# Plan-CVAE: A Planning-based Conditional Variational Autoencoder for Story Generation

**Lin Wang**[1,2], **Juntao Li**[1,2], **Dongyan Zhao**[1,2], **Rui Yan**[1,2]*
[1]Center for Data Science, Academy for Advanced Interdisciplinary Studies,
Peking University, Beijing, China
[2]Wangxuan Institute of Computer Technology, Peking University, Beijing, China
{wanglin,lijuntao,zhaody,ruiyan}@pku.edu.cn

## Abstract

Story generation is a challenging task of automatically creating natural languages to describe a sequence of events, which requires outputting text with not only a consistent topic but also novel wordings. Although many approaches have been proposed and obvious progress has been made on this task, there is still a large room for improvement, especially for improving thematic consistency and wording diversity. To mitigate the gap between generated stories and those written by human writers, in this paper, we propose a planning-based conditional variational autoencoder, namely Plan-CVAE, which first plans a keyword sequence and then generates a story based on the keyword sequence. In our method, the keywords planning strategy is used to improve thematic consistency while the CVAE module allows enhancing wording diversity. Experimental results on a benchmark dataset confirm that our proposed method can generate stories with both thematic consistency and wording novelty, and outperforms state-of-the-art methods on both automatic metrics and human evaluations.

## 1 Introduction

A narrative story is a sequence of sentences or words which describe a logically linked set of events (Mostafazadeh et al., 2016). Automatic story generation is a challenging task since it requires generating texts which satisfy not only thematic consistency but also wording diversity. Despite that considerable efforts have been made in the past decades, the requirement of thematic consistency and wording diversity is still one of the main problems in the task of story generation.

On the one hand, a well-composed story is supposed to contain sentences that are tightly connected with a given theme. To address this problem, most previous methods attempt to learn mid-level representations, such as events (Martin et al., 2018), prompts (Fan et al., 2018), keywords (Yao et al., 2019) or actions (Fan et al., 2019), to guide the sentences generation. Although these approaches have shown their encouraging effectiveness in improving the thematic consistency, most of them have no guarantee for the wording diversity. The main reason is that most of these methods are based on recurrent neural networks (RNNs), which tend to be entrapped within local word co-occurrences and cannot explicitly model holistic properties of sentences such as topic (Bowman et al., 2016; Li et al., 2018; Li et al., 2019). As a result, RNN tends to generate common words that appear frequently (Zhao et al., 2017) and this will lead to both high inter- and intra-story content repetition rates.

On the other hand, a well-composed story also needs to contain vivid and diversified words. To address the issue of wording diversity, some studies have employed models based on variational autoencoder (VAE) (Kingma and Welling, 2013) or conditional variational autoencoder (CVAE) as a possible solution. It has been proved that, through learning distributed latent representation of the entire sentences, VAE can capture global features such as topics and high-level syntactic properties, and thus can generate novel word sequences by preventing entrapping into local word co-occurrences (Bowman et al., 2016). As a modification of VAE, CVAE introduces an extra condition to supervise the generating process and

*Corresponding author: Rui Yan (ruiyan@pku.edu.cn).


Proceedings of the 19th China National Conference on Computational Linguistics, pages 892–902,
Hainan, China, October 30 – November 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

has been used in multiple text generation tasks, e.g., dialogue response generation (Zhao et al., 2017), Chinese poetry generation (Yang et al., 2018). Recent researches (Li et al., 2019; Wang and Wan, 2019) in story generation task have confirmed that CVAE can generate stories with novel wordings. Despite the promising progress, how to keep thematic consistency while improving wording diversity is still a challenging problem, since these two requirements are to some extent mutually exclusive (Li et al., 2019). Specifically, consistent stories may limit the choice of words, while diversified wordings may lead to the risk of inconsistent themes.

In this paper, we propose to conquer these two challenges simultaneously by leveraging the advantages of mid-level representations learning and the CVAE model in improving wording novelty. Specifically, we propose a planning-based CVAE model, targeting to generate stories with both thematic consistency and wording diversity. Our introduced method can be divided into two stages. In the *planning stage*, keyword extraction and expansion modules are used to generate keywords as sub-topics representations from the title, while in the *generation stage*, a CVAE neural network module is employed to generate stories under the guidance of previously generated keywords. In our method, the planning strategy aims to improve the thematic consistency while the CVAE module is expected to keep the wording diversity of the story. To evaluate our proposed method, we conduct experiments on a benchmark dataset, i.e., the Rocstories corpus (Mostafazadeh et al., 2016). Experimental results demonstrate that our introduced method can generate stories that are more preferable for human annotators in terms of thematic consistency and wording diversity, and meanwhile outperforms state-of-the-art methods on automatic metrics.

## 2 Related Work

### 2.1 Neural Story Generation

In recent years, neural network models have been demonstrated effective in natural language processing tasks (Mikolov et al., 2010; Sutskever et al., 2014; Rush et al., 2015; Roemmele et al., 2017; Liu et al., 2020; Yu et al., 2020). In story generation, previous studies have employed neural networks for enhancing the quality of generated content. Jain et al. (2017) explored generating coherent stories from independent short descriptions by using a sequence to sequence (S2S) architecture with a bidirectional RNN encoder and an RNN decoder. Since this model is insufficient for generating stories with consistent themes, to improve the thematic consistency of the generated stories, many other methods have been explored. Martin et al. (2018) argued that using events representations as the guidance for story generation is able to improve the thematic consistency of generated content. Fan et al. (2018) presented a hierarchical method that first generates a prompt from the title, and then a story is generated conditioned on the previously generated prompt. Following the idea of learning mid-level representations, Xu et al. (2018) proposed a skeleton-based model that first extracts skeleton from previous sentences, and then generates new sentences under the guidance of the skeleton. Similarly, Yao et al. (2019) explored using a storyline planning strategy for guiding the story generation process to ensure the output story can describe a consistent topic. Fan et al. (2019) further adopted a structure-based strategy that first generates sequences of predicates and arguments, and then outputs a story by filling placeholder entities. Although these methods have achieved promising results, most of them are implemented with RNNs, which tend to encounter common words problem. In recent researches, the Conditional Variational Auto-Encoder model is regarded as a possible solution for improving the wording diversity in story generation (Li et al., 2019).

### 2.2 Conditional Variational Autoencoder

The Variational Auto-Encoder (VAE) model is proposed in (Kingma and Welling, 2013). Through forcing the latent variables to follow a prior distribution, VAE is able to generate diverse text successfully by randomly sampling from the latent space (Bowman et al., 2016). Conditional Variational AutoEncoder (CVAE), as a variant of VAE, can generate specific outputs conditioned on a given input. CVAE has been used in many other related text generation tasks, such as machine translation (Zhang et al., 2016), dialogue generation (Serban et al., 2017; Zhao et al., 2017; Shen et al., 2017), and poem composing (Yang et al., 2018; Li et al., 2018). Subsequently, in recent years, CVAE has begun to be applied in

Proceedings of the 19th China National Conference on Computational Linguistics, pages 892–902,
Hainan, China, October 30 – November 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

story generation task to tackle the common wording problem. Li et al. (2019) explored adopting CVAE to generate stories with novel and diverse words, and Wang et al. (2019) alter the RNN encoder and decoder of CVAE architecture with the Transformer encoder and decoder (Vaswani et al., 2017) for the story completing task. Although the CVAE model has achieved encouraging performance on improving wording diversity, it is a still challenging problem to generate stories with both thematic consistency and diverse wordings. To solve this problem, in this paper, we propose a Plan-CVAE, which leverages the advantages of CVAE to generate diverse sentences and keeps the thematic consistency of the whole generated stories by using a planning strategy.

## 3 Preliminary

### 3.1 VAE and CVAE

A VAE model consists of two parts, an encoder which is responsible for mapping the input $x$ to a latent variable $z$, and a decoder which works by reconstructing the original input $x$ from the latent variable $z$. In theory, VAE forces $z$ to follow a prior distribution $p_\theta(z)$, generally a standard Gaussian distribution ($\mu = 0$, $\sigma = 1$). It first learns a posterior distribution of $z$ conditioned on the input $x$ via the encoder network, denoted as $q_\theta(z|x)$, and then applies the decoder network to computes another distribution of $x$ conditioned on $z$, denoted as $p_\theta(x|z)$, where $\theta$ are the parameters of the network.

The training objective of VAE is to maximize the log-likelihood of reconstructing the input $x$, denoted as $log\, p_\theta(x)$, which involves an intractable marginalization (Kingma and Welling, 2013). To facilitate model parameters learning, VAE can be trained alternatively by maximizing the variational lower bound of the log-likelihood, and the true posterior distribution $q_\theta(z|x)$ is substituted by its variational approximation $q_\phi(z|x)$, where $\phi$ denotes the parameters of $q$. The objective can be written as

$$L(\theta, \phi; x) = -KL(q_\phi(z|x)||p_\theta(z)) + E_{q_\phi(z|x)}[log\, p_\theta(x|z)] \tag{1}$$

The objective mentioned above contains two terms, where the first term $KL(\cdot)$ represents the KL-divergence loss, which encourages the model to keep the posterior distribution $q_\phi(z|x)$ close to the prior $p_\theta(z)$. The second term $E[\cdot]$ is the reconstruction loss for guiding the decoder to reconstruct the original input $x$ as much as possible.

CVAE is a modification version of VAE, it introduces an extra condition $c$ to supervise the generative process. Correspondingly, the encoder computes a posterior distribution $q_\theta(z|x, c)$, representing the probability of generating $z$ conditioned both on $x$ and $c$. Similarly, the distribution computed by decoder is $p_\theta(x|z, c)$, and the prior distribution of $z$ is $p_\theta(z|c)$. Accordingly, the objective of CVAE can be formulated as

$$L(\theta, \phi; x, c) = -KL(q_\phi(z|x, c)||p_\theta(z|c)) + E_{q_\phi(z|x,c)}[log\, p_\theta(x|z, c)] \tag{2}$$

### 3.2 Problem Formulation

We formulate the story generation task with the following necessary notations:

**Input:** A title $T = (t_1, t_2, ..., t_n)$ is given to the model to guide the story generation, where $t_i$ refers the $i$-th word and $n$ denotes the length of the given title.

**Output:** A story $S = \{S_1, S_2, ..., S_m\}$ should be generated by the model based on the given title, where $S_i$ represents the $i$-th sentence and $m$ denotes the total number of sentences in the story.

**Keywords:** A keywords sequence $K = (k_1, k_2, ..., k_m)$ is generated from title to enhance the process of story generation, where $k_i$ is the $i$-th keyword which serves as the sub-topic or extra hint for $S_i$.

## 4 Planning-based CVAE Method

### 4.1 Overview

The overview of our proposed method is shown in Figure 1. Our method contains two stages: a planning stage and a generation stage. In the planning stage, a *Keywords-Extraction module* followed by a *Keywords-Expansion module* are used. In this stage, several keywords are first extracted from the title,

Proceedings of the 19th China National Conference on Computational Linguistics, pages 892–902,
Hainan, China, October 30 – November 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

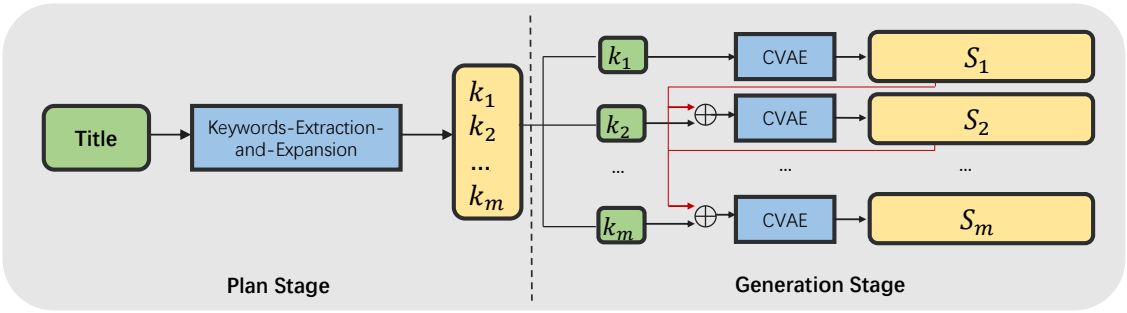



Figure 1: An overview of our proposed method.

and then the extracted keywords are expanded to match the number of sentences to be generated. In the generation stage, a *CVAE module* generates the story sentence-by-sentence conditioned on the keywords, i.e., keyword $k_i$ is used as the sub-topic or hint of sentence $S_i$.

### 4.2 Planning Stage

In the planning stage, we first utilize RAKE algorithm (Rose et al., 2010) to extract keywords from the title. Since each sentence is to be generated under the guidance of a keyword, when the number of extracted keywords is not enough, we need to expand more keywords from existing ones. We adopt a language model with a long short-term memory network (LSTM) to predict the subsequent keywords based on the previously generated keywords.

To train the model, we collect training data from the story corpus. Specifically, for each story that contains $m$ sentences in the corpus, we use RAKE to extract one keyword from one sentence. Then a keyword sequence $(k_1, k_2, ...k_m)$ corresponding to a story forms a sample in the training data. The language model is trained to maximize the log-likelihood of the subsequent keyword:

$$L(\theta) = log\, p_\theta(k_i|k_{1:i-1}) \tag{3}$$

where $\theta$ refers to the parameters of the language model, and $k_{1:i-1}$ denotes the preceding keywords.

Additionally, keywords can be directly generated by an RNN model from the title. Different from the straight-forward method, our method first extracts keywords from the title and then expands keywords to a sufficient number. Intuitively, the keywords extracted from the title possess a better consistency with the title. Thus, compared to the direct method, our method can lead to a better thematic consistency. To prove the superiority of our method, an ablation study is conducted to compare our method with the directed method, where the results are given in Table 2.

### 4.3 Generation Stage

We adopt the CVAE model for the generation stage. As demonstrated in Figure 2, the CVAE model contains an encoder and a decoder. The encoder is implemented with a bidirectional GRU network to encode both the sentences and the keywords with shared parameters. At each step, the current sentence $S_i$, preceding sentences $S_{1:i-1}$ (denoted as ctx) and the keyword $k_i$ are encoded as the concatenation of the forward and backward hidden states of the GRU, i.e. $h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$[0], $h_{ctx} = [\overrightarrow{h_{ctx}}, \overleftarrow{h_{ctx}}]$, $h_k = [\overrightarrow{h_k}, \overleftarrow{h_k}]$, respectively. $h_i$ corresponds to $x$ in Equation 2, and $[h_{ctx}, h_k]$ corresponds to $c$ in Equation 2.

Following previous work (Kingma and Welling, 2013; Zhao et al., 2017; Li et al., 2019), we hypothesize that the approximated variational posterior follows an isotropic multivariate Gaussian distribution, i.e. $q_\phi(z|x, c) \sim \mathcal{N}(\mu, \sigma I)$, where $I$ is the diagonal covariance. Thus modeling the approximated variational posterior is equal to learning $\mu$ and $\sigma$. As shown in Figure 2, a recognition network is used to learn $\mu$ and $\sigma$. Specifically, we have

$$\begin{bmatrix} \mu \\ log\, \sigma \end{bmatrix} = W_r \begin{bmatrix} x \\ c \end{bmatrix} + b_r \tag{4}$$

[0] $\rightarrow$ denotes forward and $\leftarrow$ denotes backward

Proceedings of the 19th China National Conference on Computational Linguistics, pages 892–902,
Hainan, China, October 30 – November 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

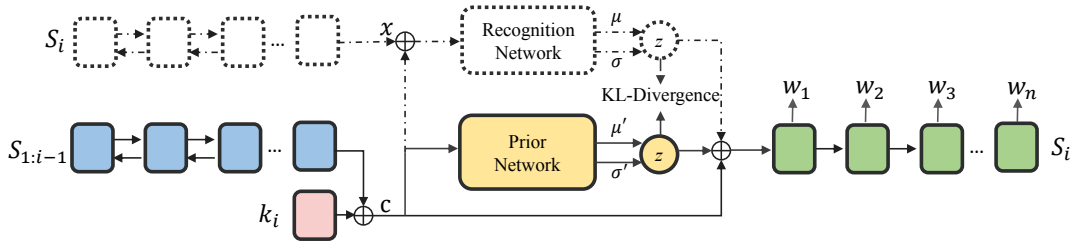



Figure 2: The architecture of the CVAE module used in the generation stage. All components are used for training, while only the components with solid lines are for testing. $\oplus$ denotes the vector concatenation operation.

where $W_r$ and $b_r$ are trainable parameters. Similarly, the prior is assumed to follow another multivariate Gaussian distribution, i.e. $p_\theta(z|c) \sim \mathcal{N}(\mu', \sigma' I)$, and $\mu'$ and $\sigma'$ are learned by the prior network in Figure 2, which is a one-layer fully-connected network (denoted as MLP) with $tanh(\cdot)$ as the activation function. Formally, it can be written as

$$\begin{bmatrix} \mu' \\ log\, \sigma' \end{bmatrix} = MLP_p(c) \tag{5}$$

The decoder is a one-layer GRU. The initial state of the decoder is computed as

$$S_{i,0} = W_d \left[z, c\right] + b_d \tag{6}$$

where $W_d$ is a matrix for dimensional transformation, $z$ is sampled from the recognition network during training and the prior network during testing. Meanwhile, a reparametrization trick (Kingma and Welling, 2013) is used to sample $z$.

Moreover, previous researches proved that CVAE intends to encounter the latent variable vanishing problem in training (Bowman et al., 2016). Thus, in our implementation, KL cost annealing (Bowman et al., 2016) and bag-of-word loss (Zhao et al., 2017) are used to tackle the problem.

## 5 Experiments

### 5.1 Dataset

We conduct experiments on the ROCStories corpus (Mostafazadeh et al., 2016), which contains 98159 stories. In our experiments, the corpus is randomly split into training, validation, and test datasets with 78527, 9816, 9816 stories. Every story in the dataset is comprised of one title and exactly five sentences, and the average word number of one story is 50.

### 5.2 Baselines

We utilize several strong and highly related methods of story generation as our baselines.

**S2S**, the sequence to sequence model (Sutskever et al., 2014) which has been widely used in multiple text generation tasks, such as machine translation and summarization. We implement it to generate stories in a sentence-by-sentence fashion, and the $i$-th sentence is generated by taking all the previous $i-1$ sentences as input.

**AS2S**, the sequence to sequence model enhanced by an attention mechanism (Bahdanau et al., 2015), which is an improved version of S2S. It takes the same generation pipeline as S2S.

**CVAE**, the CVAE model without planning strategy. This pure CVAE model takes only the previous $i-1$ sentences as the condition $c$ to generate the $i$-th sentence. This baseline is for demonstrating the performance of CVAE without planning strategy.

**Plan-and-Write**, the AS2S model with planning strategy proposed in (Yao et al., 2019). Two different schema (static and dynamic) for keywords generation are proposed in the original paper. As the authors have proved that the static one is better, we implement the static scheme as our baseline.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 892–902,
Hainan, China, October 30 – November 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Table 1: Descriptions of human evaluation metrics.

| | |
|---|---|
| **Readability** | Is the story formed with correct grammar? |
| **Consistency** | Does the story describe a consistent theme? |
| **Creativity** | Is the story narrated with diversified wordings? |

Table 2: Results of BLUE and Distinct scores. B-$n$ and D-$n$ represent the BLUE scores and Distinct scores on $n$-grams respectively. The final results are scaled to [0, 100]. The difference between Plan-CVAE* and Plan-CVAE is the former generates keywords directly from the title, while the latter generates keywords using our keywords-extraction and keyword-expansion module.

| **Model** | **Automatic Evaluation** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **B-1** | **B-2** | **B-3** | **B-4** | **D-1** | **D-2** | **D-3** | **D-4** |
| S2S | 23.65 | 9.30 | 4.07 | 1.97 | 0.90 | 4.11 | 10.70 | 19.37 |
| AS2S | 24.70 | 9.68 | 4.27 | 2.07 | 0.93 | 4.53 | 11.13 | 19.41 |
| CVAE | 28.53 | 10.21 | 3.63 | 1.39 | 1.67 | 15.82 | 46.88 | **76.64** |
| Plan-and-Write | 27.39 | 11.78 | **5.57** | **2.85** | 0.84 | 5.15 | 14.67 | 28.28 |
| Plan-CVAE* | 29.57 | 11.32 | 4.43 | 1.85 | 1.52 | 14.13 | 42.30 | 71.42 |
| Plan-CVAE | **30.25** | **12.05** | 4.89 | 2.03 | **1.75** | **16.38** | **46.98** | 75.73 |
| Human | - | - | - | - | 2.87 | 26.74 | 62.92 | 86.67 |

### 5.3 Model Settings

We train our model with the following parameters and hyper-parameters. The word embedding size is set to 300, and the vocabulary is limited to the most frequent 30000 words. The hidden state size of encoder, decoder, and prior network are 500, 500, 600 respectively. And the size of the latent variable $z$ is set to 300. To train our model, we adopt the Adam (Kingma and Ba, 2015) optimization algorithm with an initial learning rate of 0.001 and gradient clipping of 5. All initial weights are sampled from a uniform distribution $[-0.08, 0.08]$. The batch size is 80.

### 5.4 Evaluation

We utilize both automatic and human metrics to evaluate the performance of our method.

**BLUE Score.** This metric is designed for calculating the word-overlap score between the golden texts and the generated ones (Papineni et al., 2002), and has been used in many previous story generation works (Yao et al., 2019; Li et al., 2019).

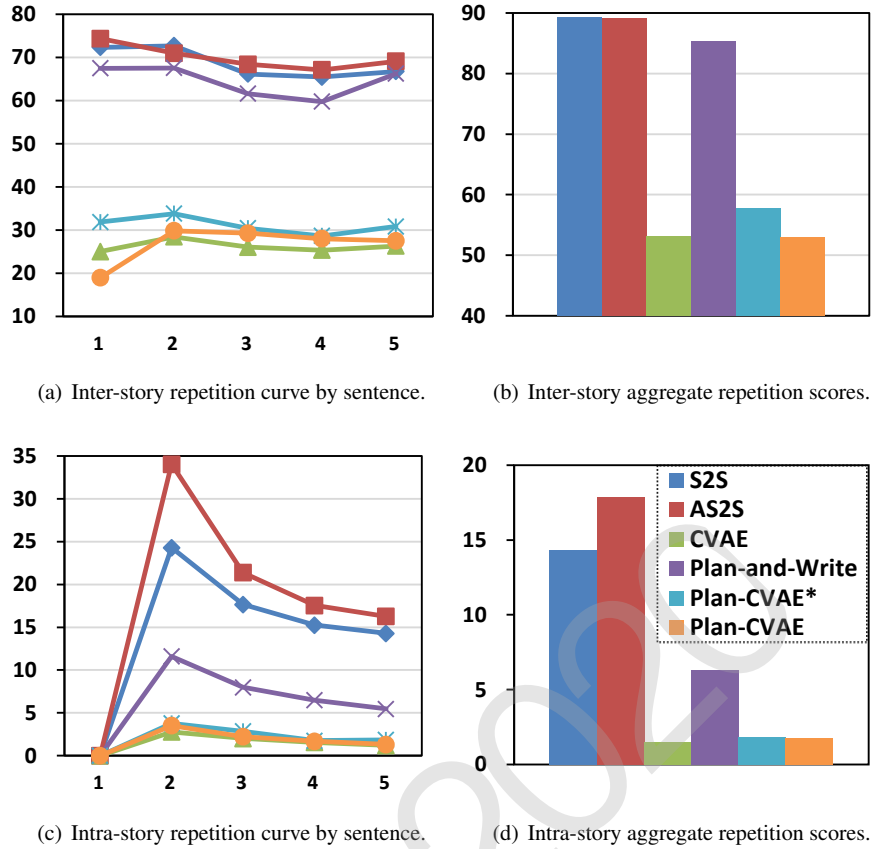
**Distinct Score.** To measure the diversity of the generated stories, we employ this metric to compute the proportion of distinct $n$-grams in the generated outputs (Li et al., 2016). Note that the final distinct scores are scaled to [0, 100].

**Inter- and intra-story repetition.** These two metrics are proposed in (Yao et al., 2019) and used for calculating the inter- and intra-story tri-grams [1] repetition rates by sentences and for the whole stories. The final results are also scaled to [0, 100].

**Human Evaluation.** We also employ three metrics for human evaluation, i.e., Readability, Consistency, and Creativity. Their descriptions are shown in Table 1. We randomly sample 100 generated stories from each baseline model and our method and then perform pairwise comparisons between our method and baselines. That is, for two stories with the same titles but generated by different two models, five well-educated human evaluators are asked to select the one they prefer on the three metrics. In comparison, no *equally good* option is given since the *equally good* option may leads to a careless comparison.

[1] Results on four and five-grams have the same trends.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 892–902,
Hainan, China, October 30 – November 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China




(a) Inter-story repetition curve by sentence.

(b) Inter-story aggregate repetition scores.

(c) Intra-story repetition curve by sentence.

(d) Intra-story aggregate repetition scores.

Figure 3: Inter- and intra-story repetition rates by sentences (curves) and for the whole stories (bars). Final results are scaled to [0, 100], the lower the better.

## 6 Results and Analysis

Table 2 and Figure 3 present the results of automatic evaluation, and Table 3 shows the results of human evaluation. Through analyzing these evaluation results, we have the following observations.

### 6.1 The Effect of the Planning Strategy

***The planning strategy is effective for improving thematic consistency.*** As shown in Table 2, BLEU-[1-4] scores of Plan-CVAE are significantly higher than the pure CVAE model. Higher BLEU scores indicate that the planning strategy can improve the word-overlapping between the generated stories and the gold standard ones, which means the generated stories are more relevant with thematically consistent cases. For the subjective feelings of humans, as indicated by the human consistency evaluation in Table 3), Plan-CVAE can generate stories with better thematic consistency than the CVAE model. Meanwhile, Plan-CVAE outperforms all baselines on thematic consistency in human evaluation, this means the CVAE model gains a significant improvement on thematic consistency by using the planning strategy.

***The planning strategy does not affect the wording diversity.*** The planning strategy aims to enhance the CVAE model with better thematic consistency while preventing poor wording diversity. Plan-CVAE has a comparable performance with CVAE and outperforms other baselines on distinct scores in Table 2 and the creativity metric in Table 3, which prove that the planning strategy does not affect the wording novelty. In Figure 3, we also have a similar observation that both Plan-CVAE* and Plan-CVAE models achieve a quite low inter- and intra-story repetition rates, which means our proposed model can learn to create stories rather than copy and concatenate frequently occurred phrases in the training corpus.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 892–902,
Hainan, China, October 30 – November 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Table 3: Results of human evaluation.

| | | | |
|---|---|---|---|
| **Readability** | | | |
| Plan-CVAE | 44% | **56%** | S2S |
| Plan-CVAE | **58%** | 42% | AS2S |
| Plan-CVAE | **67%** | 33% | CVAE |
| Plan-CVAE | 47% | **53%** | Plan-and-Write |
| **Consistency** | | | |
| Plan-CVAE | **65%** | 35% | S2S |
| Plan-CVAE | **61%** | 39% | AS2S |
| Plan-CVAE | **84%** | 16% | CVAE |
| Plan-CVAE | **58%** | 42% | Plan-and-Write |
| **Creativity** | | | |
| Plan-CVAE | **93%** | 7% | S2S |
| Plan-CVAE | **86%** | 14% | AS2S |
| Plan-CVAE | **57%** | 43% | CVAE |
| Plan-CVAE | **81%** | 19% | Plan-and-Write |

### 6.2 The Effect of CVAE

***The CVAE model can effectively improve the wording diversity.*** Plan-CVAE outperforms all baselines (excepts for CVAE) on automatic evaluations including distinct scores in Table 2 and inter- and intra-story repetition rates in Figure 3, and on creativity score of human evaluation in Table 3. Specifically, all baselines based on RNNs, i.e., S2S, AS2S, and Plan-and-Write, achieve a quite low distinct score and high inter- and intra-story repetition rates, while Plan-CVAE significantly outperforms them by nearly doubling the distinct scores, reducing the repetition rates to about half of theirs, and achieving similar scores with the pure CVAE model. Results on the creativity metric in human evaluation also indicate the same conclusion. These results support the intuition that CVAE can address the poor wording diversity problem of RNN by randomly sampling from the latent space.

***The latent variable in CVAE reduces the readability.*** CVAE improves the wording diversity by randomly sampling from the latent space. Thus, CVAE produces more uncertainty than RNNs and leads to inferior readability. This intuition is supported by the readability metric in human evaluation (Table 3).

### 6.3 Case Study

We present two groups of example stories in Table 4 to compare the performance of our proposed method with Plan-and-Write and CVAE since they perform well on either thematic consistency or wording novelty. When compared with the pure CVAE model without planning strategy, we can observe that stories generated by CVAE are formed with novel words but without consistent topics, while stories generated by Plan-CVAE describe more consistent themes and are also narrated with novel wordings. On the other hand, when compared with the planning-based RNN method, we can see that stories generated by Plan-and-Write contain many repeated words and phrases, while stories composed by Plan-CVAE consist of more diverse words.

Although our proposed method has achieved a promising improvement in thematic consistency and wording diversity than baseline models, there is still a gap between stories generated by our method and humans. We also observed some bad cases generated by our method. These bad cases reflect three major problems, i.e., logical inconsistency, lacking sentiment, and weak readability. As for the logical inconsistency problem, one can introduce an extra control to dynamically adjust the keywords planning and content generation process, or establish a polishing mechanism to check and rewrite the generated content. To solve the sentiment problem, one can utilize a sentiment planning strategy to add sentimental information into the generated stories. In order to improve readability, it is worth to use pre-training strategy on larger related corpus for story generation. Our future work will focus on these issues.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 892–902,
Hainan, China, October 30 – November 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Table 4: Two group of example stories automatically generated by our proposed Plan-CVAE model and baseline methods.

| | | |
|---|---|---|
| | **Title: Lake fun** | |
| **Plan-and-Write** | Keywords | friends → lake → water → day → home |
| | Story | I went to the lake with my friends. We went to a lake. We went to the lake. We went to a water park. We had a great time. |
| **CVAE** | Story | The kids rode outside of the neighborhood. The mom helped her to get away from the house. They played with the best and hung them out. Finally, they were a couple. Then, they played games. |
| **Plan-CVAE** | Keywords | lake → friends → water → swim → shore |
| | Story | One summer day, my friends invited me to a different lake. While walking around the lake, we moved along. It was a great summer and my dad built a lot of stuff. Now the rest of the day I can swim. I was very late and went back to shore. |
| | **Title: Tax evasion** | |
| **Plan-and-Write** | Keywords | money → taxes → pay → debt → bank |
| | Story | Bill had a lot of money. Bill had to pay for his taxes. Bill had to pay for his taxes. Bill was able to pay his taxes. Bill was able to pay his taxes . |
| **CVAE** | Story | Sam and her husband wanted to be a lawyer. He worked hard to make it with his lawyer. He just needed to buy more and more things. He made a lot of money. He told her they both hurt more. |
| **Plan-CVAE** | Keywords | tax evasion → money → taxes → pay → debt |
| | Story | Neil had recently moved to a tax preparer. He had applied for a few jobs before the loan. But he didn't notice that the bank was on his list! He was told he owed pay for tax fees. It was about to accept his taxes, but he had no interest. |

## 7 Conclusion

In this paper, we proposed a planning-based conditional variational autoencoder model (Plan-CVAE) for story generation. Our proposed method involves two stages. In the planning stage, the keyword-extraction and keyword-expansion modules are used to generate keywords from the title. As for the generation stage, a CVAE neural network module is employed to generate stories under the guidance of keywords. In our method, the planning strategy aims to improve the thematic consistency while the CVAE module is expected to keep the wording diversity of the generated story. Experimental results of both automatic and human evaluations on a benchmark dataset, i.e., ROCStories corpus, show that our method performs better than existing methods on thematic consistency and wording diversity. The case study also confirms the effectiveness of our method.

## Acknowledgements

We would like to thank the 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. 61876196 and NSFC No. 61672058) and the foundation of Key Laboratory of Artificial Intelligence, Ministry of Education, P.R. China. Rui Yan was sponsored by

Beijing Academy of Artificial Intelligence (BAAI).

## References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

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, August. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia, July. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy, July. Association for Computational Linguistics.

Parag Jain, Priyanka Agrawal, Abhijit Mishra, Mohak Sukhwani, Anirban Laha, and Karthik Sankaranarayanan. 2017. Story generation from sequence of independent short descriptions.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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, California, June. Association for Computational Linguistics.

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, Brussels, Belgium, October-November. Association for Computational Linguistics.

Juntao Li, Lidong Bing, Lisong Qiu, min D Chen, Dongyan Zhao, and Rui Yan. 2019. Learning to write creative stories with thematic consistency. In *AAAI 2019 : Thirty-Third AAAI Conference on Artificial Intelligence*.

Danyang Liu, Juntao Li, Meng-Hsuan Yu, Ziming Huang, Gongshen Liu, Dongyan Zhao, and Rui Yan. 2020. A character-centric neural model for automated story generation. In *AAAI*, pages 1725–1732.

Lara J. Martin, Prithviraj Ammanabrolu, William Hancock, Shruti Singh, Brent Harrison, and Mark O. Riedl. 2018. Event representations for automated story generation with deep neural nets. *ArXiv*, abs/1706.01331.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Melissa Roemmele, Sosuke Kobayashi, Naoya Inoue, and Andrew Gordon. 2017. An RNN-based binary classifier for the story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 74–80, Valencia, Spain, April. Association for Computational Linguistics.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley, 2010. *Automatic Keyword Extraction from Individual Documents*, pages 1 – 20. 03.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3295–3301. AAAI Press.

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: Short Papers)*, pages 504–509, Vancouver, Canada, July. Association for Computational Linguistics.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Tianming Wang and Xiaojun Wan. 2019. T-cvae: Transformer-based conditioned variational autoencoder for story completion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5233–5239. International Joint Conferences on Artificial Intelligence Organization, 7.

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4306–4315, Brussels, Belgium, October-November. Association for Computational Linguistics.

Xiaopeng Yang, Xiaowen Lin, Shunda Suo, and Ming Li. 2018. Generating thematic chinese poetry using conditional variational autoencoders with hybrid decoders. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4539–4545. AAAI Press.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Meng-Hsuan Yu, Juntao Li, Danyang Liu, Bo Tang, Haisong Zhang, Dongyan Zhao, and Rui Yan. 2020. Draft and edit: Automatic storytelling through multi-pass hierarchical conditional variational autoencoder. In *AAAI*, pages 1741–1748.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Austin, Texas, November. Association for Computational Linguistics.

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)*, pages 654–664, Vancouver, Canada, July. Association for Computational Linguistics.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 892–902,
Hainan, China, October 30 – November 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China