiments and makes the training unstable.

Therefore we combine the two modules, and the overall objective of our SCPG model is to minimize $\mathcal{S} = \mathcal{S}_1 + \lambda\mathcal{S}_2$ which utilizes the losses of different granularities to enhance both two modules. We will show this combination brings further improvement in experiments. The hyper-parameter λ is used to balance the holistic sentiment control and temporal sentiment control across lines. We set λ to 1.0 in our model. Besides, we also use the annealing trick [Yang *et al.*, 2018b] and BOW-loss [Zhao *et al.*, 2017] to alleviate the vanishing latent variable problem in VAE training.

5 Experiments

In this section, we first introduce our experimental settings, baselines and then compare these models on quality, diversity, and sentiment control accuracy of generated poems.

5.1 Data and Settings

Since our model is semi-supervised, besides the 5,000 labelled poems mentioned in Section 3, we also utilize the 146,835 unlabelled ones. For unlabelled data, we randomly select 4,500 poems for validation and testing respectively and the rest for training. As in [Yi *et al.*, 2018a], we use TextRank [Mihalcea and Tarau, 2004] to extract three keywords from each poem and build three <keyword, poem> pairs to enable the model to cope with different keywords. For labelled data, we use 500 poems for validation and testing respectively. Similarly, keywords are also extracted to construct supervised triples <keyword, poem, labels>.

[2] A corresponding extended version of the classifier $p(y|w)$ in the holistic module.

Models	Flu.	Coh.	Mea.	Poe.	Ove.
Models without sentiment controlling module					
WM	3.09	2.86	2.8	2.79	2.71
CVAE	2.50	2.41	2.33	2.39	2.24
MRL	3.20	2.96	2.88	**2.95**	**2.86**
Models with sentiment controlling module					
SBasic	2.21	2.09	1.99	1.96	1.87
SCPG-H	3.07	2.82	2.81	2.72	2.65
SCPG-T	**3.23**	2.93	2.88	2.87	2.78
SCPG-HT	**3.23**	**3.04**	**2.92**	2.83	2.78
GT	4.03	4.13	4.00	3.90	3.83

Table 2: Human evaluation results of poetry quality. SCPG-H and SCPG-T represent the SCPG model with only holistic sentiment control and only temporal sentiment control respectively. SCPG-HT denotes the model combining these two modules together. Flu., Coh., Mea., Poe. and Ove. denote Fluency, Coherence, Meaningfulness, Poeticness and Overall respectively.

Models	Jaccard Similarity
WM	3.3%
CVAE	1.8%
MRL	**1.2%**
SCPG-HT	1.5%
GT	0.05%

Table 3: Automatic evaluation of semantic diversity. Lower similarity indicates better diversity.

The dimensions of word embedding, sentiment embedding and latent variable are 256, 32, 128 respectively. The hidden state size is 512 for the encoder, decoder and content sequence; 64 for the sentiment sequence. Adam [Kingma and Ba, 2015] with mini-batches (batch size 64) is used for optimization. We also use dropout (keep ratio=0.75) to avoid overfitting. For testing, all models generate poems with beam search (beam size = 20). We first train our SCPG model using both labelled and unlabelled training sets until the perplexity on the validation set no longer decreases. Then we fine-tune the model only using the labelled data.

5.2 Baselines

We compare our model with GT (ground-truth, namely human-authored poems) and three state-of-the-art poetry generation models.

WM [Yi *et al.*, 2018b]: a working memory model which maintains user topics and generated history in a dynamical reading-and-writing way.

CVAE [Yang *et al.*, 2018b]: a conditional variational autoencoder with a hybrid decoder to learn the implicit topic information within poems lines.

MRL [Yi *et al.*, 2018a]: a reinforcement learning framework which directly models and optimizes human evaluation criteria to tackle the loss-evaluation mismatch problem in poetry generation. This model achieves the so-far best diversity.

Since we are the first work to generate sentimental poetry as aforementioned, we implement a basic supervised sentiment-controllable model called **SBasic** for comparison which simply feeds learned sentiment embeddings into the same decoding module utilized in our SCPG model. SBasic can be considered as a modified version of [Wei *et al.*, 2018].

5.3 Poetry Quality Evaluation

We first compare our SCPG with baseline models on the quality of generated poems. As the sentiment labels are not provided in this experiment, we use the classifiers trained in our model to infer a holistic sentiment and corresponding line sentiments for each poem. For SBasic, we input the same sentiment labels as inferred in SCPG for fair comparisons.

As studied in [Yi *et al.*, 2018a], perplexity or BLEU departs from the human evaluation manner, therefore we directly conduct human evaluations following the settings in [Yang *et al.*, 2018a; Yi *et al.*, 2018a; Zhang *et al.*, 2017]. We use the five criteria designed by [Manurung, 2003]: **Fluency** (is the generated poem well-formed?), **Coherence** (are the meaning and theme of the poem consistent?), **Meaningfulness** (does the poem convey some certain messages?), **Poeticness** (does the poem have some poetic attributes?), **Overall** (the general impression of the poem). Each criterion is scored on a 5-point scale ranging from 1 to 5.

We randomly select 30 keywords and generate two poems for each keyword. For GT, we randomly choose poems containing the keyword from the test set. Therefore, we generate 480 poems (30*2*8) in all for our models and baselines. We invite 10 experts, who have received professional Chinese poetry education, to conduct evaluations. We randomly split the experts into 2 groups and each evaluates all generated poems, with each expert 96 poems. Then we average the scores of these two groups on each criterion to alleviate personal bias.

As shown in Table 2, from the top part, we can observe that our SCPG-HT achieves comparable performance to the state-of-art MRL, which directly optimizes these criteria with reinforcement learning while our model focuses on the sentiment control. Besides, our model obtains the highest score on meaning, owing to that our model endows each poem with a specific sentiment while baseline models tend to create general and meaningless content without sentiments. We also get higher scores on fluency and coherence, benefiting from the temporal sentiment control module which predicts appropriate labels and brings a more natural transition among lines. However, there is still a large gap between our SCPG model and human poets.

In the bottom part of Table 2, we compare the performance of models with sentiment control. We can find that our SCPG models consistently outperform SBasic by a margin. It's because that though pre-trained with unlabelled data, SBasic only learns to stiffly load high-frequency sentimental words together which hurts the quality. Our model utilizes not only the sentiment label but also a learned latent variable z in a semi-supervised way to capture more generalized sentiment-related information, which maintains the semantics when focusing on the sentiment. Besides, SCPG-T outperforms SCPG-H since the temporal module leads to a more natural transition. By combining these two modules, SCPG-HT gets the best performance.

5.4 Poetry Diversity Evaluation

In this part, we evaluate the sentimental diversity and semantic diversity of poems generated by different models.

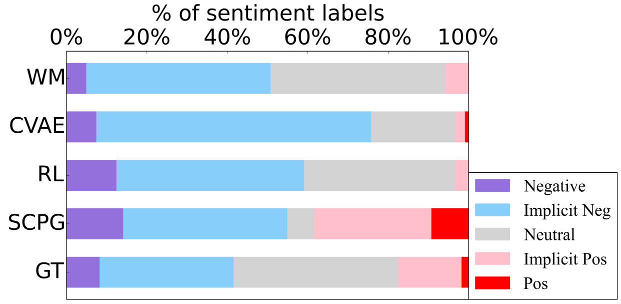


Figure 4: Normalized distributions of holistic sentiments in different models. Neg:negative sentiment; Pos: positive sentiment.

To evaluate sentimental diversity, we ask the experts to annotate the holistic sentiments of the generated poems in Section 5.3. As in Figure 4, baseline models tend to generate non-sentimental poems (sentiment collapse) or negative poems (sentiment bias), while our SCPG can create poems with more diverse and balanced sentiments. Please note that we don't manually input any sentiment signal. SCPG predicts appropriate sentiments with the keyword by itself.

To measure the semantic diversity, following [Yi *et al.*, 2018a], we utilize bigram-based Jaccard similarity. Given different input keywords, less similar generated poems can reflect higher diversity to some extent. As shown in Table 3, MRL gets the lowest similarity, benefiting from its direct optimization of the criteria. Though we don't optimize the semantic diversity directly, our SCPG gets the comparable performance to MRL. This result also verifies our suppose in Section 1 that sentiment is closely coupled with semantics.

5.5 Sentiment Control Evaluation

To evaluate the accuracy of sentiment control, we select 10 neutral keywords, such as *river*, *dream* and *clothes*, and generate two poems in each sentiment class for each keyword. Thus we generate 100 poems (10*2*5) for each model. Then we ask 5 experts to manually label the sentiments of the generated poems as mentioned in Section 3.

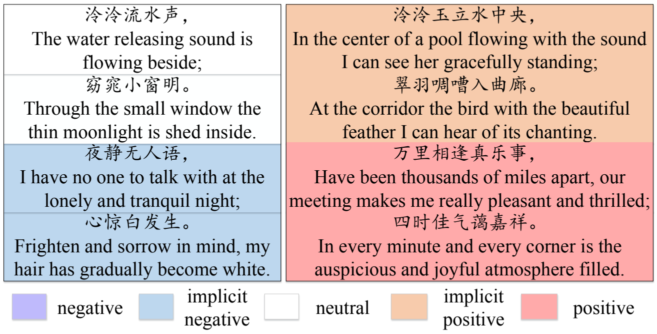
As shown in Table 4, we evaluate the control accuracy in both holistic level and line level. In the holistic level, our SCPG-HT model achieves higher accuracy in both 5 classes and 3 classes compared to SBasic. Notice that the poems generated by SBasic are in quite poor quality as shown in Table 2. In the line level, we can observe that SCPG-HT matches the control more accurately because of the constraint of holistic sentiment, while the quality of poems generated with these two models is comparable (Table 2). Besides, the overall accuracy in the line level is lower than that in the holistic level because the length of each line is limited which allows less freedom to express a specific sentiment.

5.6 Case Study

We present two poems generated by SCPG in Figure 5. We can see that these poems not only hold the holistic sentiment of the whole poem but also show a reasonable sentiment transition among lines as a human-authored piece. To conclude, our SCPG can generate fluent and coherent poems with fine-grained sentiments in the discourse level.

Models	5 Classes	3 Classes
Holistic Sentiment Accuracy		
SBasic	0.412	0.633
SCPG-HT	0.461	0.733
Line Sentiment Accuracy		
SCPG-T	0.379	0.550
SCPG-HT	0.441	0.637

Table 4: Human evaluation of sentiment control accuracy. The 3 classes are "neg" ("implicit neg" or "neg"), "pos" ("implicit pos" or "pos") and "neutral".



泠泠流水声，
The water releasing sound is flowing beside;
窈窕小窗明。
Through the small window the thin moonlight is shed inside.
夜静无人语，
I have no one to talk with at the lonely and tranquil night;
心惊白发生。
Frighten and sorrow in mind, my hair has gradually become white.

泠泠玉立水中央，
In the center of a pool flowing with the sound I can see her gracefully standing;
翠羽啁嘈入曲廊。
At the corridor the bird with the beautiful feather I can hear of its chanting.
万里相逢真乐事，
Have been thousands of miles apart, our meeting makes me really pleasant and thrilled;
四时佳气蔼嘉祥。
In every minute and every corner is the auspicious and joyful atmosphere filled.

Figure 5: Poems generated by holistic sentiments "implicit neg" and "pos" respectively given the same keyword "sound of water".

6 Conclusion and Future Work

In this paper, we introduce a fine-grained sentimental poetry corpus and propose a sentiment-controllable poetry generation model (SCPG)[3]. Based on a semi-supervised variational autoencoder, our model learns a sentiment-conditioned latent space to capture generalized sentimental semantics and control not only the holistic sentiment of the whole poem but also the temporal sentiment transition across lines. Experiments on Chinese poems show our model achieves significant improvement on sentiment control accuracy compared to the baselines. Besides, our model improves both the sentimental and semantic diversity against the state-of-the-arts without losing quality.

For future works, we will consider adopting our temporal module to generate other genres of discourse-level text and making efforts to expand our sentimental poetry corpus to better assist related research.

Acknowledgements

We would like to thank Fanchao Qi, Yankai Lin, Ruoyu Li and anonymous reviewers for their insightful comments. This research is supported by Tencent. It is also partially supported by the NSFC project under the grant no. 61661146007 and the NExT++ project, the National Research Foundation, Prime Minister's Office, Singapore under its IRC@Singapore Funding Initiative.

[3] Our source code and the Fine-grained Sentiment Poetry Corpus will be available at https://github.com/THUNLP-AIPoet. This model will also be incorporated into Jiuge, the THUNLP online poetry generation system, https://jiuge.thunlp.cn.

References

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 2015 ICLR*, San Diego, CA, 2015.

[Cagan *et al.*, 2017] Tomer Cagan, Stefan L Frank, and Reut Tsarfaty. Data-driven broad-coverage grammars for opinionated natural language generation (onlg). 2017.

[Chari, 1976] VK Chari. Poetic emotions and poetic semantics. *The Journal of Aesthetics and Art Criticism*, 34(3):287–299, 1976.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 EMNLP*, pages 1724–1734, Doha, Qatar, 2014.

[Gervás, 2001] Pablo Gervás. *An Expert System for the Composition of Formal Spanish Poetry*, pages 181–188. Springer London, 2001.

[Ghazvininejad *et al.*, 2016] Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. Generating topical poetry. In *Proceedings of the 2016 EMNLP*, pages 1183–1191, Texas, USA, 2016.

[He *et al.*, 2012] Jing He, Ming Zhou, and Long Jiang. Generating chinese classical poems with statistical machine translation models. In *Proceedings of the 2012 AAAI*, pages 1650–1656, Toronto, Canada, 2012.

[Hu *et al.*, 2017] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *Proceedings of the 2017 ICML*, pages 1587–1596. JMLR. org, 2017.

[Janowitz, 1973] KE Janowitz. The antipodes of self: Three poems by christina rossetti. *Victorian Poetry*, pages 195–205, 1973.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *Proceedings of the 2015 International Conference on Learning Representations*, San Diego, CA, 2015.

[Kingma *et al.*, 2014] Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. In *In Advances in 2014 NeurlPS*, Montreal, Canada, 2014.

[Levy, 2001] Robert P. Levy. A computational model of poetic creativity with neural network as measure of adaptive fitness. In *Proceedings of the ICCBR-01 Workshop on Creative Systems*, 2001.

[Manurung, 2003] Hisar Maruli Manurung. *An evolutionary algorithm approach to poetry generation*. PhD thesis, University of Edinburgh, 2003.

[Mihalcea and Tarau, 2004] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 EMNLP*, 2004.

[Morris-Jones, 1962] Huw Morris-Jones. The language of feelings. *The British Journal of Aesthetics*, 2(1):17–25, 1962.

[Wang and Wan, 2018] Ke Wang and Xiaojun Wan. Sentigan: Generating sentimental texts via mixture adversarial networks. In *Proceedings of the 2018 IJCAI*, pages 4446–4452, 2018.

[Wang *et al.*, 2016a] Qixin Wang, Tianyi Luo, Dong Wang, and Chao Xing. Chinese song iambics generation with neural attention-based model. In *Proceedings of the 2016 IJCAI*, pages 2943–2949, New York, USA, 2016.

[Wang *et al.*, 2016b] Zhe Wang, Wei He, Hua Wu nad Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. Chinese poetry generation with planning based neural network. In *Proceedings of COLING 2016*, pages 1051–1060, Osaka, Japan, 2016.

[Wei *et al.*, 2018] Jia Wei, Qiang Zhou, and Yici Cai. Poet-based poetry generation: Controlling personal style with recurrent neural networks. In *Proceedings of the Workshop on Computing, Networking and Communications*, pages 156–160, 2018.

[Yang *et al.*, 2018a] Cheng Yang, Maosong Sun, Xiaoyuan Yi, and Wenhao Li. Stylistic chinese poetry generation via unsupervised style disentanglement. In *Proceedings of the 2018 EMNLP*, pages 3960–3969, Brussels, Belgium, 2018.

[Yang *et al.*, 2018b] Xiaopeng Yang, Xiaowen Lin, Shunda Suo, and Ming Li. Generating thematic chinese poetry using conditional variational autoencoders with hybrid decoders. In *Proceedings of the 2018 IJCAI*, pages 4539–4545, Stockholm, Sweden, 2018.

[Yi *et al.*, 2018a] Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Wenhao Li. Automatic poetry generation with mutual reinforcement learning. In *Proceedings of the 2018 EMNLP*, pages 3143–3153, Brussels, Belgium, 2018.

[Yi *et al.*, 2018b] Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Zonghan Yang. Chinese poetry generation with a working memory mode. In *Proceedings of the 2018 IJCAI*, pages 4553–4559, Stockholm, Sweden, 2018.

[Zhang and Lapata, 2014] Xingxing Zhang and Mirella Lapata. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 EMNLP*, pages 670–680, Doha, Qatar, 2014.

[Zhang *et al.*, 2017] Jiyuan Zhang, Yang Feng, Dong Wang, Yang Wang, Andrew Abel, Shiyue Zhang, and Andi Zhang. Flexible and creative chinese poetry generation using neural memory. In *Proceedings of the 2017 ACL*, pages 1364–1373. Association for Computational Linguistics, 2017.

[Zhao *et al.*, 2017] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 2017 ACL*, pages 654–664. Association for Computational Linguistics, 2017.