Rhetorically Controlled Encoder-Decoder for Modern Chinese Poetry Generation

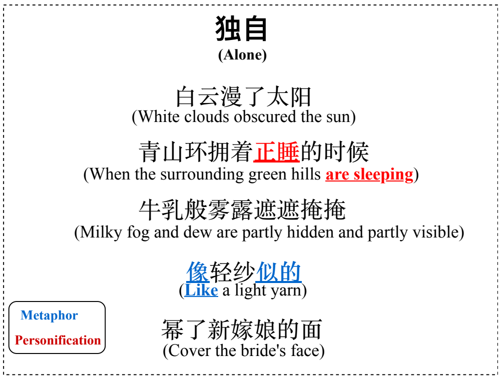
**Zhiqiang Liu[†], Zuohui Fu[‡*], Jie Cao[◇*], Gerard de Melo[‡],
Yik-Cheung Tam[†], Cheng Niu[†] and Jie Zhou[†]**
[†]Pattern Recognition Center, WeChat AI, Tencent Inc, China
[‡]Department of Computer Science, Rutgers University
[◇]School of Computing, University of Utah
zhiqliu@tencent.com, zuohui.fu@rutgers.edu, jcao@cs.utah.edu
gdm@demelo.org, {wilsontam,niucheng,withtomzhou}@tencent.com

Abstract

Rhetoric is a vital element in modern poetry, and plays an essential role in improving its aesthetics. However, to date, it has not been considered in research on automatic poetry generation. In this paper, we propose a rhetorically controlled encoder-decoder for modern Chinese poetry generation. Our model relies on a continuous latent variable as a rhetoric controller to capture various rhetorical patterns in an encoder, and then incorporates rhetoric-based mixtures while generating modern Chinese poetry. For metaphor and personification, an automated evaluation shows that our model outperforms state-of-the-art baselines by a substantial margin, while a human evaluation shows that our model generates better poems than baseline methods in terms of fluency, coherence, meaningfulness, and rhetorical aesthetics.

1 Introduction

Modern Chinese poetry, originating from 1900 CE, is one of the most important literary formats in Chinese culture and indeed has had a profound influence on the development of modern Chinese culture. Rhetoric is a vital element in modern poetry, and plays an important role in enhancing its aesthetics. Incorporating intentional rhetorical embellishments is essential to achieving the desired stylistic aspects of impassioned modern Chinese poetry. In particular, the use of metaphor and personification, both frequently used forms of rhetoric, are able to enrich the emotional impact of a poem. Specifically, a metaphor is a figure of speech that describes one concept in terms of another one. Within this paper, the term "metaphor" is considered in the sense of a general figure of speech 比喻 (*bi yu*), encompassing both metaphor in its narrower sense and similes. Personification is a figure of speech in which a thing, an idea or an animal is given human attributes, i.e., non-human objects are portrayed in such a way that we feel they have the ability to act like human beings. For example, 她笑起来像花儿一样 (*'She smiles like lovely flowers'*) with its connection between smiling and flowers highlights extraordinary beauty and pureness in describing the verb 'smile'. 夜空中的星星眨着眼睛 (*'Stars in the night sky squinting'*) serves as an example of personification, as stars are personified and described as *squinting*, which is normally considered an act of humans, but here is invoked to more vividly describe twinkling stars.



独自
(Alone)

白云漫了太阳
(White clouds obscured the sun)

青山环拥着**正睡**的时候
(When the surrounding green hills **are sleeping**)

牛乳般雾露遮遮掩掩
(Milky fog and dew are partly hidden and partly visible)

像轻纱**似的**
(**Like** a light yarn)

幂了新嫁娘的面
(Cover the bride's face)

Metaphor
Personification

Figure 1: A modern Chinese poetry with metaphor and personification.

As is well known, rhetoric encompasses a variety of forms, including metaphor, personification, exaggeration, and parallelism. For our work, we collected more than 8,000 Chinese poems and over 50,000 Chinese song lyrics. Based on the statistics given in Table 1, we observe that metaphor and personification are the most frequently used rhetorical styles in modern Chinese poetry and lyrics (see Section 4.1 for details about this data).

[*] The work was done when Zuohui Fu and Jie Cao were interns at Pattern Recognition Center, WeChat AI, Tencent Inc.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1992–2001
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

Dataset	Docs	Lines	Metaphor	Personification
Poetry	8,744	137,105	31.4%	18.5%
Lyrics	53,150	1,036,425	23.8%	13.2%

Table 1: Quantitative evaluation of the phenomena of metaphor and personification in modern Chinese poems and lyrics.

Hence, we will mainly focus on the generation of metaphor and personification in this work. As an example, an excerpt from the modern Chinese poem 独自 (*Alone*) is given in Figure 1, where the fourth sentence (highlighted in blue) invokes a metaphorical simile, while the second one (highlighted in red) contains a personification.

In recent years, neural generation models have become widespread in natural language processing (NLP), e.g., for response generation in dialogue (Le et al., 2018), answer or question generation in question answering, and headline generation in news systems. At the same time, poetry generation is of growing interest and has attained high levels of quality for classical Chinese poetry. Previously, Chinese poem composing research mainly focused on traditional Chinese poems. In light of the mostly short sentences and the metrical constraints of traditional Chinese poems, the majority of research attention focused on term selection to improve the thematic consistency (Wang et al., 2016).

In contrast, modern Chinese poetry is more flexible and rich in rhetoric. Unlike sentiment-controlled or topic-based text generation methods (Ghazvininejad et al., 2016), which have been widely used in poetry generation, existing research has largely disregarded the importance of rhetoric in poetry generation. Yet, to emulate human-written modern Chinese poems, it appears necessary to consider not only the topics but also the form of expression, especially with regard to rhetoric. In this paper, we propose a novel rhetorically controlled encoder-decoder framework inspired by the above sentiment-controlled and topic-based text generation methods, which can effectively generate poetry with metaphor and personification.

Overall, the contributions of the paper are as follows:

- We present the first work to generate modern Chinese poetry while controlling for the use of metaphor and personification, which play an essential role in enhancing the aesthetics of poetry.
- We propose a novel metaphor and personification generation model with a rhetorically controlled encoder-decoder.
- We conduct extensive experiments showing that our model outperforms the state-of-the-art both in automated and human evaluations.

2 Related Work

2.1 Poetry Generation

Poetry generation is a challenging task in NLP. Traditional methods (Gervás, 2001; Manurung, 2004; Greene et al., 2010; He et al., 2012) relied on grammar templates and custom semantic diagrams. In recent years, deep learning-driven methods have shown significant success in poetry generation, and topic-based poetry generation systems have been introduced (Ghazvininejad et al., 2017, 2018; Yi et al., 2018b). In particular, Zhang and Lapata (2014) propose to generate Chinese quatrains with Recurrent Neural Networks (RNNs), while Wang et al. (2016) obtain improved results by relying on a planning model for Chinese poetry generation.

Recently, Memory Networks (Sukhbaatar et al., 2015) and Neural Turing Machines (Graves et al., 2014) have proven successful at certain tasks. The most relevant work for poetry generation is that of Zhang et al. (2017), which stores hundreds of human-authored poems in a static external memory to improve the generated quatrains and achieve a style transfer. The above models rely on an external memory to hold training data (i.e., external poems and articles). In contrast, Yi et al. (2018a) dynamically invoke a memory component by saving the writing history into memory.

2.2 Stylistic Language Generation

The ability to produce diverse sentences in different styles under the same topics is an important characteristic of human writing. Some works have explored style control mechanisms for text generation tasks. For example, Zhou and Wang (2018) use naturally labeled emojis for large-scale emotional response generation in dialogue. Ke et al. (2018) and Wang et al. (2018) propose a sentence controlling function to generate interrogative, imperative, or declarative responses in dialogue. For the task of poetry generation, Yang et al. (2018) introduce an unsupervised style labeling to generate stylistic poetry, based on mutual information. Inspired by the above works, we regard rhetoric in poetry as a specific style and adopt a Conditional Variational Autoencoder (CVAE) model to generate rhetoric-aware poems.

CVAEs (Sohn et al., 2015; Larsen et al., 2016) extend the traditional VAE model (Kingma and Welling, 2014) with an additional conditioned label to guide the generation process. Whereas VAEs essentially directly store latent attributes as probability distributions, CVAEs model latent variables conditioned on random variables. Recent research in dialogue generation shows that language generated by VAE models benefit from a significantly greater diversity in comparison with traditional Seq2Seq models. Recently, CVAEs and adversarial training have been explored for the task of generating classical Chinese poems (Li et al., 2018).

3 Methodology

In this paper, our goal is to leverage metaphor and personification (known as rhetoric modes) in modern Chinese poetry generation using a dedicated rhetoric control mechanism.

3.1 Overview

Before presenting our model, we first formalize our generation task. The inputs are poetry topics specified by K user-provided keywords $\{w_k\}_{k=1}^K$. The desired output is a poem consisting of n lines $\{L_i\}_{i=1}^n$. Since we adopt a sequence-to-sequence framework and generate a poem line by line, the task can be cast as a text generation one, requiring the repeated generation of an i-th line that is coherent in meaning and related to the topics, given the previous $i-1$ lines $L_{1:i-1}$ and the topic keywords $w_{1:K}$. In order to control the rhetoric modes, the rhetoric label r may be provided either as an input from the user, or from an automatic prediction based on the context. Hence, the task of poetry line generation can be formalized as follows:

$$L_i^* = \arg\max_L P(L \mid L_{1:i-1}, w_{1:K}, r_i) \qquad (1)$$

As mentioned above, incorporating rhetoric into poetic sentences requires controlling for the rhetoric mode and memorizing contextual topic information. To this end, we first propose two conditional variational autoencoder models to effectively control *when* to generate rhetoric sentences, and *which rhetoric mode* to use. The first model is a *Manual Control CVAE* model (MCCVAE). It receives the user's input signal as a rhetoric label r to generate the current sentence in the poem, and is designed for user-controllable poetry generation tasks. The second model is the *Automatic Control CVAE* (ACCVAE), which automatically predicts when to apply appropriate forms of rhetoric and generates the current sentence based on contextual information.

Subsequently, to memorize pertinent topic information and generate more coherent rhetorical sentences, we propose a topic memory component to store contextual topic information. At the same time, we propose a rhetorically controlled decoder to generate appropriate rhetorical sentences. This is a mechanism to learn the latent rhetorical distribution given a context and a word, and then perform a rhetorically controlled term selection during the decoding stage. Our proposed framework will later be presented in more detail in Figure 2.

3.2 Seq2seq Baseline

Our model is based on the sequence-to-sequence (Seq2Seq) framework, which has been widely used in text generation. The encoder transforms the current input text $X = \{x_1, x_2, ..., x_J\}$ into a hidden representation $H = \{h_1, h_2, ..., h_J\}$, as follows:

$$\mathbf{h}_j = \mathrm{LSTM}(\mathbf{e}(x_j), \mathbf{h}_{j-1}), \qquad (2)$$

where LSTM is a Long Short-Term Memory Network, and $\mathbf{e}(x_j)$ denotes the embedding of the word x_j.

The decoder first updates the hidden state $\mathbf{S} = \{s_1, s_2, .., s_T\}$, and then generates the next sequence $Y = \{y_1, y_2, ..., y_T\}$ as follows:

$$\begin{aligned} \mathbf{s}_t &= \mathrm{LSTM}(\mathbf{e}(y_{t-1}), \mathbf{s}_{t-1})) \\ P(y_t \mid y_{t-1}, s_t) &= \mathrm{softmax}(W\mathbf{s}_t), \end{aligned} \qquad (3)$$

where this second LSTM does not share parameters with the encoder's network.

3.3 Proposed Models

In the following, we will describe our models for rhetorically controlled generation.

3.3.1 Manual Control (MC) CVAE

We introduce a Conditional Variational Autoencoder (CVAE) for the task of poetry generation. Mathematically, the CVAE is trained by maximizing a variational lower bound on the conditional likelihood of Y given c, in accordance with

$$p(Y \mid c) = \int p(Y \mid z, c)\, p(z \mid c)\, dz, \qquad (4)$$

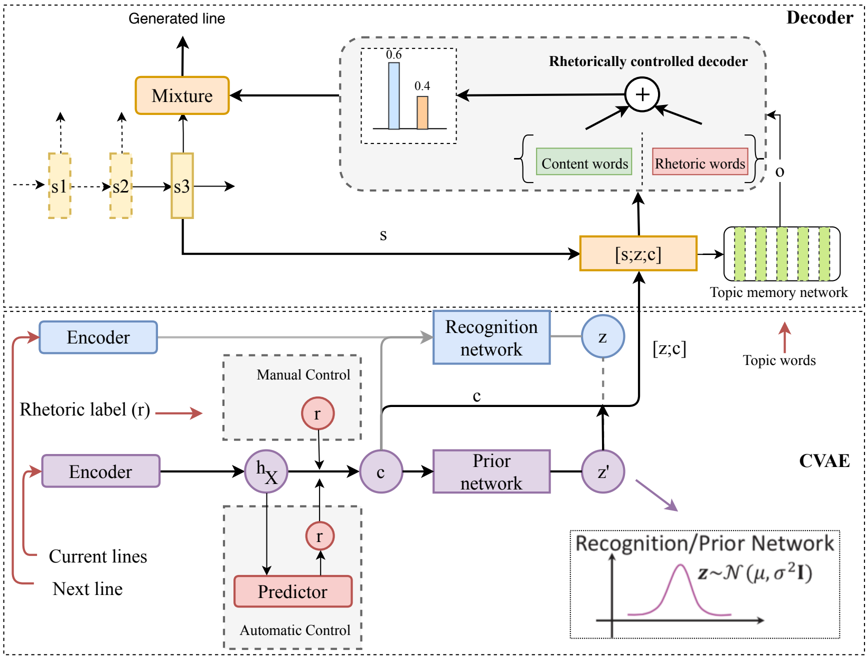


Figure 2: Illustration of our model.

where z, c, and Y are random variables, and the latent variable z is used to encode the semantics and rhetoric of the generated sentence. In our manual control model, the conditional variables that capture the input information are $c = [h_X; e(r)]$, where $e(r)$ is the embedding of the rhetorical variable r. h_X is the encoding of current poem sentences X, and the target Y represents the next sentence to be generated.

Then on top of the traditional Seq2seq model, we introduce a prior network, a recognition network, and the decoder: (i) The prior network $p_{\mathrm{P}}(z|c)$ is an approximation of $p(z|c)$. (ii) The decoder $p_{\mathrm{D}}(Y|z,c)$ is used to approximate $p(Y|z,c)$. (iii) The recognition network $q_{\mathrm{R}}(z|Y,c)$ serves to approximate the true posterior $p(z|Y,c)$. Then the variational lower bound to the loss $-\log p(Y|c)$ can be expressed as:

$$\begin{aligned} & -\mathcal{L}(\theta_{\mathrm{D}};\theta_{\mathrm{P}};\theta_{\mathrm{R}};Y,c) = \mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{\mathrm{decoderCE}} \\ & = \mathrm{KL}(q_{\mathrm{R}}(z \mid Y,c) \,||\, p_{\mathrm{P}}(z \mid c)) \\ & - \mathbb{E}_{q_R(z|Y,c)} (\log p_{\mathrm{D}}(Y \mid z,c)) \end{aligned} \tag{5}$$

Here, θ_{D}, θ_{P}, θ_{R} are the parameters of the decoder, prior network, and recognition network, respectively. Intuitively, the second term maximizes the sentence generation probability after sampling from the recognition network, while the first term minimizes the distance between prior and recognition network.

Usually, we assume that both the prior and the recognition networks are multivariate Gaussian distributions, and their mean and log variance are estimated through multilayer perceptrons (MLP) as follows:

$$\begin{aligned} \left[\mu, \sigma^2\right] &= \mathrm{MLP}_{\mathrm{posterior}}(\mathrm{LSTM}(Y), c) \\ \left[\mu', \sigma'^2\right] &= \mathrm{MLP}_{\mathrm{prior}}(c) \end{aligned} \tag{6}$$

A single layer of the LSTM is used to encode the current lines, and obtain the h_X component of c. The same LSTM structure is also used to encode the next line Y in the training stage. By using Eq. (6), we calculate the KL divergence between these distributions to optimize Eq. (5). Following the practice in Zhao et al. (2017), a reparameterization technique is used when sampling from the recognition and the prior network during training and testing.

3.3.2 Automatic Control(AC) CVAE

In the ACCVAE model, we first predict the rhetorical mode of the next sentence using an MLP that is designed as follows:

$$\begin{aligned} p(r|h_X) &= \mathrm{softmax}(\mathrm{MLP}_{\mathrm{predictor}}(h_X)) \\ r &= \arg\max p(r \mid h_X) \end{aligned} \tag{7}$$

In this case, the conditional variable c is also $[h_X; e(r)]$, where h_X is taken as the last hidden state of the encoder LSTM. The loss function is then defined as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{\mathrm{decoderCE}} + \mathcal{L}_{\mathrm{predictorCE}} \tag{8}$$

In this paper, a two-layer MLP is used for Eq. (7).

3.4 Topic Memory Component

As shown above, LSTMs are used to encode the lines of the poem. Considering the fact that Memory Networks (Sukhbaatar et al., 2015) have demonstrated great power in capturing long temporal dependencies, we incorporate a memory component for the decoding stage. By equipping it with a larger memory capacity, the memory is able to retain temporally distant information in the writing history, and provide a RAM-like mechanism to support model execution. In our poetry generation model, we rely on a special topic memory component to memorize both the topic and the generation history, which are of great help in generating appropriate rhetorical and semantically consistent sentences.

As illustrated in Figure 2, our topic memory is $M \in \mathcal{R}^{K' \times d_h}$, where each row of the matrices is a memory slot with slot size d_h and the number of slots is K'. Before generating the i-th line L_i, topic words w_k from the user and the input text are written into the topic memory in advance, which remains unchanged during the generation of a sentence.

Memory Reading. We introduce an *Addressing Function* as $\alpha = A(M, q)$, which calculates the probabilities of each slot of the memory being selected and invoked. Specifically, we define:

$$\begin{aligned} z_k &= b^T \sigma(M_k, q) \\ \alpha_k &= \text{softmax}(z_k), \end{aligned} \tag{9}$$

where σ defines a non-linear layer, q is the query vector, b is the parameter, M is the memory to be addressed, M_k is the k-th slot of M, and α_k is the k-th element in vector α. For the topic memory component, the input q should be $[s_{t-1}; c; z]$, so the topic memory is read as follow:

$$\begin{aligned} \alpha' &= A_r(M, [s_{t-1}; c; z]) \\ o_t &= \sum_{k=1}^{K'} \alpha'_k M_k, \end{aligned} \tag{10}$$

where α' is the reading probability vector, s_{t-1} represents the decoder hidden state, and o_t is the memory output at the t-th step.

3.5 Rhetorically Controlled Decoder

A general Seq2seq model may tend to emit generic and meaningless sentences. In order to create poems with more meaningful and diverse rhetoric, we propose a rhetorically controlled decoder. It assumes that each word in a poem sentence has a latent type designating it as a content word or as a rhetorical word. The decoder then calculates a word type distribution over the latent types given the context, and computes type-specific generation distributions over the entire vocabulary. The final probability of generating a word is a mixture of type-specific generation distributions, where the coefficients are type probabilities. The final generation distribution $\mathcal{P}(y_t \mid s_t, o_t, z, c)$ from the sampled word is defined as

$$\begin{aligned} &\mathcal{P}(y_t \mid s_t, o_t, z, c) = \\ &\mathcal{P}(y_t \mid \tau_t = \text{content}, s_t, o_t, z, c) \\ &\quad \mathcal{P}(\tau_t = \text{content} \mid s_t, z, c) \\ &+\mathcal{P}(y_t \mid \tau_t = \text{rhetoric}, s_t, z, c) \\ &\quad \mathcal{P}(\tau_t = \text{rhetoric} \mid s_t, z, c), \end{aligned} \tag{11}$$

where τ_t denotes the word type at time step t. This specifies that the final generation probability is a mixture of the type-specific generation probability $\mathcal{P}(y_t \mid \tau_t, s_t, z, c)$, weighted by the probability of the type distribution $\mathcal{P}(\tau_t \mid s_t, z, c)$. We refer to this decoder as a *rhetorically controlled decoder*. The probability distribution over word types is given by

$$\mathcal{P}(\tau_t \mid s_t, z, c) = \text{softmax}(W_0[s_t; z; c] + b_0),$$

where s_t is the hidden state of the decoder at time step t, $W \in R^{k \times d}$ with the dimension d. The word type distribution predictor can be trained in decoder training stage together. The type-specific generation distribution is given by

$$\begin{aligned} &\mathcal{P}(y_t \mid \tau_t = \text{content}, s_t, o_t, z, c) = \\ &\text{softmax}(W_{\text{content}}[s_t; o_t; z; c] + b_{\text{content}}) \end{aligned} \tag{12}$$

$$\begin{aligned} &\mathcal{P}(y_t \mid \tau_t = \text{rhetoric}, s_t, z, c) = \\ &\text{softmax}(W_{\text{rhetoric}}[s_t; z; c] + b_{\text{rhetoric}}), \end{aligned} \tag{13}$$

where W_{content}, $W_{\text{rhetoric}} \in R^{|V| \times d}$, and $|V|$ is the size of the entire vocabulary. Note that the type-specific generation distribution is parameterized by these matrices, indicating that the distribution for each word type has its own parameters.

Instead of using a single distribution, our rhetorically controlled decoder enriches the model by applying multiple type-specific generation distributions, which enables the model to convey more information about the potential word to be generated. Also note that the generation distribution is over the same vocabulary.

Model	Precision	Recall	F1
Metaphor	0.93	0.92	0.92
Personification	0.69	0.62	0.65
Other	0.76	0.82	0.79

Table 2: Results of the rhetoric classifier on the test sets.

3.6 Overall Loss Function

The CVAE and Seq2seq model with the rhetorically controlled decoder should be trained jointly. Therefore, the overall loss $\mathcal{L}$ is a linear combination of the KL term $\mathcal{L}_{\mathrm{KL}}$, the classification loss of the rhetoric predictor cross entropy (CE) $\mathcal{L}_{\mathrm{predictorCE}}$, the generation loss of the rhetorical controlled decoder cross entropy $\mathcal{L}_{\mathrm{decoderCE}}$, and the word type classifier (word type distribution predictor) cross entropy $\mathcal{L}_{\mathrm{word\_classifier}}$:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{\mathrm{decoderCE}} + \\ \mathcal{L}_{\mathrm{word\_classifier}} + \gamma \mathcal{L}_{\mathrm{predictorCE}} \end{aligned} \tag{14}$$

The technique of KL *cost annealing* can address the optimization challenges of vanishing latent variables in this encoder-decoder architecture. γ is set to 0 if the Manual Control CVAE is used, and 1 otherwise.

4 Experiments

4.1 Datasets and Setups

We conduct all experiments on two datasets[1]. One is a modern Chinese poetry dataset, while the other is a modern Chinese lyrics dataset. We collected the modern Chinese poetry dataset from an online poetry website[2] and crawled about 100,000 Chinese song lyrics from a small set of online music websites. The sentence rhetoric label is required for our model training. To this end, we built a classifier to predict the rhetoric label automatically. We sampled about 15,000 sentences from the original poetry dataset and annotated the data manually with three categories, i.e., *metaphor*, *personification*, and *other*. This dataset was divided into a training set, validation set, and test set. Three classifiers, including LSTM, Bi-LSTM, and Bi-LSTM with a self-attention model, were trained on this dataset. The Bi-LSTM with self-attention classifier (Yang et al., 2016) outperforms the other models and achieves the best accuracy of 0.83 on the test set. In this classifier, the sizes of word embedding, hidden state and the attention size are set to 128, 256, 30 respectively, and a two-layer LSTM is used. The results for different classes are given in Table 2.

Additionally, we select a large number of poem sentences with metaphor and personification to collect the corresponding rhetorical words. Based on statistics of word counts and part of speech, we obtained over 500 popular words associated with metaphor and personification as rhetorical words. Our statistical results show that these words cover a wide range of metaphorical and anthropomorphic features.

Meanwhile, in our entire model, the sizes of word embedding, rhetoric label embedding, hidden state are set to 128, 128, 128 respectively. The dimensionality of the latent variable is 256 and a single-layer decoder is used. The word embedding is initialized with word2vec vectors pre-trained on the whole corpus.

4.2 Models for Comparisons

We also compare our model against previous state-of-the-art poetry generation models:

- **Seq2Seq**: A sequence-to-sequence generation model, as has been successfully applied to text generation and neural machine translation (Vinyals and Le, 2015).
- **HRED**: A hierarchical encoder-decoder model for text generation (Serban et al., 2016), which employs a hierarchical RNN to model the sentences at both the sentence level and the context level.
- **WM**: A recent Working Memory model for poetry generation (Yi et al., 2018b).
- **CVAE**: A standard CVAE model without the specific decoder. We adopt the same architecture as that introduced in Zhao et al. (2017).

4.3 Evaluation Design

In order to obtain objective and realistic evaluation results, we rely on a combination of both machine evaluation and human evaluation.

Automated Evaluation. To measure the effectiveness of the models automatically, we adopt several metrics widely used in existing studies. **BLEU** scores[3] and **Perplexity** are used to quantify

[1] https://github.com/Lucien-qiang/Rhetoric-Generator

[2] http://www.shigeku.com/

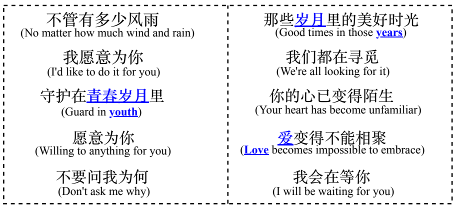
[3] The BLEU score is calculated with the standard multi-bleu.perl script.

Dataset	Model	BLEU(%)	PPL	Precision	Recall	Rhetoric-F1	Distinct-1	Distinct-2
Poetry	Seq2seq	0.38	124.55	0.49	0.45	0.47	0.0315	0.0866
	HRED	0.41	119.74	0.51	0.50	0.50	0.0347	0.0924
	CVAE	0.44	108.72	0.62	0.61	0.61	0.0579	0.1775
	WM	0.42	115.39	0.57	0.60	0.58	0.0498	0.1243
	AC model (ours)	0.43	112.28	0.64	0.65	0.64	**0.0607**	**0.1854**
	MC model (ours)	**0.47**	**95.65**	**0.68**	**0.67**	**0.67**	0.0595	0.1747
Lyrics	Seq2seq	0.52	257.06	0.37	0.34	0.35	0.0149	0.0574
	HRED	0.54	201.85	0.37	0.35	0.36	0.0193	0.0602
	CVAE	**0.59**	**147.45**	0.40	0.41	0.41	0.0231	0.0655
	WM	0.55	183.67	0.37	0.40	0.38	0.0216	0.0628
	AC model (ours)	0.58	159.78	0.41	0.41	0.41	**0.0325**	**0.0817**
	MC model (ours)	0.57	170.46	**0.45**	**0.49**	**0.47**	0.0273	0.0739

Table 3: Results of machine evaluation. **PPL** represents perplexity.

	Poetry				Lyrics			
	F	C	M	RA	F	C	M	RA
Seq2Seq	2.7	2.4	2.8	2.3	3.0	2.4	2.9	2.4
HRED	2.8	2.9	2.7	2.5	2.9	2.7	3.0	2.3
CVAE	**3.2**	2.7	3.0	3.1	**3.3**	2.6	2.9	2.9
WM	3.1	**3.4**	3.1	3.0	3.1	**3.1**	2.8	2.7
AC model (ours)	3.0	**3.4**	**3.2**	**3.5**	**3.3**	3.0	**3.1**	**3.2**

Table 4: The results of human evaluation. **F** means *Fluency*. **C** stands for *Coherence*. **M** represents *Meaningfulness* while **RA** represents *Rhetorical Aesthetics*.



不管有多少风雨
(No matter how much wind and rain)

我愿意为你
(I'd like to do it for you)

守护在青春岁月里
(Guard in **youth**)

愿意为你
(Willing to anything for you)

不要问我为何
(Don't ask me why)

(a) Seq2seq model

那些岁月里的美好时光
(Good times in those **years**)

我们都在寻觅
(We're all looking for it)

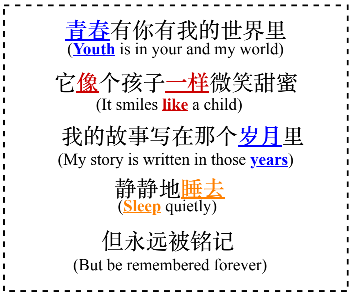
你的心已变得陌生
(Your heart has become unfamiliar)

爱变得不能相聚
(**Love** becomes impossible to embrace)

我会在等你
(I will be waiting for you)

(b) WM model

Figure 3: The results of the Seq2Seq and WM model.



青春有你有我的世界里
(**Youth** is in your and my world)

它像个孩子一样微笑甜蜜
(It smiles **like** a child)

我的故事写在那个岁月里
(My story is written in those **years**)

静静地睡去
(**Sleep** quietly)

但永远被铭记
(But be remembered forever)

Figure 4: The result of the our model.

how well the models fit the data. The **Rhetoric-F1** score is used to measure the rhetorically controlled accuracy of the generated poem sentences. Specifically, if the rhetoric label of the generated sentence is consistent with the ground truth, the generated result is right, and wrong otherwise. The rhetoric label of each poem sentence is predicted by our rhetoric classifier mentioned above (see 4.1 for details about this classifier). **Distinct-1/Distinct-2** (Li et al., 2016) is used to evaluate the diversity of the generated poems.

Human Evaluation. Following previous work (Yi et al., 2018b), we consider four criteria for human evaluation:

- **Fluency**: Whether the generated poem is grammatically correct and fluent.
- **Coherence**: Whether the generated poem is coherent with the topics and contexts.
- **Meaningfulness**: Whether the generated poem contains meaningful information.
- **Rhetorical Aesthetics**: Whether the generated rhetorical poem has some poetic and artistic beauty.

Each criterion is scored on a 5-point scale ranging from 1 to 5. To build a test set for human evaluation, we randomly select 200 sets of topic words to generate poems with the models. We invite 10 experts[4] to provide scores according to the above criteria and the average score for each criterion is computed.

4.4 Evaluation Results

The results of the automated evaluation are given in Table 3. Our MC model obtains a higher BLEU score and lower perplexity than other baselines on the poetry dataset, which suggests that the model is on a par with other models in generating grammatical sentences. Note that our AC model obtains higher Distinct-1 and Distinct-2 scores because it tends to generate more diverse and informative results.

In terms of the rhetoric generation accuracy, our model outperforms all the baselines and achieves

[4] The experts are Chinese literature students or members of a poetry association.

Rhetoric Type	Examples
Metaphor	**Input**: 光明和暗影交替在你脸面，忽闪出淡红的悠远和蓝色的幽深 (Light and shadows interlace in your face, flashing pale reddish distances and blue depths) **Topic Words**: 恋爱;光明;脸面(Love; Light; Face) **Output**:你的**眼神像**我心灵的花朵一样绽放 (Your **eyes blossom like** flowers in my heart)
Personification	**Input**: 下一次。下一次？改变它，像镜子的客观 (Another time. Another time? Change it, like the objectivity of a mirror) **Topic Words**: 灵魂;镜子;客观(Soul; Mirror; Objectivity) **Output**:它们慢慢地 走来 (They **walked** slowly)
Other	**Input**: 我的话还一句没有出口，蜜蜂的好梦却每天不同 (My words have not spoken, but the bees' dreams are different every day) **Topic Words**: 春天;蜜蜂;梦(Spring; Bees; Dreams) **Output**:我埋怨你的何时才会说完 (I blame you, when will I finish)

Table 5: The result of the rhetoric control.

the best **Rhetoric-F1** score of 0.67 on the poetry dataset, which suggests that our model can control the rhetoric generation substantially more effectively. The other baselines have low scores because they do not possess any direct way to control for rhetoric. Instead, they attempt to learn it automatically from the data, but do not succeed at this particularly well.

Table 4 provides the results of the human evaluation. We observe that on both datasets, our method achieves the best results in terms of the **Meaningfulness** and **Rhetorical Aesthetics** metrics. Additionally, we find that the WM model has higher scores in the **Coherence** metric over the two datasets, indicating that the memory component has an important effect on the coherence and relevance of the topics. The CVAE model obtains the best results in terms of the **Fluency** metric, which shows that this model can generate more fluent sentences, but it lacks coherence and meaningfulness. Overall, our model generates poems better than other baselines in terms of fluency, coherence, meaningfulness, and rhetorical aesthetics. In particular, these results show that a rhetorically controlled encoder-decoder can generate reasonable metaphor and personification in poems.

4.5 Case Study

Table 5 presents example poems generated by our model. These also clearly show that our model can control the rhetoric-specific generation. In Case 1, our model is able to follow the topics 恋爱;脸面 (*love, face*) and the metaphor label when generating the sentence, e.g., 你的眼神像心灵的花朵一样绽放 (*Your eyes blossom like flowers in my heart*). In Case 2, our model obtaining the personification signal is able to generate a personification word 走来 (*walk*).

As an additional case study, we also randomly select a set of topic words {青春 *Youth*, 爱情 *Love*, 岁月 *Years*} and present three five-line poems generated by *Seq2Seq*, *WM*, and our model, respectively, with the same topics and automatically controlled rhetoric. All the poems generated by the different models according to the same topic words are presented in Figures 3 and 4. The poem generated by our model is more diverse and aesthetically pleasing with its use of metaphor and personification, while the two other poems focus more on the topical relevance.

5 Conclusion and Future work

In this paper, we propose a rhetorically controlled encoder-decoder for modern Chinese poetry generation. Our model utilizes a continuous latent variable to capture various rhetorical patterns that govern the expected rhetorical modes and introduces rhetoric-based mixtures for generation. Experiments show that our model outperforms state-of-the-art approaches and that our model can effectively generate poetry with convincing metaphor and personification.

In the future, we will investigate the possibility of incorporating additional forms of rhetoric, such as parallelism and exaggeration, to further enhance the model and generate more diverse poems.

References

Pablo Gervás. 2001. An expert system for the composition of formal spanish poetry. In *Applications and Innovations in Intelligent Systems VIII*, pages 19–32. Springer.

Marjan Ghazvininejad, Yejin Choi, and Kevin Knight. 2018. Neural poetry translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 67–71.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. *Proceedings of ACL 2017, System Demonstrations*, pages 43–48.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines.

Erica Greene, Tugba Bodrumlu, and Kevin Knight. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 524–533.

Jing He, Ming Zhou, and Long Jiang. 2012. Generating chinese classical poems with statistical machine translation models. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Pei Ke, Jian Guan, Minlie Huang, and Xiaoyan Zhu. 2018. Generating informative responses with controlled sentence function. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1499–1508.

Diederik P Kingma and Max Welling. 2014. Autoencoding variational bayes. *stat*, 1050:10.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2016. Autoencoding beyond pixels using a learned similarity metric. In *International Conference on Machine Learning*, pages 1558–1566.

Hung Le, Truyen Tran, Thin Nguyen, and Svetha Venkatesh. 2018. Variational memory encoder-decoder. In *Advances in Neural Information Processing Systems*, pages 1515–1525.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Juntao Li, Yan Song, Haisong Zhang, Dongmin Chen, Shuming Shi, Dongyan Zhao, and Rui Yan. 2018. Generating classical chinese poems via conditional variational autoencoder and adversarial training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3890–3900.

Hisar Manurung. 2004. An evolutionary algorithm approach to poetry generation.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Oriol Vinyals and Quoc V Le. 2015. A neural conversational model.

Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie. 2018. Learning to ask questions in open-domain conversational systems with typed decoders. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2193–2203.

Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 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.

Cheng Yang, Maosong Sun, Xiaoyuan Yi, and Wenhao Li. 2018. Stylistic chinese poetry generation via unsupervised style disentanglement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3960–3969.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

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

Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Zonghan Yang. 2018b. Chinese poetry generation with a working memory model. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, page 4553‘‘4559.

Jiyuan Zhang, Yang Feng, Dong Wang, Yang Wang, Andrew Abel, Shiyue Zhang, and Andi Zhang. 2017. Flexible and creative chinese poetry generation using neural memory. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1364–1373.

Xingxing Zhang and Mirella Lapata. 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.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 654–664.

Xianda Zhou and William Yang Wang. 2018. Mojitalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1128–1137.