Generating Long Financial Report Using Conditional Variational Autoencoders With Knowledge Distillation

Ziao Wang, Yunpeng Ren, Xiaofeng Zhang, and Yiyuan Wang

Abstract—Generating financial reports from a piece of news is a challenging task due to the lack of sufficient background knowledge to effectively generate long financial reports. To address this issue, this article proposes a conditional variational autoencoders (CVAEs)-based approach that distills external knowledge from a set of news–report data. Specifically, we design an encoder-decoder architecture to learn the latent variable distribution from this set of news–report data to provide background knowledge. Next, a teacher–student network is employed to distill knowledge to refine the output of the decoder component. To evaluate the model performance, extensive experiments have been performed on two public datasets using evaluation criteria like BLEU, ROUGE, METEOR, and human evaluation. Our promising results demonstrate that our proposed approach outperforms existing state-of-the-art approaches.

Impact Statement—The financial report generation task involves generating long financial reports from concise news pieces, which is challenging due to the lack of sufficient background information. To address this issue, we propose a novel approach based on CVAEs with knowledge distillation. The teacher–student network proposed is integrated with the CVAE model to address both information loss and knowledge reasoning issues simultaneously. The experimental results demonstrate that our proposed approach is superior.

Index Terms—Conditional variational autoencoder (CVAE), financial report generation, knowledge distillation (KD), natural language generation (NLG).

I. INTRODUCTION

Text generation has been extensively studied in natural language processing (NLP). Typically, text generation tasks encompass a range of subproblems, including open-ended text generation [1], [2], [3] and conditional text generation [4], [5], [6]. Conditional long-text generation from short text is particularly challenging, especially in domain-specific applications, such as financial report generation. The primary challenge in generating financial reports based on concise news articles is the lack of sufficient information, particularly financial background knowledge. Unfortunately, this challenging task remains largely unaddressed by existing research.

In recent years, a good number of deep neural network-based text generation approaches have been proposed [7], [8], [9], [10]. Primarily, these approaches rely on deep generative models, such as variational autoencoder (VAE) and generative adversarial network (GAN). VAE-based models [11], [12], [13] learn latent variables at the sentence level, allowing them to learn the comprehensive properties of text, including style, topic, and high-level syntactic features. Additionally, a conditional variational autoencoder (CVAE) [14] with a hybrid decoder was introduced to learn topics and generate poems. Shen et al. [1] proposed a multilevel VAE-based approach, leveraging both high-level document features and low-level word features for text generation. GAN-based approaches have also gained significant attention [2], [15], [16], and much research attention has been paid to generate high-quality long text. Typically, the generator in GAN-based approaches is trained to maximize the probability of generating adversarial examples to deceive the discriminator. Yang et al. [15] extract knowledge from an external base to integrate it with the generator via a dynamic memory mechanism, and the model is adversarially trained with a multiclassifier serving as the discriminator to incorporate background knowledge.

However, the task of generating financial reports presents unique research challenges that invalidate many existing natural language generation (NLG) approaches. First, the input news is much shorter compared to the output report, as illustrated in Table I. While recent pretrained models can generate long text, the generated content may not be sufficiently informative regarding the input news. Second, human writers use their professional knowledge and reasoning to produce coherent and contextual text in financial reports. Unfortunately, the existing literature has not addressed these challenges.

To address the aforementioned issues, we propose a novel CVAE-based approach with knowledge distillation (CVAE-KD). Fig. 1 provides a graphical overview of the proposed approach. The CVAE model incorporates a carefully designed teacher–student network structure [17] to simultaneously...
We encode the input news using a bidirectional gated recurrent unit (Bi-GRU) [18], and the latent variable distribution is learnt during model learning, \( z_1 \) approximates \( z_2 \) using a designed KL-divergence term. We use a GRU component [18] in the decoder to generate the target reports from \( z_1 \). Additionally, a teacher component is adopted, which embeds a set of corresponding financial reports through a pretrained ELMO model [19]. The output logits of the decoder component serve as the student for further refinement. The main contributions of this article are summarized as follows.

1) The article presents a novel approach utilizing a CVAE with KD to generate financial reports from concise news. To the best of our knowledge, this work is among the first attempts to tackle this challenging task.
2) We employ a pretrained model as a teacher network and the student component is to learn background knowledge from the external knowledge base, and the corresponding KL-divergence loss is designed to train the model.
3) We conducted extensive experiments on two public real-world datasets and achieved superior results compared to the state-of-the-art approaches across both automatic and human evaluation metrics.

The article is organized as follows: Section II highlights the related text-to-text generation approaches; Section III presents the CVAE-KD approach in detail; Section IV reports the experimental results; and Section V discusses the proposed approach. The article concludes in Section VI.

### II. RELATED WORK

There are three main categories of text-to-text generation approaches: sequence-to-sequence (Seq2seq)-based methods, VAE-based methods, and GAN-based methods, along with diffusion-based approaches. We provide a brief review of each of these approaches in the following sections.

#### A. Seq2seq-Based Approaches

Seq2seq-based methods have been shown to have superior performance in various text-to-text generation tasks. Cho et al. [18] proposed a RNN-based Seq2seq network that uses an encoder to embed the input sequence into a fixed-length feature vector and a decoder that maps the vector to the target sequence.
To further enhance the performance, an attention mechanism [20] was proposed, which enables the decoder to automatically search relevant parts of the input sequence to predict the target word. In addition, Gu et al. [21] proposed CopyNet, which allows subsequences from the input sequence to be directly copied during decoding.

Moreover, Hu et al. [22] proposed a two-stage hybrid deep learning model to simulate the process of human writing and generate macro research reports from breaking news. To incorporate commonsense reasoning into text generation, Liu et al. [23] proposed KG-BART, a novel pretrained language generation model augmented with a knowledge graph to generate more natural and logical sentences that take into account the complex relations between concepts. Additionally, Liu et al. [24] introduced a knowledge-infused decoding algorithm that dynamically infuses external knowledge into each decoding process.

### B. Variational Autoencoder-Based Approaches

The VAE models have been widely adopted in the text generation task [25]. In particular, the RNN-based VAE model proposed by Bowman et al. [25] can generate coherent novel sentences by learning the feature representations of latent variables at the sentence level. This allows the proposed VAE model to explicitly represent the holistic properties of sentences such as style, topic, and high-level syntactic features. Recently, Hosking et al. [26] proposed a VAE-based model for paraphrase generation where the hierarchical discretized embeddings are learnt over the latent space, while Dai et al. [27] proposed an adversarial poincare variational autoencoder (APo-VAE) for text generation in a hyperbolic latent space to learn continuous hierarchical representations.

To generate long text, various VAE-based models have been proposed in the literature. For instance, Wang and Wan [13] introduced a graph transformer encoder–decoder model to generate text that well captures the higher level semantic meanings of the target text. The graph transformer encoder can leverage the relational structure of the employed knowledge graph without imposing linearization or hierarchical constraints. In addition, Yang et al. [28] proposed a topic-aware long-text generation model where discrete latent codes are designed to model the hidden topics.

### C. Generative Adversarial Network-Based Approaches

The original GAN proposed by Goodfellow et al. [29] has various applications, and inspired by it, Kusner and Hernández-Lobato [30] developed a GAN-based model that employs a Gumbel-softmax distribution to enable text generation with discrete variables. Recently, Yang and Klein [31] introduced the future discriminators for generation (FUDGE) model to generate high-quality poetry by giving a distribution of interest to a preexisting model for text generation. Dhar et al. [32] proposed a GAN-based voice conversion model that achieves high speaker similarity and good speech quality by utilizing a dense residual network architecture, adaptive learning mechanism, and boosted learning rate approach.

To improve the coherence and quality of generated text, researchers have made significant improvements to the GAN model. For example, Liu et al. [33] proposed a category-aware GAN, consisting of an efficient category-aware model for categorical text generation that reduces the gap between real and generated samples. In addition, Wu et al. [34] introduced a generative adversarial imitation learning framework for text generation, utilizing a large pretrained language model (LM) as a backbone model to provide reliable guiding signals to GAN discriminators. Moreover, Wang et al. [35] presented a controlled adversarial text generation (CAT-Gen) model that generates adversarial texts through controllable attributes, which are known to be invariant task labels to reduce perturbations in input text.

### D. Diffusion-Based Approaches

Diffusion-based generation methods are a new paradigm for generative models that have recently emerged and achieved record-breaking performance in many applications [36]. These methods transform data into noise using diffusion processes that can be reversed through learning the score function, i.e., the gradient of the log-density of perturbed data [37]. Diffusion models have also been successfully applied to NLP tasks such as text generation [38], [39]. For instance, Li et al. [39] proposed a method for controlling the behavior of pretrained language models (LMs) without retraining, utilizing diffusion models to achieve strong controllable text generation. Similarly, Gong et al. [38] presented a diffusion-based Seq2seq model that generates high-quality and diverse text using end-to-end classifier-free guided diffusion generation. In addition, Liu et al. [40] proposed an efficient approach for composable text operations in the compact latent space of text using the diffusion process to control text properties in a low-dimensional continuous space. Once the latent vectors have the corresponding properties, they are passed to the decoder to generate the text. Furthermore, He et al. [41] introduced DiffusionBERT, a new generative masked LM based on discrete diffusion models.

### III. THE PROPOSED APPROACH

#### A. Preliminaries and Problem Setup

Let $X$ be a set of news data, where each $x = (x_1, \ldots, x_M)$ contains $M$ tokens, and $Y$ be a set of generated reports, where each $y = (y_1, \ldots, y_N)$ contains $N$ tokens, where $N$ is significantly greater than $M$. We formulate the financial report generation problem as follows:

$$p(y_1, y_2, \ldots, y_N) = \prod_{k=1}^{N} p(y_k | y_{1:k-1}; x)$$  

where $p(y_1, y_2, \ldots, y_N)$ is the joint probability of all generated tokens, and the probability of each token $y_k$ is conditioned on all previous tokens $y_{1:k-1}$ and the input article $x$. Therefore, the model generates $y$ by autoregressively predicting each token in sequence given the input $x$. 

---

This response has been formatted to follow the guidelines provided, ensuring it is clear, self-contained, and adheres to the natural reading format. The content has been extracted and represented in a structured manner, preserving the logic and coherence of the original text while focusing on the key points and methodologies discussed.
B. The Proposed CVAE-KD

As shown in Fig. 2, the proposed CVAE-KD consists of three components, i.e., an encoder component, a background knowledge extraction component, and a decoder component. To encode the input news, the encoder component first employs a Bi-GRU module [18] to obtain a latent variable distribution, $z_1$. This is accomplished by estimating the mean $\mu_x$ and variance $\sigma_x$ variables through MLP layers [42]. In the background knowledge extraction component, a set of similar news $X$, along with their corresponding reports $Y$, is extracted to estimate a higher level latent variable distribution $z_2$ that contains relevant background information about the input text. Finally, the decoder component uses a GRU module to generate the output report from $z_1$, and the KD process is applied to learn the background information. We detail each component in the following sections.

C. Encoder Component

We adopt a Bi-GRU module as the encoder component. The input of this component is a piece of news $x$. We first embed $x$ using a lookup table before feeding it to the Bi-GRU component. The final output of this component is the hidden state $h_x$ of the Bi-GRU. Specifically, the update gate $c_t$, reset gate $r_t$, and candidate activation $h_t^c$ are computed as follows:

$$c_t = \sigma(W_c x + U_c h_{t-1}^c)$$

$$r_t = \sigma(W_r x + U_r h_{t-1}^c)$$

$$h_t^c = \tanh(W_x x + U h_{t-1}^c)$$

$$h_t^{xe} = (1 - c_t) h_{t-1}^{xe} + c_t h_t^c$$

$$h_t^c = [h_t^c, h_t^{xe}]$$

where $h_{t-1}^{xe}$ is the previous activation, $h_t^{xe}$ is the current activation, and $h^c$ is the concatenation of bidirectional hidden states.

Once news data have been embedded into feature vectors, we assume that the latent variable $z$ of the CVAE follows a Gaussian distribution according to [11]. We use a Gaussian distribution as it is a simple and stable probabilistic distribution that can be employed in VAEs to resample data, as highlighted in [11], [43].

In our proposed model, illustrated in Fig. 2, we employ two MLPs to estimate the distribution of the latent variable, i.e., the mean $\mu_x$ and standard deviation $\sigma_x$ as follows:

$$\mu_x = f_{\mu_x}(h_t^c)$$

$$\sigma_x = f_{\sigma_x}(h_t^c).$$

Once the latent variable distribution has been learned, we can then sample the latent variable $z_1$ from the distribution as follows:

$$z_1^i = \mu_x^i + \sigma_x^i \epsilon, \epsilon \sim N(0, I).$$

D. Background Knowledge Extraction and Distillation Component

1) Learning Latent Variable From Background Knowledge:

In this component, we begin by building an external knowledge base. We use a standard KNN algorithm to group the most similar news data $x$ to form a subset $X_s = (x_1, ..., x_Q)$ containing $Q$ pieces of news. We then extract the corresponding financial reports to form a set $Y_s = (y_1, ..., y_Q)$ containing $Q$ pieces of reports.

As in the previous steps, the embeddings of $X_s$ and $Y_s$ can be acquired using a Bi-GRU component. A Gaussian distribution
is then assumed for this background knowledge base, with the following parameter estimates:

\[
\begin{align*}
\mu_{X,Y} &= f_{\mu_x}([X_{em}; Y_{em}]) \\
\sigma_{X,Y} &= f_{\sigma_x}([X_{em}; Y_{em}]) \\
z_i &= \mu_{X,Y} + \sigma_{X,Y} \epsilon_{X,Y}, \epsilon_{X,Y} \sim N(0, I)
\end{align*}
\]

where \(X_{em}\) and \(Y_{em}\) are feature representations of each input data pair (news–report), and \(z_2\) is the latent variable sampled from the learnt Gaussian distribution.

2) KD Process: To distill knowledge, the extracted subsets \(X_s\) and \(Y_s\) are first embedded using a pretrained model ELMo, denoted as ELMo_{task}. To further enhance their discriminating power, we concatenate these ELMO embeddings with the original feature representations, yielding the following expressions for the embedded variables:

\[
\begin{align*}
X_{em} &= [X_k; \text{ELMO}_{k}^{\text{task}}] \\
Y_{em} &= [Y_k; \text{ELMO}_{k}^{\text{task}}].
\end{align*}
\]

Next, to generate the logits of the component output, we apply an MLP layer to these embeddings and then a softmax function, written as

\[
Y_{\text{logits}} = \text{softmax} \left( \frac{\tanh(W_c[Y_{em}])}{T} \right)
\]

where \(W_c\) is a weight matrix, and \(T\) is the temperature used to control the probability distribution of the logits. Typically, \(T\) is set to 1 [17].

The resulting \(Y_{\text{logits}}\) is considered as the teacher to supervise the generation of financial report \(y\) given \(x\).

Accordingly, the student is the output logits of the decoder component. Given the input \(x\), the employed decoder is to predict the probability of the next generated token given a sequence of generated tokens in \(y\), written as

\[
p(y) = \prod_{k=1}^{N} p(y_k | y_{1:k-1}, x).
\]

The difference between the student’s logits and the teacher’s logits can be measured by the following equation:

\[
L_{\text{kd}}(\theta) = - \sum_{y \in V} [P_\phi(y_t = \omega | x, y) \log P_\theta(y_t = \omega | x, y)]
\]

where \(P_\phi(y_t)\) represents the logits predicted by the teacher model, \(P_\theta(y_t)\) represents the logits predicted by the decoder, known as the student, and \(V\) represents the output vocabulary.

Since our target of the KD module is for students to learn the behavior of the teacher, we achieve this goal by minimizing the distance \(L_{\text{kd}}(\theta)\) in (17).

E. Decoder Component

A GRU model is used as the decoder component to generate a report \(y\). The output of the employed GRU is considered as the output of this component, calculated as

\[
h^*_{t} = \text{GRU}(z_1).
\]

The output \(h^*_{t}\) is then passed to a MLP module and the vocabulary distribution is learned using the softmax function represented as

\[
p(y_t) = \text{softmax}(\tanh(W_m h^*_{t} + b)).
\]

The greedy decoding strategy is adopted, where the word with the maximum probability value is selected as the final output at each iteration.

F. Model Loss

The CVAE-KD loss function proposed in this study comprises of two terms—the CVAE loss and the KD loss. Each loss term is explained in detail in the following sections. The CVAE loss is composed of two parts—the reconstruction loss and the constraint on the latent variable \(z_1\) (derived from input news \(x\)) to approximate \(z_2\) (derived from a background knowledge base). This can be represented using the following equation:

\[
L_{\text{CVAE}}(x, y; \theta, \phi) = -D_{KL}[q_\phi(z_2 | X, Y) || p_\theta(z_1 | x)] + E_{q_\phi(z_1 | x, z_2)}[\log p_\theta(y | z_1, x)]
\]

where \(p_\theta(z_1 | x)\), and \(q_\phi(z_2 | X, Y)\) are, respectively, calculated as

\[
\begin{align*}
q_\phi(z_1 | x, z_2) &= N(z_1; \mu_x, \sigma_x, \sigma_z) \\
q_\phi(z_2 | X, Y) &= N(z_2; \mu_x, \sigma_x, \sigma_y).
\end{align*}
\]

The CVAE loss mentioned in (20) encourages the model to generate a target report given the input news, while the KD loss in (17) encourages the model to learn background knowledge as a complement to the input news. To generate high-quality reports with sufficient information, we aim to jointly optimize these two objective functions. Therefore, the overall model loss for the proposed CVAE-KD is given as

\[
L_{\text{total}} = \alpha L_{\text{CVAE}} + (1 - \alpha) L_{\text{kd}}
\]

where \(\alpha\) is a learnable parameter. The model is then optimized using the Adam algorithm [44].

IV. EXPERIMENTS

This section presents the preparation of the dataset along with the evaluation criteria used. In addition, we introduce several baseline models as well as state-of-the-art approaches. Finally, we conduct extensive experiments using a widely adopted long-text generation dataset and a real-world dataset.

A. Dataset and Data Preprocessing

1) Dataset Preparation: To evaluate the performance of our model in generating long financial reports, we utilize a real-world dataset consisting of financial news and reports. To further evaluate the effectiveness and generalization of our proposed model, we also use the arXiv dataset for long-text generation [45]. The original arXiv dataset comprises over 170 million articles, but in this study, we adopt the arXiv dataset
from [45] to generate abstracts based on the article title. The News–Report dataset used in this study consists of financial news and corresponding reports crawled from three popular Chinese financial websites: Sina Finance,1 Tonghuashun Finance,2 and Eastmoney.3 The raw News–Report dataset contains 10,706 pairs of news–report data, where each piece of news is associated with a financial report.

2) Data Preprocessing: To eliminate numeric symbols and special characters from the original news data, we used the part-of-speech tagging module in the NLP tool called HanLP [46]. In Table I, we present a sample news–report pair before and after processing to highlight the structure of the dataset.

To segment the Chinese news data, we first used an open source tool called “jieba.”4 Following word segmentation, the average length of financial news and reports was found to be 28 and 331 words, respectively. As the text generation component requires the length of the input data to be the same, we truncated or lengthened the news and financial reports to the same length. The input data length is set to 30 words, and the financial report length to 200 words. We then filtered out words whose term frequency (TF) is lower than 5, which resulted in a new vocabulary set containing 17,210 words. Numeric symbols were replaced with a number token (NUM), and four other tokens were added to the vocabulary as well, namely the padding token (PAD), unknown token (UNK), start position token (START), and end position token (END). The PAD is used to fill the encoder and decoder inputs, while UNK represents words not in the vocabulary, such as entity names. START and END are added at the beginning and end of each report, respectively.

B. Experimental Settings

1) Network Parameters: The parameters for our proposed neural network model are designed as follows. The number of hidden units in the GRU component is set to 16, and the dimension of the hidden variables is set to 256. To prevent overfitting, we set the dropout rate for the decoder to 0.5. The learning rate is set to 0.001, the batch size is set to 16, and the weight for the KD loss is set to 1.

2) Experimental Settings: In the experiments, the maximum length of encoder is set to 30. To generate reports of different length, the length of decoder is, respectively, set to 100, 150, and 200 to evaluate the model performance of the proposed CVAE-KD. Then, we evaluate all models and report the experimental results in following sections.

C. Baseline Models

We evaluated the performance of our proposed model, CVAE-KD, against several baselines and state-of-the-art approaches, namely Seq2Seq [47], Seq2Seq+Attn [20], pointer-generator network [48], writing-editing network [49], plan-and-write [50], CVAE [51], and multiedit network [22]. We chose these models for comparison due to their notable performance in various text-to-text generation tasks. We describe each model in detail as follows.

1) Seq2Seq [47] is considered as a baseline model for text generation task which already achieves a superior model performance in various text-to-text generation problem.

2) Seq2Seq+Attn [20] extends the original Seq2Seq model and is originally proposed for neural machine translation task. It employs a fixed-length vector for both encoder and decoder components, facilitating an automatic soft search for relevant words to predict the next generated word, without requiring a hard segmentation task.

3) Pointer-generator network [48] is considered as the state-of-the-art model proposed for sampling words from the input source sentences via the pointing process. The model integrates the coverage mechanism to penalize the generation of repetitive words, allowing it to copy words from the source while generating new words through the generator, which is similar to our proposed model.

4) Writing-editing network [49] is a revision network of Seq2Seq model, which convert abstract generation to an iterative task. It attends to both the title and previously generated abstract drafts, and iteratively revises and polishes the abstract.

5) Plan-and-write [50] is a revised Seq2Seq model which is used to generate diverse stories. The model proposes a plan-and-write hierarchical generation framework where the model first plans a storyline before generating a story based on the storyline.

6) CVAE [51] is most related to our proposed approach, it employs latent variables to learn the distribution over potential conversational intentions and attends to the characteristics of response at the word-level, as well as the topic information at the discourse-level. CVAE-based models are among the best models for long-text generation.

7) Multiedit network [22] draws lessons from the process of human writing which first learns the outline of the input news and then generates macro financial reports from the learnt outline. The multiedit network is also a revised VAE network that has excellent performance in long-text generation.

D. Evaluation Criteria

To assess the effectiveness of our proposed method, we conducted both automatic and human evaluations. For automatic evaluation, the bilingual evaluation understudy (BLEU) [52], recall-oriented understudy for gisting evaluation (ROUGE) [53], and Metric for Evaluation of Translation with Explicit ORdering (METEOR) [54] are adopted, the definitions are, respectively, defined as follows.

\[
\text{BLEU} = \frac{\sum_{n-gram \in \text{candidate}} \text{Count}_{\text{clip}}(n - \text{gram})}{\sum_{n-gram^\prime \in \text{candidate}} \text{Count}(n - \text{gram}^\prime)}
\]
where candidate refers to the candidate sentence generated by the model, \( \text{Count}_{\text{clip}}(n - \text{gram}) \) represents the count of \( n \)-grams in the candidate sentence clipped by the maximum count of \( n \)-grams in the reference sentences, \( \text{Count}(n - \text{gram}') \) refers to the count of \( n \)-grams in the candidate sentence without clipping.

2) ROUGE-N: The ROUGE metric evaluates the similarity between the generated text and the ground truth text based on the count of identical \( N \)-grams, written as

\[
\text{ROUGE}_N = \frac{\sum_{S \in \{\text{RS}\}} \sum_{n \in S} \text{Count}_{\text{match}}(\text{gram}_N)}{\sum_{S \in \{\text{RS}\}} \sum_{n \in S} \text{Count}(\text{gram}_N)}
\]

where “RS” refers to “reference summaries,” \( \text{Count}_{\text{match}}(\text{gram}_N) \) denotes the number of \( \text{gram}_N \) contained in both the generated text and the ground truth text, and \( \text{Count}(\text{gram}_N) \) denotes the total number of \( \text{gram}_N \) contained in the ground truth text.

3) ROUGE-L: The ROUGE-L metric quantifies the similarity between the generated text and the ground truth text based on the longest common subsequence (LCS), calculated as

\[
\text{ROUGE}_L = \frac{(1 + \beta^2) R_{\text{LCS}} P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 P_{\text{LCS}}}
\]  

where \( R_{\text{LCS}} = \text{LCS}(Y, \hat{Y})/m \) measures the recall of the generated text, \( m \) is the length of the ground truth text, \( P_{\text{LCS}} = \text{LCS}(Y, \hat{Y})/n \) measures the precision of the generated text, where \( n \) is the length of the generated text, and \( \beta \) is a control parameter to balance the importance of recall and precision.

4) METEOR: The METEOR metric measures the similarity between the machine-generated sentence and the reference sentences, written as

\[
\text{METEOR} = (1 - \alpha) \cdot \text{precision} + \alpha \cdot \text{recall} \cdot F_\beta
\]

where \( F_\beta \) refers to the harmonic mean of precision and recall, and \( \alpha \) measures the importance between precision and recall.

To evaluate the performance of the Chinese phrases, we selected the BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores as the evaluation metrics due to the small size of most Chinese phrases. Furthermore, we used ROUGE-1, ROUGE-2, and ROUGE-L to evaluate the output’s summary quality. To assess the output’s fluency and consistency, we asked human evaluators to rate 500 randomly selected financial reports from the News–Report dataset and 500 generated abstracts from the arXiv dataset on a scale of 1 to 4.

### E. Experimental Results

1) Automatic Evaluation: For automatic evaluation, we evaluate both the proposed approach as well as the compared models on two datasets. The experimental results demonstrate that our approach outperforms the baseline methods regarding the BLEU, ROUGE, and METEOR evaluation criteria, as shown in Tables II and III.

Table II displays the evaluation outcomes of all the compared techniques on the News–Report dataset. Our proposed CVAE-KD model outperforms all other models. The BLEU scores of the CVAE-KD model are significantly higher than the scores of the baseline models. Notably, the CVAE-KD model’s BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores are 2.76%, 8.92%, −5.04%, and 1.20% higher than the baseline model’s best scores and 32.13%, 57.13%, 40.53%, and 64.71% higher than their average scores. Among all the baseline models, the CVAE model and the multimedia model perform best, as they have distinctly higher scores compared to the other baseline models. These findings reveal that the CVAE related model is suitable for generating long text from short text and confirm that the proposed CVAE-KD model’s effectiveness outperforms the CVAE model.

Moreover, the ROUGE and METEOR evaluation criteria results show that the CVAE-KD model performs better than other models. The proposed model outperforms the second-best model by 17.28%, 0.30%, and 7.67% regarding the ROUGE-1, ROUGE-2, and ROUGE-L scores, respectively, and 37.59%, 46.70%, and 58.16% compared to the average scores. Furthermore, the proposed model obtains a higher METEOR score by 3.25% compared to the second-highest score and 33.67% relative to the average score. Similar observations could be found on the arXiv dataset in Table III, showing that the CVAE-KD model...
TABLE III
EVALUATION RESULTS OF ALL COMPARED METHODS WITH THE arXiv DATASET

| Methods                  | BLEU 1 | BLEU 2 | BLEU 3 | BLEU 4 | ROUGE 1 | ROUGE 2 | ROUGE L | METEOR 1 | METEOR 2 | METEOR L |
|--------------------------|--------|--------|--------|--------|---------|---------|---------|---------|---------|---------|
| Seq2seq                  | 17.24  | 7.82   | 2.02   | 0.49   | 21.00   | 3.95    | 12.57   | 12.92   |
| Seq2seq+Attn             | 20.93  | 9.03   | 2.46   | 0.63   | 23.29   | 4.40    | 14.49   |
| Pointer-generator        | 17.59  | 6.86   | 1.79   | 0.37   | 20.23   | 3.73    | 11.97   | 12.40   |
| Writing-editing network  | 12.57  | 5.01   | 1.21   | 0.19   | 20.64   | 3.56    | 10.97   | 11.70   |
| Plan-and-write           | 19.35  | 4.54   | 2.14   | 0.65   | 20.65   | 3.27    | 15.71   | 14.87   |
| CVAE                     | 20.73  | 9.11   | 2.67   | 0.74   | 23.04   | 4.50    | 14.30   | 14.58   |
| Multiedit                | 20.93  | 6.07   | 2.52   | 0.78   | 24.21   | 4.46    | 16.78   | 15.66   |
| CVAE-KD                  | 20.04  | 10.09  | 1.86   | 0.80   | 25.16   | 4.81    | 17.01   | 16.83   |

The average improvement: 24.69% 45.81% 21.00% 45.45% 15.07% 20.81% 23.12% 20.48%

Improvement than the highest score of baseline: 10.08% 10.76% −4.12% 2.56% 3.92% 6.89% 1.37% 6.19%

Note: Bold indicates the highest value of each column.

TABLE IV
EVALUATION RESULTS OF ALL COMPARED METHODS IN TERMS OF HUMAN SCORE

| Methods       | News–Reports Dataset | arXiv Dataset |
|---------------|----------------------|---------------|
| Seq2seq       | 2.20                 | 2.40          |
| Seq2seq+Attn  | 2.46                 | 2.49          |
| Pointer-Generator | 2.33              | 2.47          |
| Writing-Editing Network | 2.47          | 2.33          |
| Plan-and-Write | 2.86                 | 2.90          |
| CVAE          | 2.93                 | 3.00          |
| Multiedit     | 2.08                 | 2.66          |
| CVAE-KD       | 2.95                 | 3.10          |

The average improvement: 19.16% 18.90%

Improvement than the highest score of baseline: 0.68% 3.33%

Note: Bold indicates the highest value of each column.

model achieves the highest score compared to other methods across all evaluation metrics.

2) Human Evaluation: We also conducted the human evaluation experiment to further evaluate the fluency and consistency of the generated reports. The evaluation results of all approaches’ human scores are reported in Table IV. A higher score indicates a better model performance. The results in Table IV are consistent with the automatic evaluation, confirming that the proposed CVAE-KD outperforms all other models on all evaluation criteria. In the News–Reports dataset, the human scores of CVAE-KD are 0.68% higher than those of the CVAE model (the second-best model) and 19.05% higher than the average scores. Similarly, in the arXiv dataset, the human scores of CVAE-KD are 3.33% higher than the second-highest score and 22.14% higher than the average scores, respectively.

3) Case Study: To further evaluate the effects of different models, we have compared the reports generated by the proposed CVAE-KD model and the traditional CVAE model in Tables V and VI. Table V presents the results of the news–reports dataset, while Table VI showcases the arXiv dataset. The first row represents the input news, followed by the corresponding target report in the second row. The third and fourth rows show the generated reports by the CVAE model and our CVAE-KD approach, respectively. The length of the reports is set to 200, and the generated Chinese reports are translated to English using Google Translator, with the correct words highlighted in bold. Our observations from the generated results are as follows. First, the CVAE-KD approach exhibited more accurate results compared to the traditional CVAE, as indicated by a higher hit rate. This result underscores the effectiveness of KD in enhancing the decoder of CVAE-based models. Second, the generated reports by CVAE-KD were found to be more coherent than other approaches. For instance, “The growth rate of manufacturing investment has fallen” generated by the CVAE-KD report was more accurate than “the U.S. inflationary pressure” generated by CVAE. These observations suggest that the reports generated by CVAE-KD have better quality. However, we noticed a few repeated words or sentences in the CVAE-KD report, such as “We believe that we believe that we will” in the last sentence of Table V, which requires further research.

4) Effect of the Report Length: This experiment aims to evaluate whether the length of generated reports affects the model’s performance. Based on the previous experiment, we kept the length of the input news unchanged and set the length of the generated reports to 100, 150, and 200. The BLEU scores, ROUGE scores, and METEOR scores of the proposed CVAE-KD model generating reports of varying lengths are reported in Table VII. For simplicity, we denote report(200) as the generated report with a length of 200 words, and similarly report(100) and report(150).

As shown in Table VII, for the BLEU evaluation criterion, it is observed that the CVAE-KD model’s performance gradually deteriorates when generating longer reports in both datasets. In the News–Reports dataset, the BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores of report(200) are 2.5%, 4.8%, 9.1%, and
From the perspective of the U.S.’s own situation, the U.S. economic prospects are improving but the economic rebound will be less than expected. Private consumption in the U.S. depends on employment conditions and wage growth. When the number of employees is at a high level, the unemployment rate has dropped. At the same time, weaker U.S. inflation will further slow down wage growth. Therefore, subject to wage growth and declining inflation, the recovery of private consumption in the United States will be less than expected. Secondly, in terms of private investment, US housing inventories are at a low level, and housing prices continue to rise, which will stimulate the rapid growth of new housing construction and real estate investment; but another On the one hand, from the perspective of the inventory cycle, the U.S. manufacturing inventory replenishment is nearing completion. It is expected that the monthly inventory cycle will peak and fall, which will drag down manufacturing investment and therefore the U.S. fixed asset investment in the third quarter. Performance will be difficult to regain the strong first quarter.

we believe that, of, month month day forecast; ())) month forecast; (", and economic growth expectations and economic The growth rate fell back to—The economic growth rate fell back to the US dollar index, which was a month-on-month decline; the US dollar index was a month-on-month decline and the US dollar index was a quarter-on-quarter decline, and the year-on-year growth rate fell, but it will continue in the future. The growth rate of manufacturing is in line with “market expectations and policies, etc.” Under the policy: The central bank’s meeting on the market’s inflation in the middle of last month: This will be the index to increase inflation. Interest will be the main reason. The company will become a global enterprise and global enterprise field in the future. We believe that we believe that we will also have a global enterprise field in the future.

We describe a uniform all-sky survey of bright blazars, selected primarily by their flat radio spectra, that is designed to provide a large catalog of likely γ-ray active galactic nuclei (AGNs). The defined sample has 1625 targets with radio and X-ray properties similar to those of the EGRET blazars, spread uniformly across the b > 10 degree sky. We also report progress toward optical characterization of the sample; of objects with known R < 23.85% have been classified and 81% have measured redshifts. One goal of this program is to focus attention on the most interesting (e.g., high-redshift, high-luminosity...) sources for intensive multiwavelength study during the observations by the Large Area Telescope (LAT) on GLAST.

The Standard Model of the Standard Model (SM) Higgs bosons at the LHC is predicted to be dedicated to the Standard Model (SM). We also describe the prospects for probing the LHC and the ILC, including the full Higgsless and the ILC are discussions of the Standard Model (SM) Higgs bosons are discussed. of the sample are discussed for the LHC and future colliders. experiments are presented for the discovery of a new observable Higgs, of the results are presented for the LHC and future colliders. We also report the full range of the LHC, and the LHC are discussed. of the results are presented for the discovery of a new observable Higgs, and the LHC, are discussed. of the sample are presented for the discovery of a new generation of the Standard Model (SM) and formulae, are discussed. of the LHC, the sources for including the new physics of the Standard Model (SM) Higgs bosons are discussed. of the LHC, including the new physics study of the Standard Model.

We present a new all-sky survey for computing the full length scale of the braneworld model with a late-time null scalar field. We assume that the Gauss–Bonnet term is proportional to the instantaneous curvature of the universe with This sample takes into account the constraints on the metric tensor and the baryon number density, similar to the mass and energy density. We show that the model could be used to describe the inflation models with a non trivial and the universe. We also report the progress that the modified equation is modified to the equation of state of the optical characterization of the sample. It is found that there is no ZZ section in the angular parameter region. We also discuss the objects have been classified as a model for the scenario. of the model is indeed better than the curvature scalar curvature (where is the energy density, high redshift, which is the conformal Ricci curvature R on the metric does not depend on the curvature of the potential, the energy scales as the mass of the Kaluza–Klein radius. During the observations, for the energy scale of the field does not change.

13.2% higher than those of report(150), which is the second-best model. Furthermore, report(200) is the worst model with respect to all BLEU criteria. Similar results were observed in the arXiv dataset, where the CVAE-KD model’s performance decreased as the report length increased. We also noticed that the BLEU scores of all reports gradually decreased, which is consistent with our expectation that the model might not be accurate enough to generate longer phrases.
TABLE VII

RESULTS OF THE PROPOSED CAVE-KD ON GENERATING REPORTS OF DIFFERENT LENGTH

| Dataset   | Methods | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-1 | ROUGE-2 | ROUGE-L | METEOR |
|-----------|---------|--------|--------|--------|--------|---------|---------|---------|--------|
| News–Reports | 100     | 51.06  | 24.10  | 14.53  | 8.66   | 19.08   | 2.98    | 7.64    | 9.92   |
|            | 150     | 49.83  | 23.00  | 13.32  | 7.65   | 18.59   | 2.91    | 7.55    | 9.79   |
|            | 200     | 46.67  | 20.32  | 12.81  | 8.00   | 18.27   | 2.64    | 6.95    | 9.42   |
| arXiv     | 100     | 25.93  | 11.36  | 2.96   | 0.60   | 33.38   | 6.24    | 20.07   | 18.33  |
|           | 150     | 23.04  | 10.09  | 1.86   | 0.50   | 31.84   | 5.92    | 17.23   | 17.77  |
|           | 200     | 21.07  | 8.96   | 1.77   | 0.32   | 25.16   | 3.81    | 17.01   | 16.63  |

TABLE VIII

EVALUATION RESULTS OF THE PROPOSED MODEL WITH DIFFERENT BACKGROUND INFORMATION SCALE

| Q   | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-1 | ROUGE-2 | ROUGE-L | METEOR |
|-----|--------|--------|--------|--------|---------|---------|---------|--------|
| 5   | 15.45  | 6.02   | 1.96   | 0.67   | 16.93   | 2.94    | 6.12    | 8.57   |
| 10  | 15.20  | 5.70   | 1.79   | 0.60   | 16.62   | 2.70    | 5.67    | 8.40   |
| 15  | 18.27  | 7.08   | 2.14   | 0.68   | 20.02   | 3.31    | 7.16    | 10.18  |
| 20  | 16.75  | 6.01   | 1.89   | 0.41   | 18.27   | 2.64    | 6.95    | 9.42   |

Note: Bold indicates the highest value of each column.

TABLE IX

RESULTS OF THE ABLATION STUDY

| Methods          | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-1 | ROUGE-2 | ROUGE-L | METEOR |
|------------------|--------|--------|--------|--------|---------|---------|---------|--------|
| CV AE-KD (without KD) | 42.46  | 15.76  | 8.87   | 4.86   | 15.38   | 1.84    | 6.84    | 7.88   |
| CV AE-KD         | 46.67  | 20.32  | 12.81  | 8.00   | 18.27   | 2.64    | 6.95    | 9.42   |
| Improvement      | 9.92%  | 28.93% | 44.42% | 64.61% | 18.79%  | 43.48%  | 1.61%   | 19.54% |

Note: Bold indicates the highest value of each column.

From these observations, we can conclude that generating a long report from a short text is a difficult task. The longer the target report length, the worse the performance of the generated text. These observations are in agreement with the fact that the decoding ability of the employed GRU decreases as the loss of estimating the hidden states of GRU will accumulate monotonically. The results presented in Table VII verify the challenges associated with generating longer reports. Additionally, these objective evaluation results partially validate the effectiveness of the proposed KD-based approach.

5) Effect of the Range of Background Information: To determine the most appropriate range and evaluate the effect of the background information range in the proposed model, we varied the number of text in the news set, \( X \), and report set, \( Y \), denoted by \( Q \). The number of texts was set to 5, 10, 15, and 20, respectively. Additionally, we evaluated the proposed models with the BLEU, ROUGE, and METEOR metrics for different background information scales. The evaluation results are highlighted in Table VIII. It is well noticed that the model performs the best when \( Q = 15 \), indicating a sweet spot for the background information. If the number of texts is too low, the latent variables \( z_1 \) cannot learn sufficient background information for decoding the text. On the other hand, if there is an excess of background information, the KD component is unable to filter out the most relevant knowledge. Therefore, identifying a reasonable amount of background information is crucial.

6) Ablation Study of KD: To assess the effectiveness and importance of the KD component in our proposed model, we conducted an experiment by removing the KD component and revising the model loss function. This comparison model, named “CV AE-KD without KD,” learns background knowledge only through the approximation of \( z_1 \) and \( z_2 \) and does not include the KD component shown in Fig. 2. The loss function of the comparison model contains only the CVAE loss.

The comparison results are presented in Table IX. The table shows that the model performance of CVAE with a KD component is superior to that of the CVAE without the KD component, particularly for the BLEU scores. These results confirm that KD makes the Kullback–Leibler (KL) approximation of \( z_1 \) learn \( z_2 \) more effectively. Furthermore, this ablation study indicates that our proposed CVAE-KD model can partially overcome the problem of missing information in generating long text from short text. The KD component plays a significant role in the proposed model, with the primary purpose being to strengthen the learning effect of the background knowledge in the decoding process.

V. DISCUSSION

Our proposed model addresses the problem of generating financial reports using an end-to-end neural network architecture. As is known, such structure might be able to be easily adapted to other domains by simply fine-tuning the model using the domain-oriented dataset. By further considering the proposed teacher–student architecture, the background knowledge of other domains could be easily learnt through the teacher component, and this is verified by our experimental results on
the arXiv dataset, e.g., both automatic and human evaluation results achieve the SOTA performance. The intuitive design of incorporating background knowledge is to mimic the writing process of human writers. For these human writers, a piece of short news is too sufficient to write a long financial report. External domain knowledge is needed for this scenario. However, such background knowledge dynamically varies with time. Thus, it is a natural choice to build such background knowledge base in a real-time manner, i.e., querying similar reports from existing news-report dataset for each input news. The effect of the size of the background knowledge base is evaluated in the experiments. Although our proposed approach has achieved the SOTA performance, there still exist several limitations that need to be addressed in the near future. First, the factual error for the generated numbers remains a challenging issue even for the recent GPT-4. To fix such error, the new Bing searches the Web and queries the information. However, it is not suitable for an end-to-end model, and thus, a postprocessing step might be needed for this issue which needs further research exploration.

VI. CONCLUSION

Long-text generation is a significant yet challenging task, particularly for domain-specific applications, such as financial report generation. In this article, we propose a novel approach based on CVAEs, which first extracts external knowledge from the historical news-report data, and then, a teacher–student architecture is designed to guide the long financial report generation given a piece of short news. Extensive experiments have been performed on two public datasets and the proposed approach achieves superior performance when compared with both baselines and the SOTA models w.r.t. a number of evaluation criteria such as BLEU, ROUGE, METEOR, and human scores.

REFERENCES

[1] D. Shen et al., “Towards generating long and coherent text with multi-level latent variable models,” in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, 2019, pp. 2079–2089.
[2] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, and J. Wang, “Long text generation via adversarial training with leaked information,” in Proc. 32nd AAAI Conf. Artif. Intell./30th Innovative Appl. Artif. Intell. Conf./8th AAAI Symp. Educ. Artif. Intell., 2018, pp. 5141–5148.
[3] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. Le, and R. Salakhutdinov, “Transformer-XL: Attentive language models beyond a fixed-length context,” in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, 2019, pp. 2978–2988.
[4] Y. Gao, W. Zhao, and S. Eger, “SUPERF: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 1347–1354.
[5] Y. Liu and M. Lapata, “Text summarization with pretrained encoders,” in Proc. Conf. Empirical Methods Natural Lang. Process./9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), 2019, pp. 3730–3740.
[6] Z. Cao, W. Li, S. Li, and F. Wei, “Retrieve, rerank and rewrite: Soft template based neural summarization,” in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics, 2018, pp. 152–161.
[7] M. Mager et al., “GPT-2: A language-model-first approach for AMR-to-text generation,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 1846–1852.
[8] X. Feng, M. Liu, J. Liu, B. Qin, Y. Sun, and T. Liu, “Topic-to-essay generation with neural networks,” in Proc. 27th Int. Joint Conf. Artif. Intell., 2018, pp. 4078–4084.
[9] I. Sorodoc, K. Gulordava, and G. Boleda, “Probing for referential information in language models,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 4177–4189.
[10] S. Xu, H. Li, P. Yuan, Y. Wu, X. He, and B. Zhou, “Self-attention-guided copy mechanism for abstractive summarization,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 1355–1362.
[11] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in Proc. 2nd Int. Conf. Learn. Represent. (ICLR), Banff, AB, Canada, April 14–16, 2014, arXiv:1312.6114v10.
[12] A. D. McCarthy, X. Li, J. Gu, and N. Dong, “Addressing posterior collapse with mutual information for improved variational neural machine translation,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 8512–8525.
[13] T. Wang and X. Wan, “T-CVAE: Transformer-based conditioned variational autoencoder for story completion,” in Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI), 2019, pp. 5233–5239.
[14] X. Yang, X. Lin, S. Suo, and M. Li, “Generating thematic Chinese poetry using conditional variational autoencoders with hybrid decoders,” in Proc. 27th Int. Joint Conf. Artif. Intell., 2018, pp. 4539–4545.
[15] P. Yang, L. Li, F. Luo, T. Liu, and X. Sun, “Enhancing topic-to-essay generation with external commonsense knowledge,” in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, Jul. 2019, pp. 2002–2012.
[16] R. Zhang et al., “Improving adversarial text generation by modeling the distant future,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 2516–2531.
[17] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in Proc. NIPS Deep Learn. Representation Learn. Workshop, 2015.
[18] K. Cho et al., “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), Oct. 2014, pp. 1724–1734.
[19] J. Gu, Z. Lu, H. Li, and V. O. Li, “Incorporating copying mechanism in sequence-to-sequence learning,” in Proc. 54th Annu. Meeting Assoc. Comput. Linguistics, 2016, pp. 1631–1640.
[20] W. Hu, X. Zhang, and Y. Ren, “Generating financial reports from macro news via multiple edits neural networks,” in Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases, 2020, pp. 667–682.
[21] Y. Liu, Y. Wan, L. He, H. Peng, and P. S. Yu, “KGBART: Knowledge graph-augmented BART for generative commonsense reasoning,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 6418–6425.
[22] R. Liu et al., “Knowledge infused decoding,” in Proc. Int. Conf. Learn. Representations, 2022.
[23] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, “Generating sentences from a continuous space,” in Proc. 24th SIGKDD Conf. Comput. Natural Lang. Learn., Aug. 2016, pp. 10–21.
[24] T. Hosking, H. Tang, and M. Lapata, “Hierarchical sketch induction for paraphrase generation,” in Proc. 60th Annu. Meeting Assoc. Comput. Linguistics, 2022, pp. 2489–2501.
[25] S. Dai, Z. Gan, Y. Cheng, C. Tao, L. Carin, and J. Liu, “Apo-VAE: Text generation in hyperbolic space,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., 2021, pp. 416–431.
[26] E. Yang, M. Liu, D. Xiong, Y. Zhang, Y. Chen, and J. Xu, “Long text generation with topic-aware discrete latent variable model,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2022, pp. 8100–8107.
[27] I. Goodfellow et al., “Generative adversarial nets,” in Proc. 27th Int. Conf. Neural Inf. Process. Syst., 2014, pp. 2672–2680.
[28] M. J. Kasner and J. M. Hernández-Lobato, “GANs for sequences of discrete elements with the Gumbel-softmax distribution,” 2016, arXiv:1611.04051.
[29] K. Yang and D. Klein, “FUDGE: Controlled text generation with future discriminators,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., 2021, pp. 3531–3535.
[30] S. Dhar, N. D. Jana, and S. Das, “An adaptive-learning-based generative adversarial network for one-to-one voice conversion,” IEEE Trans. Artif. Intell., vol. 4, no. 1, pp. 92–106, Feb. 2023.
[31] Z. Liu, J. Wang, and Z. Liang, “CatGAN: Category-aware generative adversarial networks with hierarchical evolutionary learning for category text generation,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 8425–8432.
C.-W. Huang, J. H. Lim, and A. C. Courville, “A variational perspective on diffusion-based generative models and score matching,” in Proc. Adv. Neural Inf. Process. Syst., vol. 35, 2022, pp. 4328–4343.

X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, “Diffusion-lm improves controllable text generation,” in Proc. Int. Conf. Learn. Representations, 2023.

Z. He, T. Sun, Q. Tang, K. Wang, X. Huang, and X. Qiu, “Diffusion-BERT: Improving generative masked language models with diffusion models,” in Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (Long Papers), vol. 1, 2023.

H. Ramchoun, M. A. J. Idrissi, Y. Ghanou, and M. Ettaouil, “Multilayer perceptron: Architecture optimization and training with mixed activation functions,” in Proc. 2nd Int. Conf. Big Data, Cloud Appl., 2017, pp. 1–6.

A. Pagnoni, K. Liu, and S. Li, “Conditional variational autoencoder for neural machine translation,” 2018, arXiv:1812.04405.

D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” Proc. Int. Conf. Learn. Representations, 2014.

C. B. Clement, M. Bierbaum, K. P. O’Keeffe, and A. A. Alemi, “On the use of arXiv as a dataset,” 2019, arXiv:1905.00075.

H. He and J. D. Choi, “The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2021, pp. 5555–5577.

I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in Proc. 27th Int. Conf. Neural Inf. Process. Syst., vol. 2, 2014, pp. 3104–3112.

A. See, P. J. Liu, and C. D. Manning, “Get to the point: Summarization with pointer-generator networks,” in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics, 2017.

Q. Wang, Z. Zhou, L. Huang, S. Whitehead, and K. Knight, “Paper abstract writing through editing mechanism,” Proc. 56th Annu. Meeting Assoc. Comput. Linguistics, 2018, pp. 260–265.

L. Yao, N. Peng, R. Weischedel, K. Knight, D. Zhao, and R. Yan, “Plan-and-write: Towards better automatic storytelling,” in Proc. 33rd AAAI Conf. Artif. Intell./31st Innovative Appl. Artif. Intell. Conf/9th AAAI Symp. Educ. Adv. Artif. Intell., 2019, pp. 7378–7385.

T. Zhao, R. Zhao, and M. Eisen扎,” “Learning discourse-level diversity for neural dialog models using conditional variational autoencoders,” in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics, 2017, pp. 654–664.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, 2002, pp. 311–318.

C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in Proc. Text Summarization Branches Out, 2004, pp. 74–81.

S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization, 2005, pp. 65–72.

J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On faithfulness and factuality in abstractive summarization,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 1906–1919.

Ziao Wang received the bachelor’s degree in computer science from Dalian University of technology, Dalian, China, in 2015, and the master’s degree in computer science from Harbin Institute of Technology, Shenzhen, China, in 2018, where he is currently working toward the Ph.D. degree in computer science with the Department of Computer Science.

His research interests include the areas of natural language processing and graph neural network.

Yunpeng Ren received the bachelor’s degree in engineering from Harbin Institute of Technology, Shenzhen, China, in 2014, and the Ph.D. degree in science from Peking University, Beijing, China, in 2019.

He completed the postdoctoral research at the School of Computer Science, Harbin Institute of Technology, in 2022. He is currently conducting research work at Century Securities. His research interests include the areas of natural language processing, financial technology, and security market.

Xiaofeng Zhang received the M.Sc. degree in control theory and control engineering from Harbin Institute of Technology, Shenzhen, China, in 1999, and the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, China, in 2008.

He has worked with R&D Center, Peking University Founder Group, Beijing, China, and with E-business Technology Institute, Hong Kong University, Hong Kong, China. He is currently a Professor with the Department of Computer Science, Harbin Institute of Technology. His research interests include data mining, machine learning, and graph mining.

Yiyuan Wang received the bachelor’s degree in software engineering from Harbin Institute of Technology, Weihai, China, in 2019, and the master’s degree in computer science from Harbin Institute of Technology, Shenzhen, China, in 2022.

His research interest includes natural language generation.