Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.051745

ARTICLE

Generating Factual Text via Entailment Recognition Task

Jinqiao Dai, Pengsen Cheng and Jiayong Liu*

School of Cyber Science, Sichuan University, Chengdu, 610207, China

*Corresponding Author: Jiayong Liu. Email: ljy@scu.edu.cn

Received: 14 March 2024 Accepted: 14 May 2024 Published: 18 July 2024

ABSTRACT

Generating diverse and factual text is challenging and is receiving increasing attention. By sampling from the latent space, variational autoencoder-based models have recently enhanced the diversity of generated text. However, existing research predominantly depends on summarization models to offer paragraph-level semantic information for enhancing factual correctness. The challenge lies in effectively generating factual text using sentence-level variational autoencoder-based models. In this paper, a novel model called fact-aware conditional variational autoencoder is proposed to balance the factual correctness and diversity of generated text. Specifically, our model encodes the input sentences and uses them as facts to build a conditional variational autoencoder network. By training a conditional variational autoencoder network, the model is enabled to generate text based on input facts. Building upon this foundation, the input text is passed to the discriminator along with the generated text. By employing adversarial training, the model is encouraged to generate text that is indistinguishable to the discriminator, thereby enhancing the quality of the generated text. To further improve the factual correctness, inspired by the natural language inference system, the entailment recognition task is introduced to be trained together with the discriminator via multi-task learning. Moreover, based on the entailment recognition results, a penalty term is further proposed to reconstruct the loss of our model, forcing the generator to generate text consistent with the facts. Experimental results demonstrate that compared with competitive models, our model has achieved substantial improvements in both the quality and factual correctness of the text, despite only sacrificing a small amount of diversity. Furthermore, when considering a comprehensive evaluation of diversity and quality metrics, our model has also demonstrated the best performance.

KEYWORDS

Text generation; entailment recognition task; natural language processing; artificial intelligence

1 Introduction

Text generation models are increasingly expected to be naturalistic, diverse, and factually consistent due to their popularity and use in human life [1–3]. Recent works related to text generation design different models according to specific downstream task requirements, including machine translation [4], category text generation [5], and grammatical correctness [6]. These studies are used to generate different text by controlling some local attributes or formatting expressions. What has been relatively less explored in text generation research is the ability to control the generation of a current sentence in terms of the factual content and semantics of the entire sentence.

Controlling the generation of text based on its factual content and semantics is generally caught in a dilemma. On the one hand, to ensure the quality and fluency of the generated text, text generation models usually tend to generate safe, bland, repetitive sentences [7]. On the other hand, introducing other exciting research can help text generation models escape blandness and enhance the diversity of the generated text, such as generative adversarial network (GAN) [8] or variational autoencoder (VAE) [9]. However, this also inevitably introduces perturbations that cause the model to generate "hallucinating" words or sentences that are irrelevant to factual content [10]. Specifically, in the text summarization task, a study has statistically analyzed that 30% of the generated summaries are "hallucinating" and unusable [11]. Neither of the two cases mentioned above provides a compelling experience. Therefore, it is a challenging task to achieve a balance between text quality and diversity while simultaneously enhancing the factual correctness of the generated text.

Research on improving factual correctness primarily focuses on the summarization task and is divided into two different modes, as illustrated in Fig. 1. One is extracting factual descriptions from the source article and then generating summaries from the fact descriptions [10,11]. The other is introducing a natural language inference (NLI) system to generate more factual summaries [12,13]. However, these existing researches are aimed at paragraph-level and chapter-level text summarization tasks, which can enrich the output text with a large amount of input content. In sentence-level text generation scenarios, rich semantic information is missing, and the ability to generate factual text is less explored.

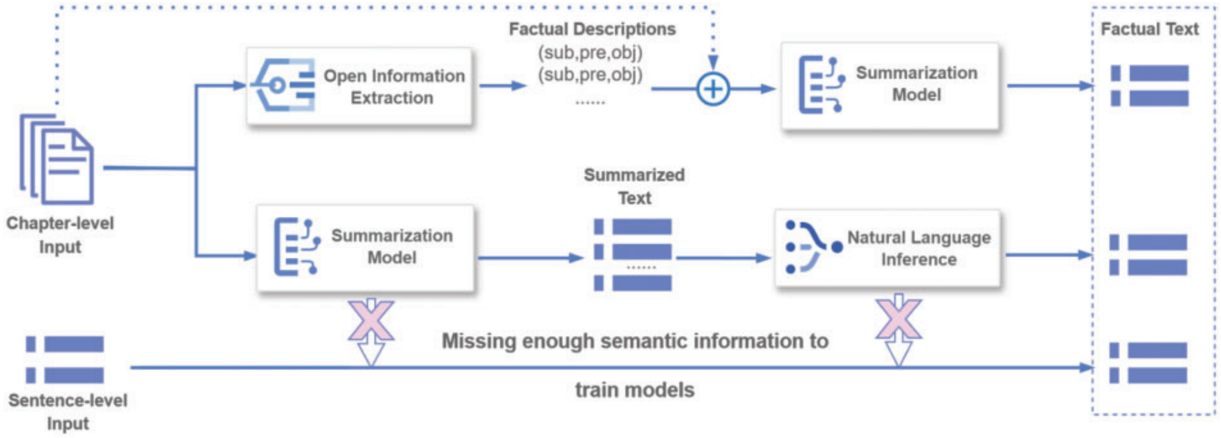

Figure 1: Generating factual text based on different input. In this process, open information extraction extracts triples representing factual descriptions from the input. The summarization model summarizes chapter-level input into sentence-level text. Natural language inference optimizes the summarized text to generate factual content

From the above analysis, it is evident that applying methods from summarization models to enhance factual correctness in sentence-level generative models still faces two challenges: 1) How to enable the model to acquire additional semantic information to generate diverse and factual text when provided with only sentence-level input; 2) How to integrate NLI systems into sentence-level models to enhance the factual correctness of generated text.

To address these challenges, we propose a novel model called fact-aware conditional variational autoencoder (FA-CVAE). To enrich the input semantics and increase the diversity, we introduce a conditional variational autoencoder (CVAE) [14] network as a generator to obtain the latent vector of the input text and generate text from the latent vector. In addition to VAE, the CVAE network introduces conditional variables, requiring the model to reconstruct training samples under various input conditions. Specifically, in the CVAE network for this work, the input facts are treated as conditional variables. A prior network is designed to derive the latent vector representing the semantic distribution of input facts. Subsequently, the recognition network integrates the input information with the training data to derive latent vectors that represent the semantic meaning of both texts. By calculating the KL divergence, CVAE aims to align the semantic distribution of the input facts closely with the recognition network, thereby generating text based on the input text. To improve the factual correctness, the traditional discriminator was improved via multi-tasking learning that allows the discriminator to perform two similar tasks simultaneously. One is a traditional real/false discriminate task [15], aiming to get a better qualified text. The other is an entailment recognition task [16] to discriminate the factual consistency of the input and output text. Furthermore, a penalty term is proposed to feed the discriminator results to the generator for backpropagation, forcing the generator to generate text more consistent with the inputs. The adversarial training strategy optimizes the generator and the discriminator alternately.

In summary, our contributions are as follows:

- A sentence-level model for generating factual text is proposed, employing the CVAE network to encode factual content and generate diverse text based on these input facts.
- A penalty term and a learning objective are proposed, employing multi-task learning to incorporate the discriminator with the textual entailment task and improving the factual correctness and quality of the generated text.
- Extensive experiments are performed on different datasets, demonstrating the capability of our model to achieve favorable results in balancing text quality, diversity, and improving the factual correctness.

2 Related Works

2.1 Neural Text Generation Models

Different works have extended the end-to-end generation framework to generate contentful and diverse text and proposed different text generation models. GAN-based methods have been widely used in text generation to address the exposure bias problem in RNN-based methods [17], SeqGAN [8] and LeakGAN [18] are early methods that have successfully used GAN in the field of text generation. By using policy gradients to update weights, they successfully solved the problem of the inability to backpropagate. More derivative models based on GAN have been proposed, such as SentiGAN [19] and CatGAN [20]. However, the text generated by GANs usually suffers from problems of mode collapse and training instability, the strategy of adversarial training is transferable to other methods.

Another method that is widely used along with GAN is VAE. Specifically, our work is more relevant to CVAE [14]. The goal of CVAE is to control a given attribute of the output text (for example, style, topic) by providing labels as additional input to a regular VAE. Hu et al. [21] combined VAE and attribute discriminators for learning the attribute content of input text. Other CVAE-based work has been improved to varying degrees for implementing different text generation tasks, such as advertising text generation [22], and story completion [23]. Recently, Russo et al. [24] proposed a CVAE text generation model that can control multiple attributes simultaneously. Inspired by providing additional labels to control specific attributes, we consider a complete sentence as an additional input to control the semantics and factual content of the output text.

2.2 Factual Correctness in Text

The more factual the generated text is, the more reliable the generation model will be. The study of factual correctness has only begun to attract interest in recent years. Cao et al. [11] and Goodrich et al. [25] extracted factual information from the original articles by introducing open information extraction (OIE), then generated summaries along with the article and the factual information. Recently, Zhang et al. [10] defined a factual correctness score function and adopted reinforcement learning to generate summaries with higher readability and factuality. However, the above methods rely more on the accuracy of factual information extraction, and if the factual information is biased, the summaries will also deviate from the articles.

Since the source text can entail the generated summaries, another way to improve factual correctness is to introduce an NLI system [16], such as textual entailment recognition. Li et al. [12] introduced multi-task learning to simultaneously perform text summarization and NLI. Instead of multi-task learning, Pasunuru et al. [26] proposed two NLI-based reward functions via reinforcement learning. Nevertheless, their experiments do not evaluate whether the techniques improve summarization correctness. Thus, Falke et al. [13] proposed a more direct approach to adopt a textual entailment model to evaluate the entailment relationship between articles and generated summaries, then re-rank the summaries on this basis. However, this approach only filters the generated summaries and does not directly improve the ability of the summarization model. Therefore, to improve the factual correctness of the generated text, it is crucial to integrate the NLI task as a fundamental element of the model and involve it in the training process. Moreover, summarization models can provide extensive training data at paragraph or chapter-level for implementing NLI task. For sentence-level models to utilize the NLI task effectively, one possible solution is to employ controllable text generation models like CVAE. This process includes integrating extra semantic information as conditions to support the model in utilizing the NLI task. Additionally, some researchers [27] leveraged large language models (LLMs) to enhance the factual correctness of the generated text.

Based on the aforementioned analysis, we extend CVAE with textual entailment to improve the factual correctness of generated text. Additional inputs are added to the CVAE model to enrich the semantic and factual content, and a penalty term is proposed to constrain the generator to generate factual text. Improving the factual correctness while retaining diverse text.

3 The Proposed Model

3.1 Task Definition and Model Overview

To define the factual text generation task reasonably, we refer to some concepts from the entailment recognition task [16]. Specifically, the entailment recognition task takes a pair of sentences and predicts whether the facts in the first sentence necessarily deduce the facts in the second. These two sentences are denoted as premise and hypothesis, respectively. The goal of text generation in this work is to take the premise as the input sentence $c = (c_1, c_2, \ldots, c_T)$ and generate the corresponding hypothesis $x = (x_1, x_2, \ldots, x_T)$ as the output sentence, which x is sampled using Eq. (1).

$$P(x|c) = \prod_{t=1}^{T} p(x_t|c, x < t) \tag{1}$$

However, generating x directly from c cannot fully reproduce the factual content contained in the hypothesis. We improve the task by fusing the semantics of the premise and hypothesis into a latent variable z to obtain a mathematical representation of the generated text finally x using Eq. (2).

$$P(x|z,c) = \prod_{t=1}^{T} p(x_t|z,c,x<t) \tag{2}$$

The overall framework is shown in Fig. 2, the model is mainly composed of two parts. The CVAE part is used in Section 3.2 to generate text with premise semantics and hypothesis semantics. The discriminator is used in Section 3.3 to improve the quality and factual correctness of the generated text. Based on the results of the discriminator, a penalty term is proposed in Section 3.4 to reconstruct the loss of the CVAE, forcing the CVAE to generate text consistent with the premise facts. Ultimately, the learning objective of FA-CVAE is proposed in Section 3.4. This objective integrates the KL term of the CVAE, the loss based on the real/fake discriminate task and the penalty term based on the entailment recognition task.

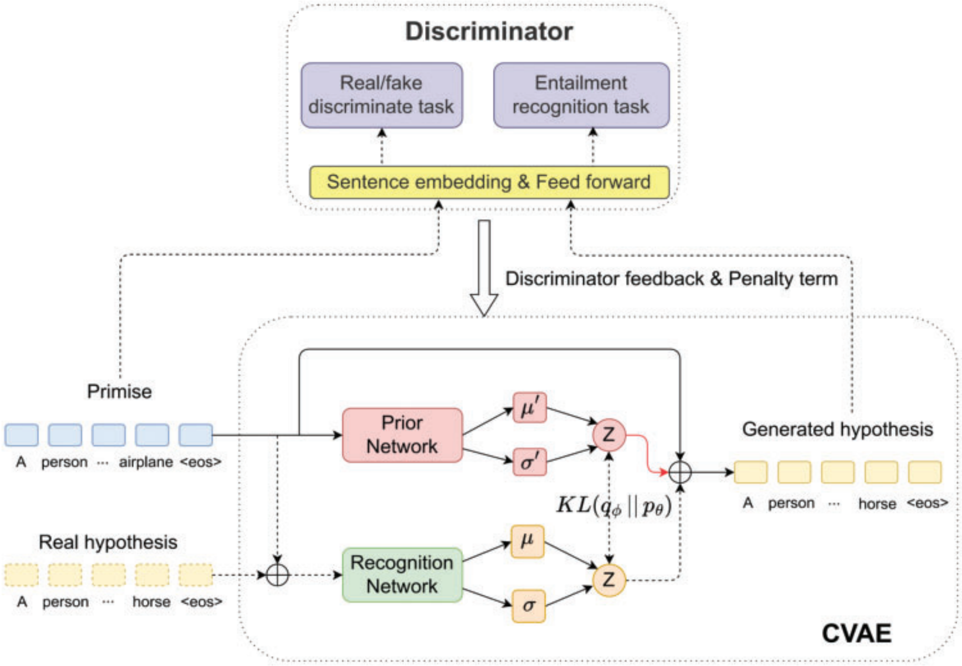

Figure 2: The overall process of FA-CVAE. The CVAE part utilizes a prior network and a recognition network to obtain latent variables for premises and hypotheses, thereby training the model to generate hypotheses based on premises. The discriminator part evaluates the generated hypotheses against the input premises and provides feedback to CVAE

It should be noted that in Fig. 2, μ and σ denote the mean and standard deviation of z, $\oplus$ denotes the concatenation of vectors. The whole CVAE and discriminator are applied in the training process except the solid red line. And in the prediction process, only the black and solid red lines are applied.

3.2 Conditional Variational Autoencoder (CVAE)

The CVAE focuses on fusing the text semantics of the premise and hypothesis, enabling the model to generate the corresponding hypothesis based on the premise. Therefore, we define the conditional distribution $p(x, z|c) = p(x|z, c)p(z|c)$, where x represents the hypothesis to assist in providing additional semantic content, and c represents the input text which is the premise. The training target is to approximate $p(x|z, c)$ and $p(z|c)$ via deep neural networks (parameterized using θ).

And the generative process of x can be described as sampling a latent variable z from the prior network $p_\theta(z|c)$ and then generating x through the decoder $p_\theta(x|z, c)$. Where z represents the true semantics distribution of the premise and hypothesis $p(z|x, c)$ and follows multivariate Gaussian distribution with a diagonal covariance matrix. Thus, we introduce Stochastic Gradient Variational Bayes (SGVB) framework [28] to train CVAE by maximizing the variational lower bound of the conditional log likelihood. Meanwhile introducing a recognition network $q_\phi(z|x, c)$ to approximate the semantics distribution $p(z|x, c)$. The variational lower bound of the CVAE is written using Eq. (3).

$$L_{CVAE}(\theta, \phi; x, c) = -KL\left(q_\phi(z|x, c)\,||p_\theta(z|c)\right) + \mathrm{E}_{q_\phi(z|c,x)}\left[\log p_\theta(x|z, c)\right] \le \log p(x|c) \tag{3}$$

In Eq. (3), $\log p_\theta(x|z, c)$ represents the reconstruction loss used during training to train the decoder with the recognition network. Through this training process, it enables the decoder of CVAE to generate hypotheses based on the latent variables z obtained from the recognition network $q_\phi(z|x, c)$. Within this training process, the decoder is unable to use the latent variables acquired from the prior network $p_\theta(z|c)$ to generate hypotheses. To address this limitation, the KL divergence is incorporated. CVAE minimizes the KL divergence between the latent variables of the prior network and the recognition network, this process is described in Eq. (3) as the KL term, $KL(q_\phi(z|x, c)\,||p_\theta(z|c))$. Finally, the KL term prompts the CVAE model to generate corresponding hypotheses based on the latent variables z obtained from the prior network.

To implement the above process, we designed the CVAE structure as shown in Fig. 2, which contains an encoder and a decoder. The encoder encodes premise c and hypothesis x into fixed-size vectors by a recurrent neural network (RNN). The last hidden states of premise and hypothesis are denoted as h^c and h^x, respectively. Then inspired by previous work [7], we assume z follows isotropic Gaussian distribution, and the recognition network $q_\phi(z|x, c) \sim N(\mu, \sigma^2 I)$ takes the concatenation of h^c and h^x as input. μ and σ are key parameters and computed using Eq. (4).

$$\begin{bmatrix} \mu \\ \log\left(\sigma^2\right) \end{bmatrix} = W_q \begin{bmatrix} h^x \\ h^c \end{bmatrix} + b_q \tag{4}$$

where W_q and b_q are trainable parameters. The prior network $p_\theta(z|c) \sim N(\mu', \sigma'^2 I)$ takes h^c as conditional input. The parameters of the prior network are calculated using Eq. (5) in RNN.

$$\begin{bmatrix} \mu' \\ \log\left(\sigma'^2\right) \end{bmatrix} = RNN_p(h^c) \tag{5}$$

The decoder then takes (z, c) as input and predicts the output hypothesis x by RNN. The loss of the CVAE is subsequently computed using Eq. (3) to train the CVAE network. It is worth noting that both prior network and reconstructed network are involved in the training process, and the hypothesis is predicted from $q_\phi(z|x, c) \sim N(\mu, \sigma^2 I)$ using the reparameterization trick [28]. And only the prior network is used in the testing process, the hypothesis is predicted from $p_\theta(z|c) \sim N(\mu', \sigma'^2 I)$. Finally, the generated hypothesis x is passed to the discriminator along with premise c.

3.3 Discriminator via Multi-Task Learning

The discriminator part is designed to improve the factual correctness and quality of the generated hypotheses. In traditional adversarial training, discriminator D only discriminates whether the text (generated by generator G) is fake text or not. In the task of this work, an entailment recognition task is introduced to distinguish the factual consistency between the generated hypotheses and the premises. To learn the losses of real/fake discriminate and entailment recognition tasks, multi-task learning is employed to improve the traditional discriminator. The detailed structure of the discriminator is shown in Fig. 3.

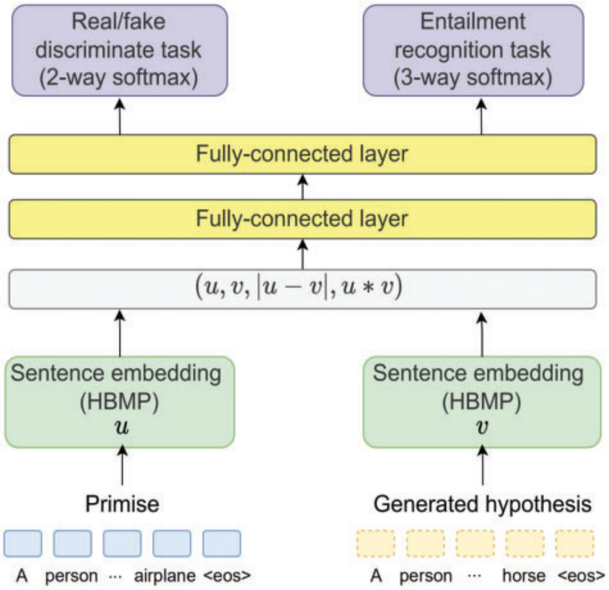
Real/fake discriminate task (2-way softmax)
Entailment recognition task (3-way softmax)
Fully-connected layer
Fully-connected layer
$(u, v, |u - v|, u * v)$
Sentence embedding (HBMP) u
Sentence embedding (HBMP) v
Primise
Generated hypothesis
A person ··· airplane <eos>
A person ··· horse <eos>

Figure 3: Illustration of the discriminator structure. The embedding layer and the fully-connected layer of the discriminator serve as shared layers for multi-task learning. During the training process, the discriminator takes input premises and generated hypotheses, and utilizes multi-task learning to simultaneously perform real/fake discrimination and entailment recognition tasks

In detail, the generated hypotheses and the premises are encoded as vectors u and v, respectively. To get better sentence representations, a hierarchical BiLSTM max pooling architecture (HBMP) [16] was introduced as a sentence embedding layer, and the output embeddings are concatenated as (u, v). Subsequently, the absolute difference $|u - v|$ and the inner product $(u * v)$ are individually calculated. These results are then combined to form representation q, as denoted by Eq. (6).

$$q = (u, v, |u - v|, u * v) \tag{6}$$

where $|u - v|$ represents the degree of closeness between two sentences in Euclidean space, and $(u * v)$ indicates the degree of similarity between the two sentences. This approach enhances the discriminator's ability to capture the relationship between the two sentences. Then feed q into a 3-layer multilayer perceptron, the first two layers utilize dropout and LeakyReLU activation functions. The last layer is divided into a 2-way softmax real/fake discriminate task and a 3-way softmax entailment recognition task. The real/fake discriminate task is used to distinguish whether the hypotheses are real or fake by minimizing Eq. (7).

$$L_R = -\mathrm{E}_{x\sim P_r}\left[\log D\left(x, c\right)\right] - \mathrm{E}_{(z,c)\sim P_g}\left[\log\left(1 - D\left(G\left(z, c\right), c\right)\right)\right] \tag{7}$$

In Eq. (7), $\log D(x, c)$ represents the loss of real/fake discriminate task obtained by feeding the real hypothesis and premise to the discriminator. $G(z, c)$ represents the process of generating hypotheses using the CVAE network described in Section 3.2. Based on this, $\log(1 - D(G(z, c), c))$ represents the loss obtained by feeding the generated hypothesis and premise to the discriminator. By minimizing L_R, the discriminator can refine its capacity to distinguish between real and fake text.

The entailment recognition task needs to correctly determine the relationship between premises and hypotheses for both real and generated samples by minimizing Eq. (8).

$$L_E = -\mathrm{E}_{x\sim P_r}\left[\log P\left(y|x, c\right)\right] - \mathrm{E}_{(z,c)\sim P_g}\left[\log P\left(y|G\left(z, c\right), c\right)\right] \tag{8}$$

In Eq. (8), $P(y|x, c)$ represents the probability, in the entailment recognition task, that the real hypothesis entails the same facts as the premise, denoted as the entailment score of the hypothesis. $P(y|G(z, c), c)$ represents the entailment score of the generated hypothesis. Similar to Eqs. (7) and (8) demonstrates the loss obtained by the discriminator while performing the entailment recognition task. By minimizing L_E, the discriminator can refine its capacity to distinguish the factual consistency between the hypotheses and the premises.

Based on the real/fake discriminate and entailment recognition results, a penalty term and the learning objective will be proposed in the next section.

3.4 Penalty Term and Learning Objective of FA-CVAE

To constrain the generator to generate hypotheses more consistent with the premises, we design a penalty term based on the entailment recognition results. The purpose of the penalty term is to penalize the generation of hypotheses with low entailment scores during the training process and encourage the model to generate hypotheses with high entailment scores. Specifically, the penalty term V_E^G first computes the specific value of the penalty term for the generated text based on Eq. (9).

$$V_E^G = 1 - P\left(y|G\left(z, c\right), c\right) \tag{9}$$

where $P(y|G(z, c))$ represents the entailment score of the generated hypotheses, and higher entailment scores indicate higher factual correctness of the generated hypotheses. Hence, based on the result of Eq. (9), it can be observed that the higher the factual correctness of the generated hypotheses, the lower the value of its penalty term. Subsequently, the penalty term acts as the weighting factor for the reconstruction loss in Section 3.2 of CVAE, resulting in a penalty-based reconstruct loss function, denoted as Eq. (10).

$$L_{GE} = V_E^G \cdot G\left(z, c\right) = G\left(z, c\right) - G\left(z, c\right) P\left(y|G\left(z, c\right), c\right) \tag{10}$$

From Eq. (10), it can be observed that when the generator generates hypotheses with high entailment scores, the penalty-based loss is small. Conversely, when generating hypotheses with low entailment scores, the loss is larger. Therefore, in order to minimize the reconstruction loss, the model tends to generate text with high entailment scores during the training process. Ultimately, with the assistance of the penalty term, the model effectively enhances the factual correctness of the generated hypotheses.

And for the real/fake discriminate task results, the loss L_{GR} based on the results is denoted as Eq. (11).

$$L_{GR} = -\mathrm{E}_{(z,c)\sim P_g}\left[\log\left(D\left(G\left(z, c\right), c\right)\right)\right] \tag{11}$$

By minimizing this loss, CVAE is encouraged to generate text that closely resembles real hypotheses as much as possible, thereby improving the quality of the generated text.

Based on the losses constructed by the above sections, the final loss of the discriminator is $L_E + L_R$. And the loss of generating hypotheses is finally obtained by integrating the variational lower bound of CVAE and the feedback of the discriminator using Eq. (12).

$$L = L_{KL} + L_{GE} + \lambda L_{GR} \tag{12}$$

In Eq. (12), L_{KL} represents the KL term of CVAE in Eq. (3), L_{GE} represents the loss based on the penalty term, L_{GR} represents the loss based on the real/fake discriminate task results utilizing adversarial learning, and λ represents the balance parameter that adjusts the influence of the real/fake discrimination task on the model. Ultimately, by minimizing the loss function L during training process, a model capable of generating diverse and high-quality hypotheses is obtained.

4 Experiment

4.1 Model Parameter Settings

For the CVAE module, the word embedding size is set to 300. RNNs in the encoder and decoder were set as a single-layer GRU with a hidden dimension size of 256 and a maximum length of 30 words. The latent size of z is set to 128, and Adam is selected as the optimizer. To avoid posterior collapse, a proportional-integral differential controller technique [29] is used in the KL term. For the discriminator module, the sentence embedding layer is the same as HBMP, and the balancing parameter λ is set to 0.1, following the conventional setting in previous work [21]. Other parameter settings are the same as CVAE. The generator and the discriminator are pre-trained for 100 epochs, and the adversarial training strategy is repeated until the FA-CVAE is converged. The experiment is constructe d using the PyTorch framework and trained on an NVIDIA GeForce RTX 3090. The source code of this paper can be accessed at: https://github.com/djqqiao/FA-CVAE (Accessed: May 13, 2024).

4.2 Datasets

From Sections 3.2 and 3.3, it can be inferred that in order to train the CVAE network and the entailment recognition task in this paper, the dataset used for training needs to ensure that the input text and the output text contain the same factual content. Since our work is based on the concept of NLI for text generation, where the inputs and outputs evolved from the entailment recognition task. Thus, the NLI dataset is also used as the training data for our text generation.

- **The Stanford Natural Language Inference (SNLI)** [30] is a collection of human-written English sentence pairs consisting of several premise-hypothesis sentence pairs with a relation label from (entailment, neutral, contradiction). The original dataset consisted of 570,702 sentence pairs after filtering out the sentence pairs that did not have a relationship label. The dataset thus generated contained 550,147 premise sentences and 550,147 hypothesis sentences with a vocabulary size of 20,188.
- **The Multi-Genre Natural Language Inference (MNLI)** [31] is a sub-task of General Language Understanding Evaluation (GLUE) [32] and has the same data structure as SNLI. The original dataset consisted of 392,702 sentence pairs after filtering out the sentence pairs that did not have a relationship label. The dataset thus generated contained 392,207 premise sentences and 392,207 hypothesis sentences with a vocabulary size of 75,731.

- **CommonGen** [33] provides a set of common concepts (e.g., dog, frisbee, catch, throw) and requires to generate sentences using these concepts (e.g., "a man throws a frisbee and his dog catches it"). This dataset is primarily aimed at relational reasoning and generating sentences with background commonsense knowledge. The CommonGen dataset comprises 32,651 training samples and 993 test samples, with each text containing 3 to 5 concepts.

For the SNLI and MNLI datasets, we randomly selected 1000 premises from the training set to use as the test set. During the training process, all premise-hypothesis pairs from the training sets were used as inputs to train the FA-CVAE model. During the testing process, only premises from the test set will be used as inputs to the model, while the hypotheses from the test set will serve as references for computing evaluation metrics. For the CommonGen dataset, the concept sets are treated as premises, while the rest of the process remains consistent with the SNLI and MNLI datasets.

4.3 Comparative Methods

This section compares our model with several text generation models and text summarization models. It is important to note that the text summarization task differs from our work in input content. In this task, we uniformly use the dataset provided in Section 4.2 for experiments and adjust the structure of some comparative experiments to fit the task of this paper.

- **AS2S-basline** [4]: A basic sequence-to-sequence model with attention mechanism, often used as a baseline model in various text generation tasks.
- **CVAE** [14]: Advanced controlled generation model that generates text by sampling from the latent space. Characteristically, the generated data is stable and not prone to mode collapse.
- **MTL-ERAML** [12]: A sequence-to-sequence model that introduces multi-task learning and an entailment-aware encoder and decoder. The proposed model can improve the factual correctness of the generated summaries.
- **FAR-ASS** [10]: A relatively new text summarization model that proposes two different reward terms and adopts reinforcement learning to improve the readability and factual correctness of generated summaries.
- **ChatProtect** [27]: A relatively new model investigates self-contradiction in large language models, offering a novel prompting-based framework to improve the factual correctness of generated.

To guarantee the consistency of the experiment, we made a slight adjustment to the original FAR-ASS by replacing the fact tuple with the entire premise and hypothesis sentence. The reward is then changed to the output of the entailment recognition task. In addition to models of related works, a constructed model named MTL-CVAE is further proposed based on the idea of related works for comparison experiments and the discussion section.

- **MTL-CVAE:** Inspired by the idea of MTL-ERAML, we replace the sequence-to-sequence model of MTL-ERAML with CVAE to verify the effectiveness of multi-task learning in the CVAE framework.

4.4 Evaluation Results of the Generated Text

To comprehensively evaluate the effectiveness of FA-CVAE and comparative methods. We evaluate the generated text based on text fluency, quality, diversity and overall performance.

- **Perplexity** (PPL) [34]: Perplexity measures the average probability of correctly predicting the next word in a sequence, often used to evaluate the fluency of language models.

- **ROUGE** [35]: ROUGE measures the recall of generated hypothesis by comparing overlapping n-grams between the generated text and references.
- **BLEU** [36]: BLEU measures the precision of the generated hypothesis. It uses the training set as a reference to calculate the BLEU value of the generated hypothesis.
- **Self-BLEU** [37]: Self-BLEU measures the diversity of the generated hypothesis. It calculates the BLEU value of the generated hypothesis by using all hypotheses generated from the premise as references.
- **Harmonic-BLEU** (BLEU$_{\mathrm{HA}}$): Harmonic-BLEU measures the F1 score of the generated hypothesis. It is the harmonic average value of BLEU and Self-BLEU, which is defined using Eq. (13).

$$\mathrm{BLEU}_{HA} = \frac{2 * \mathrm{BLEU} * (1 - \text{Self-BLEU})}{\mathrm{BLEU} + (1 - \text{Self-BLEU})} \tag{13}$$

In the aforementioned evaluation methods, higher BLEU and ROUGE scores indicate higher quality of the generated text, while lower PPL and Self-BLEU scores signify better fluency and diversity. In order to calculate the harmonic average correctly, 1-Self-BLEU is used as the calculation term of harmonic average instead of using Self-BLEU directly. Furthermore, we uniformly employ 2-gram overlapping units to calculate BLEU and ROUGE scores, denoted as BLEU-2, Self-BLEU-2, BLEU$_{\mathrm{HA}}$-2, and ROUGE-2, while utilizing the longest common subsequence for calculating ROUGE scores, denoted as ROUGE-L. For measuring perplexity, we follow the previous work [34] to calculate the PPL score. The experimental results are shown in Tables 1–3.

Table 1: Comparison of quality of generated hypotheses on SNLI

Method	BLEU-2	Self-BLEU-2	BLEU$_{\mathrm{HA}}$-2	ROUGE-2	ROUGE-L	PPL
AS2S-baseline	0.229	0.864	0.170	0.176	0.422	75.1
CVAE	0.219	0.711	0.249	0.164	0.385	83.4
MTL-ERAML	0.273	0.823	0.214	0.216	0.463	61.4
FAR-ASS	0.282	0.833	0.209	0.241	0.514	63.1
ChatProtect	0.142	**0.425**	0.234	0.103	0.326	**39.8**
MTL-CVAE	0.184	0.524	0.266	0.132	0.351	108
FA-CVAE	**0.314**	0.737	**0.286**	**0.261**	**0.567**	62.2

Table 2: Comparison of quality of generated hypotheses on MNLI

Method	BLEU-2	Self-BLEU-2	BLEU$_{\mathrm{HA}}$-2	ROUGE-2	ROUGE-L	PPL
AS2S-baseline	0.218	0.778	0.220	0.163	0.383	78.2
CVAE	0.207	0.752	0.225	0.145	0.417	81.3
MTL-ERAML	0.245	0.782	0.230	0.186	0.441	63.6
FAR-ASS	0.256	0.798	0.213	0.191	0.456	65.4
ChatProtect	0.130	0.376	0.216	0.094	0.312	**42.5**

(Continued)

Table 2 (continued)

Method	BLEU-2	Self-BLEU-2	BLEU_{HA}-2	ROUGE-2	ROUGE-L	PPL
MTL-CVAE	0.123	**0.346**	0.212	0.089	0.301	113
FA-CVAE	**0.278**	0.765	**0.254**	**0.219**	**0.483**	64.3

Table 3: Comparison of quality of generated hypotheses on CommonGen

Method	BLEU-2	Self-BLEU-2	BLEU_{HA}-2	ROUGE-2	ROUGE-L	PPL
AS2S-baseline	0.198	0.917	0.116	0.092	0.312	81.6
CVAE	0.181	0.731	0.316	0.164	0.385	85.4
MTL-ERAML	0.217	0.922	0.114	0.105	0.320	74.0
FAR-ASS	0.224	0.922	0.115	0.241	0.110	71.1
ChatProtect	0.450	0.422	0.231	0.113	0.315	**42.0**
MTL-CVAE	0.104	**0.510**	0.171	0.073	0.209	109
FA-CVAE	**0.252**	0.754	**0.248**	**0.121**	**0.342**	66.7

Compared with CVAE, FA-CVAE is approximately 5% to 7% higher in terms of the BLEU-2, ROUGE-2 and ROUGE-L metrics, proving that introducing discriminators via adversarial training can effectively improve the quality of the generated text. In terms of diversity, FA-CVAE loses a small portion of diversity but has a certain degree of improvement in quality. Compared with baseline and other related works. FA-CVAE has different degrees of improvement in both quality and diversity. Regarding diversity, text generated by FA-CVAE outperforms all related works, proving the effectiveness of using CVAE networks. Moreover, we noticed that ChatProtect demonstrates exceptional fluency and diversity, but it shows relatively poor performance on BLEU and ROUGE metrics. This inconsistency may stem from ChatProtect being constructed using large language models. These models, trained on vast corpora, can capture intricate contextual information in language, thereby enhancing the generation of more natural and coherent text. When the model is presented with identical inputs, the vast corpora act as a knowledge base, aiding the model in generating diverse text. However, BLEU and ROUGE metrics primarily evaluate differences between generated text and reference text. In such cases, the generation of diverse outputs can decrease similarity to the reference text, resulting in lower performance.

Finally, the overall evaluation shows that FA-CVAE achieves the best results in terms of harmonic average values. Demonstrating that FA-CVAE can generate text with high quality while preserving the diversity of the generated text. In addition, it is essential to note that the constructed model MTL-CVAE shows a large drop in both datasets, while a similar problem has been encountered in our recent research in category text generation [5]. The reasons for this will be discussed in detail in Section 4.7.

4.5 Ablation Study of the Discriminator

To validate the discriminator's effectiveness applied by multi-task learning, ablation experiments are conducted for two different loss functions used in adversarial training. The baseline model for the ablation experiment is CVAE, and different loss functions are added sequentially to the CVAE. The experimental results are shown in Tables 4–6.

Table 4: Ablation experiments on of adversarial training on SNLI

Method	BLEU-2	Self-BLEU-2	BLEU$_{\mathrm{HA}}$-2	ROUGE-2	ROUGE-L	PPL
CVAE	0.219	**0.711**	0.249	0.164	0.385	83.4
CVAE + L_{GE}	0.282	0.713	0.284	0.245	0.517	70.6
CVAE + L_{GR}	0.301	0.726	**0.286**	0.253	0.538	65.1
FA-CVAE	**0.314**	0.737	**0.286**	**0.261**	**0.567**	**62.2**

Table 5: Ablation experiments on of adversarial training on MNLI

Method	BLEU-2	Self-BLEU-2	BLEU$_{\mathrm{HA}}$-2	ROUGE-2	ROUGE-L	PPL
CVAE	0.207	**0.752**	0.225	0.145	0.417	81.3
CVAE + L_{GE}	0.240	0.758	0.241	0.181	0.437	72.5
CVAE + L_{GR}	0.254	0.761	0.246	0.189	0.452	67.6
FA-CVAE	**0.278**	0.765	**0.254**	**0.219**	**0.483**	**64.3**

Table 6: Ablation experiments on of adversarial training on CommonGen

Method	BLEU-2	Self-BLEU-2	BLEU$_{\mathrm{HA}}$-2	ROUGE-2	ROUGE-L	PPL
CVAE	0.181	**0.731**	0.316	0.164	0.385	85.4
CVAE + L_{GE}	0.223	0.743	0.238	0.110	0.324	71.2
CVAE + L_{GR}	0.245	0.749	0.247	0.115	0.338	68.0
FA-CVAE	**0.252**	0.754	**0.248**	**0.121**	**0.342**	**66.7**

Where L_{GE} denotes the reconstruction loss based on the penalty term. L_{GR} denotes the discriminative loss of the real/fake discriminator. In the ablation experiment, L_{GE} constraints the generator to generate text that matches the premise facts. L_{GR} constraints the generator to generate text with the same distribution as the real samples. It can be concluded that each of the two losses can improve the quality of the generated text, and CVAE + L_{GR} is slightly higher than the CVAE + L_{GE} model in terms of quality. This is because the reference text in the BLEU calculation is based on the real samples, L_{GR} is naturally more advantageous. FA-CVAE uses both losses, resulting in higher quality despite losing some diversity.

In general, using both losses separately effectively improves the quality of the generated text, and combining the two losses through a multi-task framework enables the model to generate text with higher quality while preserving the diversity of the generated text.

4.6 Factual Correctness of Generated Text

To effectively evaluate the factual correctness of the generated text, two approaches were taken to perform the evaluation. One is the fact score, which uses a pre-trained RoBERTa model [38] to evaluate the scores of the premises and the corresponding hypotheses. The other is manual evaluation, which randomly selects 100 samples from the generated hypotheses and employs five postgraduates to determine whether the generated hypotheses are consistent with the premises. The results are shown in Table 7.

Table 7: Comparison of quality of generated hypotheses on MNLI

Method	Factual score		Manual evaluation	
	SNLI	MNLI	SNLI	MNLI
AS2S-baseline	0.521	0.488	0.492	0.473
CVAE	0.508	0.471	0.474	0.462
MTL-ERAML	0.713	0.610	0.695	0.637
FAR-ASS	0.716	0.613	0.694	0.628
ChatProtect	0.737	0.627	0.714	0.646
CVAE + L_{GR}	0.707	0.602	0.681	0.606
CVAE + L_{GE}	0.721	0.619	0.709	0.632
FA-CVAE	**0.742**	**0.628**	**0.734**	**0.657**

In terms of factual scores, text generated by the baseline and CVAE models have little factual correctness. However, all models optimized for factual content were able to improve the factual score of the generated text significantly, with improvements ranging from 12% to 22%. Compared to other related works, text generated by FA-CVAE show a slight improvement in factual correctness. Demonstrating that using adversarial training and the discriminator can generate factual text consistent with the premises. The ablation experiments on FA-CVAE show that the introduction of the real/fake discriminate task and entailment recognition task can improve the factual correctness of the generated text. Using multi-task learning to combine these two tasks can further improve factual correctness.

In manual evaluation, factual correctness is lower than factual scores because some ambiguous sentence pairs are mostly marked as "neutral" or "contradictory" in the manual evaluation process. For example ("A man on a horse jumped over a broken down airplane.", "There is a person in the plane"), this sentence pair received a fact score of 0.79. However, it was mostly marked as "contradiction". By manual evaluation, FA-CVAE is still better than the baseline and other related works, consistent with the conclusions obtained from the factual score.

4.7 Possibility of Using Multi-Task Learning Directly in CVAE

Both multi-task learning and adversarial training use auxiliary tasks to improve the effectiveness of the original model. In this section, we try to replace the adversarial training strategy in CVAE and train the model directly using multi-task learning. The main task of multi-task learning is the text generation task dominated by the CVAE model, and the auxiliary task is the entailment recognition task. The overall structure of MTL-CVAE is similar to MTL-RAML, and the encoder of CVAE is used as the shared layer, sharing the same parameters.

The experimental results are shown in Tables 1 and 2, the quality of the generated text decreases dramatically. This is because training the generation and recognition tasks simultaneously introduces a significant interference to the KL term. In multi-task learning, the generation task training the posterior fits the prior distribution, while the recognition task training the posterior fits the distribution of recognition results. The conflict between these two tasks eventually leads to the inability of the KL term to converge [5]. The curve of the KL term is shown in Fig. 4.

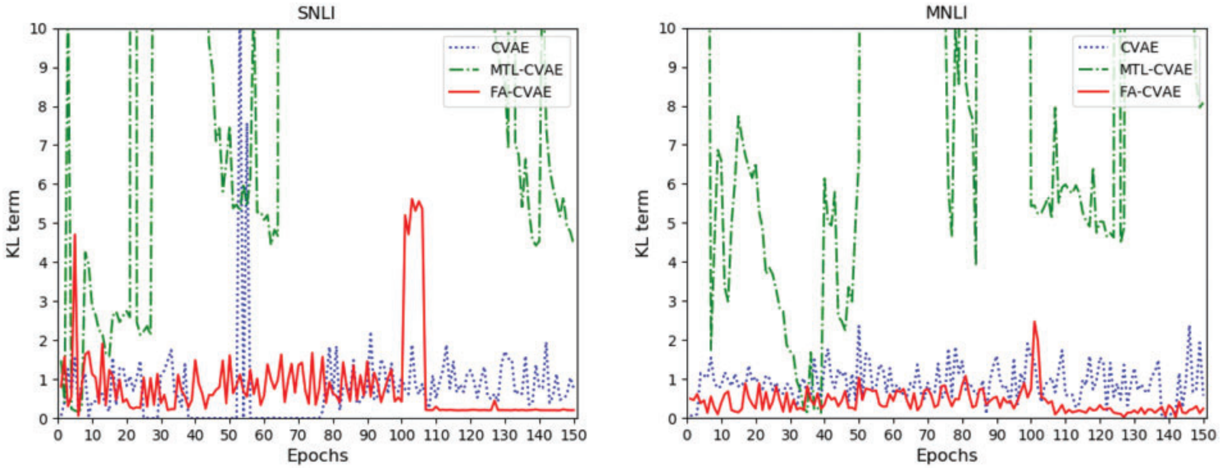

Figure 4: The KL term of different CVAE models in SNLI and MNLI datasets. The horizontal axis represents the epochs of training different CVAE models, the vertical axis represents the variation of the KL term in the CVAE loss during training. During the first 100 training epochs of FA-CVAE, it undergoes a pre-training process similar to CVAE. From epochs 100 to 150, FA-CVAE introduces a discriminator for adversarial training. It should be noted that due to the significant fluctuation range of MTL-CVAE, only the variation of the KL term within the range of 0 to 10 is displayed

The KL term of MTL-CVAE fluctuates drastically and is difficult to converge, resulting in significant noise in the generation task. In our recent work in category text generation [5], we found a similar problem, and the solution in that work was to use the VRNN structure and remove the KL term. To avoid changing the structure of CVAE while solving the fluctuation problem of the KL term, FA-CVAE adopts adversarial training, where the generation and recognition tasks are trained independently and separately, avoiding the influence of the recognition task on the KL term. It can be seen that during the pre-training phase of FA-CVAE, the fluctuation of the KL term is the same as CVAE. And during the adversarial training phase, the KL term first rises, then falls rapidly, and finally remains at a stable level. We attribute this situation to the optimization process of CVAE. During the initial stages of adversarial training, the CVAE network is influenced by the discriminator and attempts to focus more on deceiving it, this may lead to the cognitive network optimizing the latent variable z towards a certain direction, resulting in an increase in the KL term. As training progresses, the discriminator becomes more powerful, and better at distinguishing between generated samples and real ones. The generator may gradually adjust its strategy, learning how to balance different losses under adversarial learning. Therefore, as training continues, the KL term decreases.

Therefore, it is concluded that the multi-task learning framework cannot be used directly in CVAE but can be used on the discriminator in adversarial training to optimize the model indirectly.

4.8 Case Study of FA-CVAE

For a more visual illustration of the text generated by FA-CVAE, some of the generated texts are shown in Table 8. In case 1, FA-CVAE generates the hypothesis of “riding a horse” from the premise of “person on a horse”, and based on the premise of “jumps over a broken down airplane”, the hypothesis of "performing stunts" is deduced. In case 2, FA-CVAE can also generate the corresponding hypothesis of "two ladies were hugging" based on the factual content of "two women, holding food carryout containers, hug". In addition, FA-CVAE has generated some interesting content in "Hypothesis 3". In case 1, FA-CVAE deduces the hypothesis of "skydiving" based on "jump over". And in case 2, the hypothesis "the women are lovers" is deduced based on "hug each other". These hypotheses are an exaggeration of the facts. Although leading to the generation of hypotheses that are inconsistent with the facts, the object of hypotheses and premises are the same. We consider these hypotheses interesting and acceptable.

Table 8: Case study of FA-CVAE

	Case 1	Case 2
Premise	A person on a horse jumps over a broken down airplane.	Two women, holding food carryout containers, hug.
Golden reference	A person is outdoors, on a horse.	Two women hug each other.
Hypothesis 1	There is a person riding a white/green horse.	Two women are hugging each other.
Hypothesis 2	A person is performing stunts on a horse.	Two ladies were hugging.
Hypothesis 3	A person skydiving on horseback.	The women are lovers.

5 Conclusion

This paper investigates the factual correctness problem in sentence-level text generation models. To address this, we propose the FA-CVAE model. Essentially, FA-CVAE constructs a CVAE network and employs multi-task learning to integrate entailment recognition tasks within the discriminator. Subsequently, it employs adversarial training to provide feedback from the discriminator's results to the CVAE network. Through extensive and thorough experimentation, we have derived the following conclusions: 1) By introducing the CVAE network, it is possible to generate factual text (hypotheses) from sentence-level input, resulting in text of a certain quality and diversity. 2) The discriminator constructed using multi-task learning can effectively enhance the quality and factual correctness of text generated by the CVAE network. 3) When considering both the quality and diversity of generated text, FA-CVAE demonstrates favorable performance, indicating its ability to balance the quality and diversity of the generated text. 4) Incorporating entailment recognition tasks directly onto the CVAE through multi-task learning can result in the KL term being difficult to converge, thereby impacting the quality of the generated text.

This study successfully extends the application of the NLI system from paragraph or chapter-level summarization models to sentence-level generation models. However, the scope of this study was limited in terms of training data. FA-CVAE requires that the input and reference output in the training data contain identical factual content. Such constraints hinder its effective extension to all sentence-level text generation tasks. In future work, one potential approach to addressing this issue is to integrate our method with large language models. Leveraging the powerful corpora of large language models can provide abundant factual content for training entailment recognition tasks, thereby enhancing the factual correctness of the generated text.

Acknowledgement: We thank the anonymous reviewers for their useful comments.

Funding Statement: This research is supported by the Science and Technology Department of Sichuan Province (No. 2021YFG0156).

Author Contributions: The authors confirm contribution to the paper as follows: Jinqiao Dai: Methodology, Software, Validation, Investigation, Data Curation, Visualization, Writing-Original Draft. Pengsen Cheng: Validation, Investigation, Data Curation, Writing-Review & Editing. Jiayong Liu: Resources, Supervision, Project Administration, Funding Acquisition. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets analyzed during the current study are available in the SNLI and MNLI repository, the links to these datasets are https://nlp.stanford.edu/projects/snli/ and https://gluebenchmark.com/tasks. The Source code that supports the findings of this study are available on request from the first author or the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] E. Reiter and R. Dale, "Building applied natural language generation systems," *Nat. Lang. Eng.*, vol. 3, no. 1, pp. 57–87, 1997. doi: 10.1017/S1351324997001502.

[2] Z. W. Ji, N. Lee, and R. Frieske, "Survey of hallucination in natural language generation," arXiv preprint arXiv:2202.03629, 2022.

[3] H. Li, W. J. Yu, and W. T. Ma, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," arXiv preprint arXiv:2311.05232, 2023.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[5] P. S. Cheng, J. Q. Dai, and J. Y. Liu, "CatVRNN: Generating category text via multi-task learning," *Knowl.-Based Syst.*, vol. 244, no. 2, pp. 108491, 2022. doi: 10.1016/j.knosys.2022.108491.

[6] J. Ji, Q. Wang, K. Toutanova, and Y. Gong, "A nested attention neural hybrid model for grammatical error correction," arXiv preprint arXiv:1707.02026, 2017.

[7] T. Zhao, R. Zhao, and M. Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *Proc. 57th Annu. Meet. Assoc. Comput. Linguist. (ACL'17)*, Vancouver, Canada, 2017, pp. 654–664.

[8] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence generative adversarial nets with policy gradient," in *Proc. 31st Proc. AAAI Conf. Art. Intell. (AAAI'17)*, 2017, vol. 31, no. 1.

[9] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz and S. Bengio, "Generating sentences from a continuous space," in *Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn. (CoNLL'17)*, Berlin, Germany, 2016, pp. 10–21.

[10] M. Zhang, G. Zhou, W. Yu, and W. Liu, "FAR-ASS: Fact-aware reinforced abstractive sentence summarization," *Inf. Process. Manage.*, vol. 58, no. 3, pp. 102478, 2021. doi: 10.1016/j.ipm.2020.102478.

[11] Z. Cao, F. Wei, and W. Li, "Faithful to the original: Fact aware neural abstractive summarization," arXiv preprint arXiv:1711.04434, 2017.

[12] H. Li, J. Zhu, and C. Zong, "Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization," in *Proc. 27th Int. Conf. Comput. Linguist. (COLING'18)*, Santa Fe, New Mexico, USA, 2018, pp. 1430–1441.

[13] T. Falke and L. F. R. Ribeiro, "Ranking generated summaries by correctness: An interesting but challenging application for natural language inference," in *Proc. 57th Annu. Meet. Assoc. Comput. Linguist. (ACL'19)*, Florence, Italy, 2019, pp. 2214–2220.

[14] K. Sohn, H. Lee, and X. Yan, " Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, 2015, vol. 28, pp. 3483–3491.

[15] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020. doi: 10.1145/3422622.

[16] A. Talman, A. Yli-Jyrä, and J. Tiedemann, "Sentence embeddings in nli with iterative refinement encoders," *Nat. Lang. Eng.*, vol. 25, no. 4, pp. 467–482, 2019. doi: 10.1017/S1351324919000202.

[17] S. Bengio, O. Vinyals, and N. Jaitly, "Schedules sampling for sequence prediction with recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2015, vol. 28, pp. 1771–1179.

[18] J. Guo, S. Lu, H. Cai, W. Zhang, and Y. Yu, "Long text generation via adversarial training with leaked information," in *Proc. 32nd Proc. AAAI Conf. Art. Intell. (AAAI'18)*, 2018, vol. 32, no. 1.

[19] K. Wang and X. Wan, "SentiGAN: Generating sentimental text via mixture adversarial networks," in *Proc. 27th Int. Joint Conf. Art. Intell. (IJCAI'18)*, 2018, pp. 4446–4452.

[20] Z. Liu, J. Wang, and Z. Liang, "CatGAN: Category-aware generative adversarial networks with hierarchical evolutionary learning for category text generation," in *Proc. 34th Proc. AAAI Conf. Art. Intell. (AAAI'20)*, 2020, vol. 34, no. 5, pp. 8425–8432.

[21] Z. T. Ting, Z. C. Yang, and X. D. Liang, "Toward controlled generation of text," in *Proc. 34th Int. Conf. Mach. Learn. (PMLR'17)*, 2017, vol. 70, pp. 1587–1596.

[22] Z. H. Shao, M. L. Huang, and J. T. Wen, "Long and diverse text generation with planning-based hierarchical variational model," in *Proc. 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP'19)*, Hong Kong, China, 2019, pp. 3257–3268.

[23] Z. Shao, M. Huang, J. Wen, and W. Xu, "Transformer-based conditioned variational autoencoder for story completion," in *Proc. 29th Int. Joint Conf. Art. Intell. (IJCAI'19)*, 2018, pp. 5233–5239.

[24] G. Russo, N. Hollenstein, and N. Musat, "Control, generate, augment: A scalable framework for multi-attribute text generation," in *Findings Assoc. Comput. Linguist. (EMNLP'20)*, 2020, pp. 351–366.

[25] B. Goodrich, V. Rao, and P. J. Liu, "Assessing the factual accuracy of generated text," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. (KDD'19)*, New York, USA, 2019, pp. 166–175.

[26] R. Pasunuru and M. Bansal, "Multi-reward reinforced summarization with saliency and entailment," in *Proc. 2018 Conf. N. Am. Chapter Assoc. Comput. Linguist. (NAACL'18)*, New Orleans, Louisiana, USA, 2018, pp. 646–653.

[27] N. Mundler, J. X. He, and S. Jenko, "Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation," arXiv preprint arXiv:2305.15852, 2024.

[28] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2022.

[29] H. J. Shao, S. C. Yao, and D. C. Shun, "ControlVAE: Controllable variational autoencoder," in *Proc. 37th Int. Conf. Mach. Learn. (PMLR'20)*, 2020, vol. 119, pp. 8655–8664.

[30] S. R. Bowan, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proc. 2015 Conf. Empir. Methods Nat. Lang. Process. (EMNLP'15)*, Lisbon, Portugal, 2015, pp. 632–642.

[31] A. Williams, N. Nangia, and S. R. Bowan, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proc. 2018 Conf. N. Am. Chapter Assoc. Comput. Linguist. (NAACL'18)*, New Orleans, Louisiana, USA, 2018, pp. 1112–1122.

[32] A. Wang, A. Singh, and J. Michael, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. (EMNLP'18)*, Brussels, Belgium, 2018, pp. 353–355.

[33] B. Y. C. Lin, W. C. C. Zhou, and M. Shen, "CommonGen: A constrained text generation challenge for generative commonsense reasoning," in *Findings Assoc. Comput. Linguist. (EMNLP'20)*, 2020, pp. 1823–1840.

[34] J. B. Zhao, Y. Kim, and K. Zhang, "Adversarially regularized autoencoders," in *Proc. 35th Int. Conf. Mach. Learn. (PMLR'18)*, 2018, vol. 80, pp. 5902–5911.

[35] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, *Findings Assoc. Comput. Linguist*, Barcelona, Spain, 2004, pp. 74–81.

[36] K. Papineni, S. Roukos, T. Ward, and T. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meet. Assoc. Comput. Linguist. (ACL'02)*, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318.

[37] R. Shu, H. Nakayama, and K. Cho, "Generating diverse translations with sentence codes," in *Proc. 57th Annu. Meet. Assoc. Comput. Linguist. (ACL'19)*, Florence, Italy, 2019, pp. 1823–1827.

[38] Y. H. Liu, M. Ott, and N. Goyal, "RoBERTa: A Robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.