Generating Classical Chinese Poems via Conditional Variational Autoencoder and Adversarial Training

Juntao Li[1,2,†], **Yan Song**[3], **Haisong Zhang**[3],
Dongmin Chen[1], **Shuming Shi**[3], **Dongyan Zhao**[1,2], **Rui Yan**[1,2,*]
[1]Beijing Institute of Big Data Research,
Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China
[2]Institute of Computer Science and Technology, Peking University, Beijing, China
[3]Tencent AI Lab
{lijuntao,zhaody,dongminchen,ruiyan}@pku.edu.cn
{clksong,hansonzhang,shumingshi}@tencent.com

Abstract

It is a challenging task to automatically compose poems with not only fluent expressions but also aesthetic wording. Although much attention has been paid to this task and promising progress is made, there exist notable gaps between automatically generated ones with those created by humans, especially on the aspects of term novelty and thematic consistency. Towards filling the gap, in this paper, we propose a conditional variational autoencoder with adversarial training for classical Chinese poem generation, where the autoencoder part generates poems with novel terms and a discriminator is applied to adversarially learn their thematic consistency with their titles. Experimental results on a large poetry corpus confirm the validity and effectiveness of our model, where its automatic and human evaluation scores outperform existing models.

1 Introduction

In mastering concise, elegant wordings with aesthetic rhythms in fixed patterns, classical Chinese poem is a special cultural heritage to record personal emotions and political views, as well as document daily or historical events. Being a fascinating art, writing poems is an attractive task that researchers of artificial intelligence are interested in (Tosa et al., 2008; Wu et al., 2009; Netzer et al., 2009; Oliveira, 2012; Yan et al., 2013, 2016a; Ghazvininejad et al., 2016, 2017; Singh et al., 2017; Xu et al., 2018), partially for the reason that poem generation and its related research could benefit other constrained natural language generation tasks. Conventionally, rule-based models (Zhou et al., 2010) and statistical machine translation (SMT) models (He et al., 2012) are proposed for this task. Recently, deep neural models are employed to generate fluent and natural poems (Wang et al., 2016a; Yan, 2016; Zhang et al., 2017a). Although these models look promising, they are limited in many aspects, e.g., previous studies generally fail to keep thematic consistency (Wang et al., 2016c; Yang et al., 2017) and improve term[1] novelty (Zhang et al., 2017a), which are important characteristics of poems.

In classical Chinese poem composing, thematic consistency and term novelty are usually mutually exclusive conditions to each other, i.e., consistent lines may bring duplicated terms while intriguing choices of characters could result in thematic diversities. On one hand, thematic consistency is essential for poems; it is preferred that all lines concentrate on the same theme throughout a poem. Previous work mainly focused on using keywords (Wang et al., 2016c; Hopkins and Kiela, 2017) to plan a poem so as to generate each line with a specific keyword. Such strategy is risky for the reason that the keywords are not guaranteed consistent in a topic, especially when they are generated or extracted from an inventory (Wang et al., 2016c). On the other hand, Chinese poems are generally short in length, with every character carefully chosen to be concise and elegant. Yet, prior poem generation models with recurrent neural networks (RNN) are likely to generate high-frequency characters (Zhang et al., 2017a), and the resulted poems are trivial and boring. The reason is that RNN tends to be entrapped within local word co-occurrences, they normally fail to capture global characteristic such as topic or hierarchical semantic properties (Bowman et al., 2016).

To address the aforementioned shortcomings, RNN is extended to autoencoder (Dai and Le, 2015) for improving sequence learning, which has

*Corresponding author: Rui Yan (ruiyan@pku.edu.cn)
†Work was partially done at Tencent AI Lab.

[1]We use term and character interchangeably in this paper.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3890–3900
Brussels, Belgium, October 31 - November 4, 2018.

been proven to be appealing in explicitly modeling global properties such as syntactic, semantic, and discourse coherence (Li et al., 2015). Moreover, boosting autoencoder with variational inference (Kingma and Welling, 2014), known as variational autoencoder (VAE), can generate not only consistent but also novel and fluent term sequences (Bowman et al., 2016). To generalize VAE for versatile scenarios, conditional variational autoencoders (CVAE) are proposed to supervise a generation process with certain attributes while maintaining the advantages of VAE. It is verified in supervised dialogue generation (Serban et al., 2017; Shen et al., 2017; Zhao et al., 2017) that CVAE can generate better responses with given dialogue contexts. Given the above background and to align it with our expectations for poem generation, it is worth trying to apply CVAE to create poems. In the meantime, consider that modeling thematic consistency with adversarial training is proven to be promising in controlled text generation (Hu et al., 2017), models for semantic matching can be potentially improved with an explicit discriminator (Wu et al., 2017), so does poem generation.

In this paper, we propose a novel poem generation model (CVAE-D) using CVAE to generate novel terms and a discriminator (D) to explicitly control thematic consistency with adversarial training. To the best of our knowledge, this is the first work of generating poems with the combination of CVAE and adversarial training. Experiments on a large classical Chinese poetry corpus confirm that, through encoding inputs with latent variables and explicit measurement of thematic information, the proposed model outperforms existing ones in various evaluations. Quantitative and qualitative analysis indicate that our model can generate poems with not only distinctive terms, but also consistent themes to their titles.

2 Preliminaries

2.1 VAE and CVAE

In general, VAE consists of an encoder and a decoder, which correspond to the encoding process where input x is mapped to a latent variable z, i.e., $x \mapsto z$, and the decoding process where the latent variable z is reconstructed to the input x, i.e., $z \mapsto x$. In detail, the encoding process computes a posterior distribution $q_\theta(z \mid x)$ given the input x. Similarly, the decoding process can be formulated as $p_\theta(x \mid z)$, representing the probability distribution of generating input x conditioned on z, where z has a regularized prior distribution $p_\theta(z)$, i.e. a standard Gaussian distribution. Herein θ represents the parameters of both encoder and decoder. Importantly, presented by Kingma and Welling (2014), on the condition of large datasets and intractable integral of the marginal likelihood $p_\theta(x)$, the true posterior $q_\theta(z \mid x)$ is simulated by a variational approximation $q_\phi(z \mid x)$ in modeling the encoding process, where ϕ is the parameters for q.

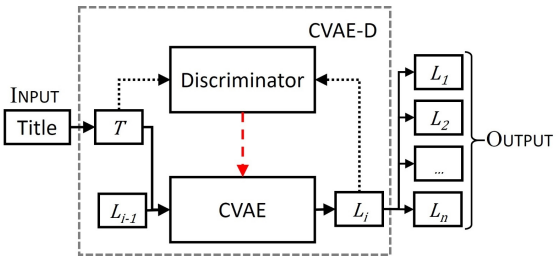


Figure 1: The overall framework of our poem generation model. Solid arrows present the generation process of each line L_i on the condition of the previous line L_{i-1} and title T. Black dotted arrows represent the adversarial learning for thematic consistency. The red dashed arrow refers to the back-propagation of the discriminator to the CVAE.

In learning a VAE, its objective is to maximize the log-likelihood $\log p_\theta(x)$ over input x. To facilitate learning, one can target on pushing up the variational lower bound of $\log p_\theta(x)$:

$$\begin{aligned}\mathbb{L}(\theta, \phi; x) = &- \mathrm{KL}(q_\phi(z|x) \parallel p_\theta(z)) \\ &+ \mathbf{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]\end{aligned} \tag{1}$$

such that the original $\log p_\theta(x)$ is also optimized. Herein the KL-divergence term $\mathrm{KL}(\cdot)$ can be viewed as the regularization for encouraging the approximated posterior $q_\phi(z|x)$ to be close to the prior $p_\theta(z)$, e.g. standard Gaussian distribution. $\mathrm{E}[\cdot]$ is the reconstruction loss conditioned on the approximation posterior $q_\phi(z|x)$, which reflects how well the decoding process goes.

CVAE extends VAE with an extra condition c to supervise the generation process by modifying the. The objective of CVAE is thus to maximize the reconstruction log-likelihood of the input x under the condition of c. Following the operation for VAE, we have the corresponding variational lower bound of $p_\theta(x|c)$ formulated as

$$\begin{aligned}\mathbb{L}(\theta, \phi; x, c) = &- \mathrm{KL}(q_\phi(z|x, c) \parallel p_\theta(z|c)) \\ &+ \mathbf{E}_{q_\phi(z|x,c)}[\log p_\theta(x|z, c)]\end{aligned} \tag{2}$$

which is similar to Eq.1 except that all items are introduced with c, such as $q_\phi(z|x, c)$ and $p_\theta(z|c)$,

referring to the conditioned approximate posterior and the conditioned prior, respectively.

2.2 Problem Formulation

Following the text-to-text generation paradigm (Ranzato et al., 2015; Kiddon et al., 2016; Hu et al., 2017; Ghosh et al., 2017), our task has a similar problem setting with conventional studies (Zhang and Lapata, 2014; Wang et al., 2016c), where a poem is generated in a line-by-line manner that each line serves as the input for the next one, as illustrated in Figure 1. To formulate this task, we separate its input and output with necessary notations as follows.

The **INPUT** of the entire model is a title, $T = (e_1, e_2, \ldots, e_N)$, functionalized as the theme of the target poem[2], where e_i refers to i-the character's embedding and N is the length of the title. The first line L_1 is generated only conditioned on the title T, once this step is done, the model takes the input of the previous generated line as well as the title at each subsequent step, until the entire poem is completed.

The overall **OUTPUT** is an n-line poem, formulated as $(L_1, L_2, \ldots, L_n)$, where $L_i = (e_{i,1}, e_{i,2}, \ldots, e_{i,m})$ denotes each line in the poem, with $e_{i,j}$ referring to the embedding of a character at i-th line on j-th position, $\forall\ 1 \leq i \leq n$, $1 \leq j \leq m$. Particularly for classic Chinese poems, there are strict patterns, which require $m = 5$ or $m = 7$, and $n = 4$[3] or $n = 8$[4]. Once a template is chosen, m and n are fixed. In this paper, we mainly focus on $n = 4$.

3 The Model

As illustrated in Figure 1, our CVAE-D consists of two parts, CVAE and a discriminator, where their details are elaborated in the following subsections.

3.1 The CVAE

The CVAE includes an encoder and a decoder, plays as the core part in our model that generates classic Chinese poems. The encoder encodes both the title and lines with shared parameters by a bidirectional RNN (Schuster and Paliwal, 1997) with gated recurrent units (GRU) (Chung et al., 2014). Through the encoder, at each step, the previous line L_{i-1}, current line L_i and title T are represented as concatenated forward and backward[5] vectors $h_{i-1} = [\overrightarrow{h}_{i-1}, \overleftarrow{h}_{i-1}]$, $h_i = [\overrightarrow{h}_i, \overleftarrow{h}_i]$ and $t = [\overrightarrow{t}, \overleftarrow{t}]$, respectively. Note that h_i corresponds to x, while the concatenation of h_{i-1} and t functionalized as c in Eq. 2, i.e., $c = [h_{i-1}, t]$. Following previous work (Kingma and Welling, 2014; Zhao et al., 2017; Yang et al., 2017), we assume that the variational approximate posterior is a multivariate Gaussian $\mathcal{N}$ with a diagonal covariance structure $q_\phi(z|x,c) = \mathcal{N}(\mu, \sigma^2 \boldsymbol{I})$. Thus μ and σ are the key parameters[6] to be learned, and they are computed by

$$\begin{bmatrix} \mu \\ \log(\sigma^2) \end{bmatrix} = W_q \begin{bmatrix} x \\ c \end{bmatrix} + b_q \tag{3}$$

where W_q and b_q are trainable parameters. Similarly, the prior $p_\theta(z|c)$ can be formulated as another multivariate Gaussian $\mathcal{N}(\mu', \sigma'^2 \boldsymbol{I})$; its parameters are then calculated by a single-layer fully-connected neural network (denoted as MLP) with the $tanh(\cdot)$ activation function,

$$\begin{bmatrix} \mu' \\ \log(\sigma'^2) \end{bmatrix} = \mathrm{MLP}_p(c) \tag{4}$$

The decoder uses a one-layer RNN with GRU that takes $[z, c]$ as the input to predict each line L_i. The hidden states of the GRU, $(s_1, s_2, \ldots, s_m)$[7],

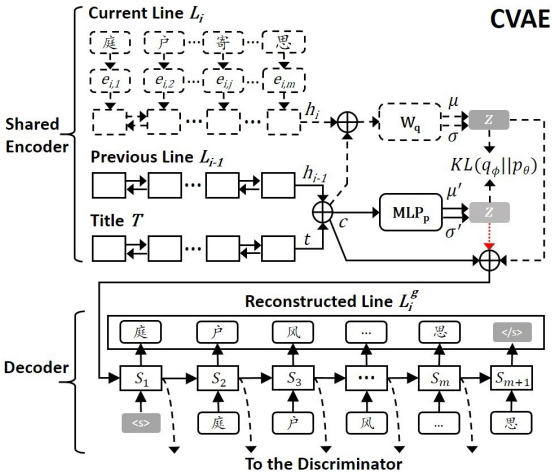


Figure 2: The CVAE for poem generation. $\oplus$ denotes the vector concatenation operation. Only the part with solid lines and the red dotted arrow is applied in prediction, while the entire CVAE is used in training process except the red dotted arrow part.

[2] We directly treat the title as the theme for each poem in this paper instead of transferring it to a few keywords as that was done in Yang et al. (2017).

[3] The quatrain.

[4] The eight-line regulated verse.

[5] $\rightarrow$ and $\leftarrow$ refer to forward and backward, respectively.

[6] μ and σ^2 represent the mean and variance of $\mathcal{N}(\mu, \sigma^2 \boldsymbol{I})$.

[7] Note that s_{m+1} is not passed to the discriminator.

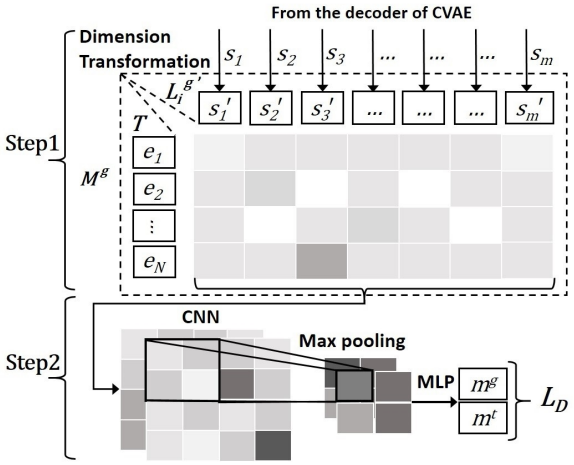


Figure 3: The Discriminator.

are not only used to generate the reconstructed lines, but also passed to the discriminator for learning thematic consistency.

The entire encoder and the decoder are used throughout the training process, with only part of the encoder (objects with solid lines in Figure 2) and the decoder applied in prediction. It is worth noting that, θ and ϕ mentioned in §2.1 are not explicitly corresponded to any particular neural networks described in this section. Instead, the probability process denoted by θ corresponds to the decoding and part of the encoding process, so does ϕ, i.e., $\phi = \{W_q, b_q\}$.

3.2 The Discriminator

The discriminator is introduced in our model to evaluate thematic consistency between the input title and the generated poem lines. The loss from this discriminator is then back-propagated to the decoder of the CVAE to enhance its training. In this paper, we employ a procedure that consists of two steps. First, we compute an interaction (or matching) matrix according to a generated line ${L_i}^g$ and the title T, where ${L_i}^g$ is the reconstructed result of L_i. Then, we utilize a convolutional neural network (CNN) to learn the matching score between ${L_i}^g$ and T, where the score is interpreted as the degree of thematic consistency. Specifically, in the discriminator, we treat ${L_i}^g$ and L_i as the negative and positive instance, referring to thematically inconsistent and consistent case, respectively.

In detail, for the first step, we use the state sequence[8] of the decoder to represent ${L_i}^g$, i.e., ${L_i}^g = (s_1, s_2, \ldots, s_m)$. A dimension transformation is then conducted on ${L_i}^g$, to align ${L_i}^g$ and T:

$$s_i^{'} = ReLU(W_d s_i + b_d) \tag{5}$$

where $ReLU$ is the rectified linear units activation function (Nair and Hinton, 2010), with trainable parameters W_d and b_d. In doing so, the dimension of $s_i^{'}$ is identical to character embeddings. The transformed line is then denoted as ${L_i}^{g'} = (s_1^{'}, s_2^{'}, \ldots, s_m^{'})$. Thus the interaction matrix between ${L_i}^{g'}$ and T is then formulated as

$$M^g = {L_i}^{g'} \cdot T \tag{6}$$

where $M^g \in \mathbb{R}^{N \times m}$; "$\cdot$" denotes the matrix multiplication.

In the second step, a CNN is used to extract features from the interaction matrix. The resulted feature matrix is calculated by $\mathcal{F} = CNN(M^g)$. Then, we apply a max-overtime pooling (Collobert et al., 2011) over $\mathcal{F}$ to capture the most salient information. After this operation, an MLP with one hidden layer is used to flatten the feature matrix and generate the final matching score $m^g \in (0, 1)$ via a $sigmoid$ activation function.

In addition to m^g, the matching score m^t between the positive sample L_i and T is computed in a process similar to the above procedure, except the dimension transformation because character embeddings in both title T and L_i share the same dimension.[9]

Finally, following the routine of generative adversarial networks (GAN) (Goodfellow et al., 2014), the discriminator is trained to measure the thematic consistency of generated lines and the ground truth lines according to the matching scores m^g and m^t, with the objective function

$$\mathcal{L}_D = \log(m^g) + \log(1 - m^t) \tag{7}$$

minimized. Note that the discriminator is only applied during the training process, where the parameters of the encoder and decoder are enhanced by the feedback of the discriminator.

[8] Following previous work (Goyal et al., 2016; Hu et al., 2017) using adversarial training, using state sequence instead of the outputs is because the discrete nature of the outputs hinders gradient calculation.

[9] Different from ${L_i}^g$, L_i is represented directly by its sequence of character embeddings, for the reason that the discriminator is only connected with the decoder while L_i does not go through it. Otherwise, if the encoder states are passed to the discriminator, the loss would be back-propagated to the encoder and disturb CVAE training accordingly.

	Poem #	Line #	Vocab #	Token #
PTD	56,549	371,754	7,685	2,093,740
PSD	253,237	1,497,348	9,959	9,008,418
Total	309,786	1,869,102	10,306	11,102,158

Table 1: Corpus statistics of PTD and PSD. **Vocab #** and **Token #** refer to vocabulary size and total number of tokens, respectively, in terms of character.

3.3 Training the Model

The overall objective of CVAE-D is to minimize

$$\mathcal{L}_{\text{CVAE-D}} = \mathcal{L}_{CVAE} - \lambda \mathcal{L}_D \tag{8}$$

with respect to parameters of the CVAE, where $\mathcal{L}_{CVAE}$ is the loss of CVAE, corresponding to $-\mathbb{L}(\theta, \phi; x, c)$. In doing so, $\mathcal{L}_D$ is maximized with regard to parameters of the discriminator, referring to that the generated poems are thematic consistent and able to confuse the discriminator. Herein λ is a balancing parameter. We train the CVAE and the discriminator alternatively in a two-step adversarial fashion similar to that was done in Zhang et al. (2017c). This training strategy is repeated until the $\mathcal{L}_{\text{CVAE-D}}$ is converged.

4 Experiment Setup

4.1 Datasets

To learn our poem generation model, we collect two corpora for experiments: a collection of classic Chinese poems from Tang dynasty (PTD), and the other from Song dynasty (PSD). Statistics of the two corpora are reported in Table 1. Note that for classical Chinese poem, the dominant genres are quatrain and eight-line regulated verse with either 5 or 7 characters in each line. As a result, our model is targeted to generate poems within these two genres, especially the quatrain. All titles of poems are treated as their themes. We randomly choose 1,000 and 2,000 poems for validation and test, respectively, with the rest poems for training.

4.2 Baselines

In addition to our CVAE-D, several highly related and strong methods are conducted as baselines in our experiments, including:

S2S, the conventional sequence-to-sequence model (Sutskever et al., 2014), which has proven to be successful in neural machine translation (NMT) and other text generation tasks.

AS2S and its extension **Key-AS2S** and **Mem-AS2S**, where AS2S is the S2S model integrated with attention mechanism (Bahdanau et al., 2014). Key-AS2S and Mem-AS2S are AS2S with keywords planning (Wang et al., 2016c) and a memory module (Zhang et al., 2017a), respectively. Particularly, they are dedicated models designed for Chinese poem generation.

Criterion	Description
Consistency	Whether a poem displays a consistent theme.
Fluency	Whether a poem is grammatically satisfied.
Meaning	How meaningful the content of a poem is.
Poeticness	Whether a poem has the attributes of poetry.
Overall	Average scores of the above four criteria.

Table 2: Human evaluation criteria.

GAN, a basic implementation of generative adversarial networks (Goodfellow et al., 2014) for this task on top of S2S. This baseline is added to investigate the performance of introducing a discriminator to simple structures other than CVAE.

CVAE[10] and its extension **CVAE-Key**, where the former is the conventional CVAE model and the latter refers to the combination of CVAE and keywords planning (Yang et al., 2017). The CVAE baseline is used for investigating how poem generation can be done with only CVAE, while CVAE-Key aims to provide a comparison to our model with a different technique for thematic control.

4.3 Model Settings

All baselines and the CVAE-D are trained with the following hyper-parameters. The dimension of character embedding is set to 300 for the most frequent 10,000 characters in our vocabulary. The hidden state sizes of the GRU encoder and decoder are set to 500. All trainable parameters, e.g., W_q and W_d, are initialized from a uniform distribution $[-0.08, 0.08]$. We set the mini-batch size to 80 and employ the Adam (Kingma and Ba, 2014) for optimization. We utilize the gradient clipping strategy (Pascanu et al., 2013) to avoid gradient explosion, with the gradient clipping value set to 5.

In addition to the shared hyper-parameters, we have particular settings for CVAE-D. The layer size of MLP_p is set to 400. The dimension of latent variable z is set to 300. For the CNN used in the discriminator, its kernel size is set to $(5, 5)$, with the stride size k to 2. We follow the conventional setting (Hu et al., 2017; Creswell et al.,

[10] We do not include VAE as our baseline since VAE cannot perform a supervised generation process.

	Automatic Evaluation							Human Evaluation				
Model	**BLEU-1**	**BLEU-2**	**Sim**	**Dist-1**	**Dist-2**	**Dist-3**	**Dist-4**	**Con.**	**Flu.**	**Mea.**	**Poe.**	**Ovr.**
S2S	13.8	2.48	14.7	2.50	16.2	34.9	50.0	1.79	1.84	1.71	1.60	1.74
AS2S	15.5	2.59	14.8	2.30	15.2	31.4	44.3	1.92	1.71	1.80	1.74	1.79
Key-AS2S	15.8	1.92	19.8	3.00	16.3	33.0	45.6	2.21	2.15	1.92	2.23	2.13
MeM-AS2S	16.0	1.48	22.0	3.40	51.4	87.9	96.8	1.70	2.23	2.09	2.89	2.23
GAN	17.7	2.54	22.5	2.50	16.8	35.3	49.6	2.36	2.08	2.01	2.08	2.13
CVAE	17.0	1.73	13.7	4.70	52.3	90.6	99.0	1.69	2.16	2.14	2.58	2.14
CVAE-Key	16.4	1.83	31.0	4.31	43.0	80.6	95.8	1.83	2.29	2.08	2.53	2.18
CVAE-D	**18.1**	**2.85**	**36.3**	**5.20**	**59.2**	**94.2**	**99.8**	**2.58**	**2.35**	**2.34**	**2.96**	**2.56**

Table 3: Results of automatic and human evaluations. **BLEU-1** and **BLEU-2** are BLEU scores on unigrams and bigrams ($p < 0.01$); **Sim** refer to the similarity score; **Dist-**n corresponds to the distinctness of n-gram, with $n = 1$ to 4; **Con., Flu., Mea., Poe., Ovr.** represent consistency, fluency, meaning, poeticness, and overall, respectively.

2017) to set the balancing parameter λ to 0.1.[11]

4.4 Evaluation Metrics

To comprehensively evaluate the generated poems, we employ the following metrics:

BLEU: The BLEU score (Papineni et al., 2002) is an effective metric, widely used in machine translation, for measuring word overlapping between ground truth and generated sentences. In poem generation, BLEU is also utilized as a metric in previous studies (Zhang and Lapata, 2014; Wang et al., 2016a; Yan, 2016; Wang et al., 2016b). We follow their settings in this paper.

Similarity: For thematic consistency, it is challenging to automatically evaluate different models. We adopt the embedding average metric to score sentence-level similarity as that was applied in Wieting et al. (2015). In this paper, we accumulate the embeddings of all characters from the generated poems and that from the given title, and use $cosine$ to compute the similarity between the two accumulated embeddings.

Distinctness: As an important characteristic, poems use novel and unique characters to maintain their elegance and delicacy. Similar to that proposed for dialogue systems (Li et al., 2016), this evaluation is employed to measure character diversity by calculating the proportion of distinctive [1,4]-grams[12] in the generated poems, where final distinctness values are normalized to [0,100].

Human Evaluation: Since writing poems is a complicated task, there always exist incoordinations between automatic metrics and human experiences. Hence, we conduct human evaluation to assess the performance of different models. In doing so, each poem is assessed by five annotators who are well educated and have expertise in Chinese poetry. The evaluation is conducted in a blind review manner, where each annotator has no information about the generation method that each poem belongs to. Following previous work (He et al., 2012; Zhang and Lapata, 2014; Wang et al., 2016c; Zhang et al., 2017a), we evaluate generated poems by four criteria, namely, consistency, fluency, meaning, and poeticness. Each criterion is rated from 1 to 3, representing bad, normal, good, respectively. The details are illustrated in Table 2.

5 Experimental Results

5.1 Quantitative Analysis

Table 3 reports the results of both automatic and human evaluations. We analyze the results from the following aspects.

5.1.1 The effect of CVAE

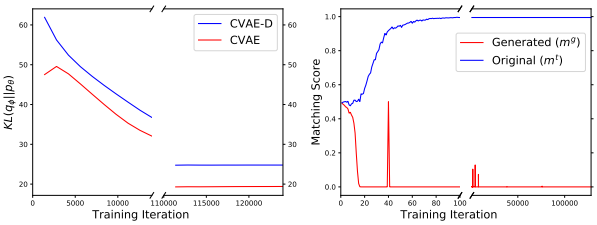
This study is to investigate whether using latent variable and variational inference can improve the diversity and novelty of terms in generated poems. There are two main observations.

CVAE significantly improves term novelty. As illustrated in Table 3, CVAE outperforms all baselines significantly in terms of distinctness. With diversified terms, the aesthetics scores also confirm that CVAE can generate poems that correspond to better user experiences. Although Mem-AS2S can generate a rather high distinctness score, it requires a more complicated structure in learning and generating poems. The results confirm the effectiveness of CVAE in addressing the issue of term duplications that occurred in RNN.

CVAE cannot control thematic consistency of generated poems. Recall that thematic consistency and term diversity are usually mutually ex-

[11] We tried different values for λ, varying from 0.001 to 1, which result in similar performance of the CVAE-D.

[12] Defined as the number of distinctive n-grams divided by the total number of n-grams, shown as Dist-1, Dist-2, Dist-3, Dist-4 in Table 3.



(a) KL divergence (b) Matching score

Figure 4: KL divergences of CVAE and CVAE-D (a) and matching scores of the generated and original lines from the discriminator (b). All curves are drawn against training iterations.

clusive, CVAE produces the worst result in thematic consistency, which is confirmed in Table 3 by the similarity score in automatic evaluation and the consistency score in human evaluation.

5.1.2 The Influence of the Discriminator

As previously stated, introducing a discriminator with adversarial training is expected to bring positive effect on thematic consistency. We investigate the influence of discriminator with two groups of comparison, i.e., CVAE-D v.s. CVAE, GAN v.s. S2S. Following observations are made in this investigation, which confirm that adversarial learning is an effective add-on to existing models for thematics control, without affecting other aspects.

The discriminator effectively enhances poem generation with thematic information. When the discriminator is introduced, CVAE and S2S model are capable of generating thematically consistent poems, as illustrated by the similarity and meaning scores in Table 3. The BLEU results also confirm that the discriminator can improve the overlapping between generated poems and the ground truth, which serves as thematic consistent cases.

The extra discriminator does not affect base models on irrelevant merits. For any base model, e.g., S2S and CVAE, when adding a discriminator, it is expected that it can bring help on thematic consistency while limiting any inferior effects on other evaluations. This is confirmed in the results, e.g., for distinctness, CVAE-D and GAN are comparable to CVAE and S2S.

5.1.3 The Performance of CVAE-D

Overall, the CVAE-D model substantially outperforms all other models in all metrics. Especially for term novelty and thematic consistency, CVAE-D illustrates an extraordinary balance between them, with observable improvements on both sides. This balance is mainly contributed

书窗碧桃
Daydream in my garden
庭户风光寄所思，
The view in the garden brings up the fantasy,
伊人重过惜残枝。
As if my love dances in the scenery.
窗前花开不知味，
Hence blossom can never arouse my curiosity,
唯有落红入我诗。
With only fading memory in the poetry.

Figure 5: An example poem generated by the CVAE-D model. Note that the translation is performed in delivering the meaning instead of the verbatim manner.

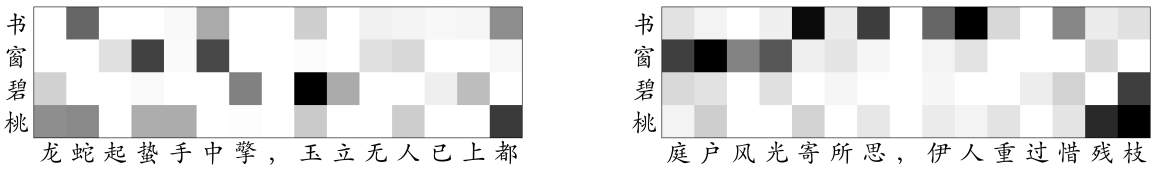
from the proposed framework that seamlessly integrates CVAE and the discriminator. Except for the automatic and human evaluation scores, the fact is also supported by the training loss of $\mathrm{KL}(q_\phi(z|x,c) \parallel p_\theta(z|c))$ and $\mathcal{L}_\mathcal{D}$ as shown in Figure 4, where 1) the KL-divergence of CVAE-D has an analogous trend with CVAE, referring to that the CVAE part in CVAE-D is trained as good as an independent CVAE; 2) the discriminator captures the distinctness of thematic consistency between the generated lines and the ground truth lines at the very early stage of training.

5.2 Qualitative Analysis

In addition to evaluating CVAE-D with quantitative results, we also conduct case studies to illustrate its superiority. Figure 5 gives an example of the CVAE-D generated poems, which well demonstrates the capability of our model. The entire poem elegantly expresses a strong theme of "missing my love".[13] It is clearly shown that the choices of the characters, such as 庭 (yard), 枝 (branch), 花 (flower), 红 (red), etc., match with the given title to a certain extent with no one repetitively used. To further investigate how different models perform on thematic consistency, we visualize the correspondence between generated poems (the first two lines) and the given title with heatmaps in Figure 6, where Figure 6(a) and Figure 6(b) illustrate the results yielded by CVAE and CVAE-D, respectively.[14] Obviously, the overall color in Figure 6(a) is lighter than that in Figure 6(b), which

[13] "Seeing an object makes one miss someone" is a popular theme in Classical Chinese poems.

[14] Grids in the heatmap represent the correlations between the fine-tuned embeddings of the characters in the title and the generated lines. Since the embeddings are updated in the training process, a better model leads to higher correlations among the embeddings of related characters.



(a) CVAE (b) CVAE-D

Figure 6: Heatmaps derived from CVAE (a) and CVAE-D (b), in illustrating the correlation between the characters in lines and the title. The horizontal axis refers to characters of the first two lines generated by different models; the vertical axis corresponds to characters in the title. Darker color indicates higher thematic consistency.

may indicate that most of the characters generated by CVAE are not addressed with thematic attentions over the given title. On the opposite, CVAE-D presents darker color in the grids on all related characters, which further reveals the effectiveness of CVAE-D in improving thematic consistency of a poem with respect to its title.

It is observed that there are also inferior cases generated by our model. A notable example pattern is that some fine-grained attributes, e.g., sentiment, emotion, are not well aligned across lines, where some lines may deliver different mood from others. Since our model does not explicitly control such attributes, thus one potential solution to address this issue is to introduce other features to model such information, which requires a special design to adjust the current model. We also notice there exists a few extraordinary bad cases where their basic characteristics, such as wording, fluency, etc., are unacceptable. This phenomenon is randomly observed with no patterns, which could be explained by the complexity of the model and the fragile natural of adversarial training (Goodfellow et al., 2014; Li et al., 2017). Careful parameter setting and considerate module assemble could mitigate this problem, thus lead to potential future work of designing more robust frameworks.

6 Related Work

Deep Generative Models. This work can be seen as an extension of research on deep generative models (Salakhutdinov and Hinton, 2009; Bengio et al., 2014), where most of the previous work, including VAE and CVAE, focused on image generation (Sohn et al., 2015; Yan et al., 2016b). Since GAN (Goodfellow et al., 2014) is also a successful generative model, there are studies tried to integrate VAE and GAN (Larsen et al., 2016). In natural language processing, many recent deep generative models are applied to dialogue systems Serban et al. (2017); Shen et al. (2017); Zhao et al. (2017) and text generation with (Hu et al., 2017; Yu et al., 2017; Lin et al., 2017; Zhang et al., 2017b; Guo et al., 2018). To the best of our knowledge, this work is the first one integrating CVAE and adversarial training with a discriminator for text generation, especially in a particular text genre, poetry.

Automatic Poem Generation. According to methodology, previous approaches can be roughly classified into three categories: 1) rule and template based methods (Tosa et al., 2008; Wu et al., 2009; Netzer et al., 2009; Zhou et al., 2010; Oliveira, 2012; Yan et al., 2013); 2) SMT approaches (Jiang and Zhou, 2008; Greene et al., 2010; He et al., 2012); 3) deep neural models (Zhang and Lapata, 2014; Wang et al., 2016b; Yan, 2016). Compared to rule-based and SMT models, neural models are able to learn more complicated representations and generate smooth poems. Most recent studies followed this paradigm. For example, Wang et al. (2016c) proposed a modified encoder-decoder model with keyword planning; Zhang et al. (2017a) adopted memory-augmented RNNs to dynamically choose each term from RNN output or a reserved inventory. To improve thematic consistency, Yang et al. (2017) combined CVAE and keywords planning. Compared to them, our approach offers an alternative way for poem generation that can produce novel terms and consistent themes via an integrated framework, without requiring special designed modules or post-processing steps.

7 Conclusions

In this paper, we proposed an effective approach that integrates CVAE and adversarial training for classical Chinese poem generation. Specifically, we used CVAE to generate each line of a poem with novel and diverse terms. A discriminator was then applied with adversarial training to explicitly control thematic consistency. Experiments conducted on a large Chinese poem corpus illus-

trated that through the proposed architecture with CVAE and the discriminator, substantial improvement was observed on the results from our generated poems over those from the existing models. Further qualitative study on given examples and some brief error analyses also confirmed the validity and effectiveness of our proposed approach.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (No. 2017YFC0804001), the National Science Foundation of China (NSFC No. 61672058; NSFC No. 61876196). Rui Yan was sponsored by CCF-Tencent Open Research Fund and Microsoft Research Asia (MSRA) Collaborative Research Program.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

Yoshua Bengio, Eric Thibodeau-Laufer, Guillaume Alain, and Jason Yosinski. 2014. Deep Generative Stochastic Networks Trainable by Backprop. In *International Conference on Machine Learning*, pages 226–234, Beijing, China.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating Sentences from a Continuous Space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv preprint arXiv:1412.3555*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Antonia Creswell, Anil A Bharath, and Biswa Sengupta. 2017. Conditional Autoencoders with Adversarial Information Factorization. *arXiv preprint arXiv:1711.05175*.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised Sequence Learning. In *Advances in Neural Information Processing Systems*, pages 3079–3087, Montreal, Canada.

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, Austin,USA.

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.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A Neural Language Model for Customizable Affective Text Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 634–642, Vancouver, Canada.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in neural information processing systems*, pages 2672–2680, Montreal, Canada.

Anirudh Goyal, Alex Lamb, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. 2016. Professor Forcing: A new Algorithm for Training Recurrent Networks. In *Advances In Neural Information Processing Systems*, pages 4601–4609, Barcelona, Spain.

Erica Greene, Tugba Bodrumlu, and Kevin Knight. 2010. Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation. In *Conference on Empirical Methods in Natural Language Processing*, pages 524–533, Massachusetts, USA.

Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2018. Long Text Generation via Adversarial Training with Leaked Information. *AAAI*.

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*, pages 1650–1656, Toronto, Canada.

Jack Hopkins and Douwe Kiela. 2017. Automatically Generating Rhythmic Verse with Neural Networks. In *Meeting of the Association for Computational Linguistics*, pages 168–178, Vancouver, Canada.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward Controlled Generation of Text. In *International Conference on Machine Learning*, pages 1587–1596, Sydney, Australia.

Long Jiang and Ming Zhou. 2008. Generating Chinese Couplets using a Statistical MT Approach. In *COLING 2008, International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, Uk*, pages 377–384, Manchester, United Kingdom.

Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally Coherent Text Generation with Neural Checklist Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339, Austin, USA.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P Kingma and Max Welling. 2014. Auto-Encoding 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, New York, USA.

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, San Diego, USA.

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A Hierarchical Neural Autoencoder for Paragraphs and Documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 1106–1115, Beijing, China.

Jiwei Li, Will Monroe, Tianlin Shi, Sèbastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial Learning for Neural Dialogue Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen, Denmark.

Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. 2017. Adversarial Ranking for Language Generation. In *Advances in Neural Information Processing Systems*, pages 3155–3165.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th international conference on machine learning*, pages 807–814, Haifa, Israel.

Yael Netzer, David Gabay, Yoav Goldberg, and Michael Elhadad. 2009. Gaiku: Generating Haiku with Word Associations Norms. *Computational Approaches to Linguistic Creativity*, page 32.

Hugo Gonçalo Oliveira. 2012. PoeTryMe: a Versatile Platform for Poetry Generation. *Computational Creativity, Concept Invention, and General Intelligence*, 1:21.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL*, pages 311–318, Philadelphia, USA.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the Difficulty of Training Recurrent Neural Networks. In *International Conference on Machine Learning*, pages 1310–1318, Atlanta,USA.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence Level Training with Recurrent Neural Networks. *arXiv preprint arXiv:1511.06732*.

Ruslan Salakhutdinov and Geoffrey E Hinton. 2009. Deep Boltzmann Machines. In *AISTATS*, pages 448–455.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *AAAI*, pages 3295–3301, San Francisco, USA.

Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A Conditional Variational Framework for Dialog Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 504–509, Vancouver, Canada.

Divya Singh, Margareta Ackerman, and Rafael Pérez y Pérez. 2017. A Ballad of the Mexicas: Automated Lyrical Narrative Writing. In *Eighth International Conference on Computational Creativity, ICCC, Atlanta*.

Kihyuk Sohn, Xinchen Yan, and Honglak Lee. 2015. Learning Structured Output Representation using Deep Conditional Generative Models. In *International Conference on Neural Information Processing Systems*, pages 3483–3491, Montreal, Canada.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in neural information processing systems*, pages 3104–3112, Montreal, Canada.

Naoko Tosa, Hideto Obara, and Michihiko Minoh. 2008. Hitch Haiku: An Interactive Supporting System for Composing Haiku Poem. In *International Conference on Entertainment Computing*, pages 209–216. Springer.

Qixin Wang, Tianyi Luo, and Dong Wang. 2016a. Can Machine Generate Traditional Chinese Poetry? A Feigenbaum Test. In *International Conference on Brain Inspired Cognitive Systems*, pages 34–46.

Qixin Wang, Tianyi Luo, Dong Wang, and Chao Xing. 2016b. Chinese Song Iambics Generation with Neural Attention-based Model. In *International Joint Conference on Artificial Intelligence*, pages 2943–2949, New York, 2016.

Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 2016c. 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, Osaka, Japan.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards Universal Paraphrastic Sentence Embeddings. *arXiv preprint arXiv:1511.08198*.

Xiaofeng Wu, Naoko Tosa, and Ryohei Nakatsu. 2009. New Hitch Haiku: An Interactive Renku Poem Composition Supporting Tool Applied for Sightseeing Navigation System. In *International Conference on Entertainment Computing*, pages 191–196. Springer.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 496–505, Vancouver, Canada.

Linli Xu, Liang Jiang, Chuan Qin, Zhe Wang, and Dongfang Du. 2018. How Images Inspire Poems: Generating Classical Chinese Poetry from Images with Memory Networks. *arXiv preprint arXiv:1803.02994*.

Rui Yan. 2016. i, Poet: Automatic Poetry Composition through Recurrent Neural Networks with Iterative Polishing Schema. In *International Joint Conference on Artificial Intelligence*, pages 2238–2244, New York, USA.

Rui Yan, Han Jiang, Mirella Lapata, Shou De Lin, Xueqiang Lv, and Xiaoming Li. 2013. i, Poet: Automatic Chinese Poetry Composition through a Generative Summarization Framework under Constrained Optimization. In *International Joint Conference on Artificial Intelligence*, pages 2197–2203, Beijing, China.

Rui Yan, Cheng-Te Li, Xiaohua Hu, and Ming Zhang. 2016a. Chinese couplet generation with neural network structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2347–2357.

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2016b. Attribute2image: Conditional Image Generation from Visual Attributes. In *European Conference on Computer Vision*, pages 776–791, Amsterdam, Netherlands. Springer.

Xiaopeng Yang, Xiaowen Lin, Shunda Suo, and Ming Li. 2017. Generating Thematic Chinese Poetry using Conditional Variational Autoencoder. *arXiv preprint arXiv:1711.07632*.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In *AAAI*, pages 2852–2858, San Francisco, USA.

Jiyuan Zhang, Yang Feng, Dong Wang, Yang Wang, Andrew Abel, Shiyue Zhang, and Andi Zhang. 2017a. Flexible and Creative Chinese Poetry Generation Using Neural Memory. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1364–1373, Vancouver, Canada.

Xingxing Zhang and Mirella Lapata. 2014. Chinese Poetry Generation with Recurrent Neural Networks. In *Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Doha, Qatar.

Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. 2017b. Adversarial Feature Matching for Text Generation. In *International Conference on Machine Learning*, pages 4006–4015.

Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. 2017c. Aspect-augmented Adversarial Networks for Domain Adaptation. *Transactions of the Association of Computational Linguistics*, 5(1):515–528.

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, pages 654–664, Vancouver, Canada.

Chang Le Zhou, Wei You, and Xiao Jun Ding. 2010. Genetic Algorithm and Its Implementation of Automatic Generation of Chinese SONGCI. *Journal of Software*, 21(3):427–437.