Topic-to-Essay Generation with Comprehensive Knowledge Enhancement

Zhiyue Liu, Jiahai Wang(✉), and Zhenghong Li

School of Computer Science and Engineering, Sun Yat-sen University
Guangzhou, China
{liuzhy93,lizhh98}@mail2.sysu.edu.cn, wangjiah@mail.sysu.edu.cn

Abstract. Generating high-quality and diverse essays with a set of topics is a challenging task in natural language generation. Since several given topics only provide limited source information, utilizing various topic-related knowledge is essential for improving essay generation performance. However, previous works cannot sufficiently use that knowledge to facilitate the generation procedure. This paper aims to improve essay generation by extracting information from both internal and external knowledge. Thus, a topic-to-essay generation model with comprehensive knowledge enhancement, named TEGKE, is proposed. For internal knowledge enhancement, both topics and related essays are fed to a teacher network as source information. Then, informative features would be obtained from the teacher network and transferred to a student network which only takes topics as input but provides comparable information compared with the teacher network. For external knowledge enhancement, a topic knowledge graph encoder is proposed. Unlike the previous works only using the nearest neighbors of topics in the commonsense base, our topic knowledge graph encoder could exploit more structural and semantic information of the commonsense knowledge graph to facilitate essay generation. Moreover, the adversarial training based on the Wasserstein distance is proposed to improve generation quality. Experimental results demonstrate that TEGKE could achieve state-of-the-art performance on both automatic and human evaluation.

Keywords: Topic-to-essay generation · Knowledge transfer · Graph neural network · Adversarial training.

1 Introduction

Topic-to-essay generation (TEG) is a challenging task in natural language generation, which aims at generating high-quality and diverse paragraph-level text under the theme of several given topics. Automatic on-topic essay generation would bring benefits to many applications, such as news compilation [10], story generation [3], and intelligent education. Although some competitive results for TEG have been reported in the previous works using deep generative models [4, 15, 20], the information gap between the source topic words and the targeted essay blocks their models from performing well. The comparison of information flow between
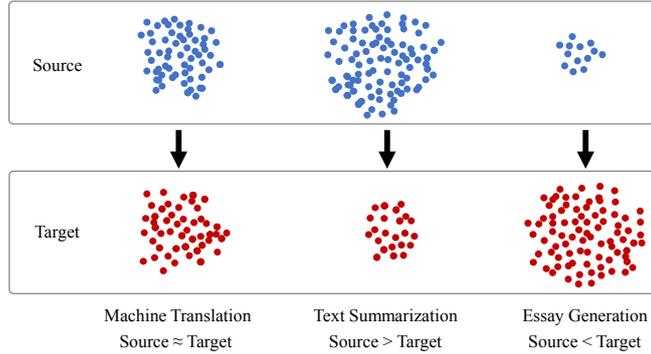
Fig. 1. Toy illustration of the information volume on different text generation tasks.

TEG and other text generation tasks is illustrated in Fig. 1 [20]. For machine translation and text summarization, the source provides enough information to generate the targeted text. However, for the TEG task, the information provided by only the topic words is much less than that contained in the targeted text during generation, making the generated essays low-quality.

The proper utilization of various topic-related knowledge is essential to enrich the source information, which has not been sufficiently explored. Incorporating the external knowledge from related common sense into the generation procedure is an efficient way to improve the TEG performance. However, in the commonsense knowledge graph, previous works [15,20] only consider the nearest neighbor nodes of topic words, and neglect the multi-hop neighbors which would bring more structural and semantic information. Moreover, without considering external knowledge, their models cannot fully exploit the relation between topics and essays to assist the generation procedure.

This paper proposes a topic-to-essay generation model with comprehensive knowledge enhancement, named TEGKE. By extracting both internal and external knowledge, TEGKE greatly enriches the source information. Besides, the adversarial training based on the Wasserstein distance is proposed to further enhance our model. Thus, there are three key parts, including internal knowledge enhancement, external knowledge enhancement, and adversarial training.

For internal knowledge enhancement, our model is based on the auto-encoder framework including a teacher network and a student network. Inspired by the conditional variational auto-encoder (CVAE) framework, the teacher network takes both topic words and related essays as source information to get informative latent features catching the high-level semantics of the relation between topics and essays. Then, a decoder could better reconstruct the targeted essay conditional on these features. Since only topic words could be used as the input source during inference, the informative features (i.e., internal knowledge) from the teacher network would be transferred to the student network. Different from CVAE that trains the recognition network and the prior network to be close to each other in the latent space, the teacher network in TEGKE maintains an

independent training procedure. Then, the student network is forced to be close to the teacher network. That is, the student could take only topics as input but output comparable informative latent features compared with the teacher.

For external knowledge enhancement, ConceptNet [17] is employed as the commonsense knowledge base. Different from the previous works only using the nearest neighbors of topics, a topic-related knowledge graph is extracted from ConceptNet, which consists of multi-hop neighbors from the source topic words. Then, a topic knowledge graph encoder is proposed to perform on the multi-hop knowledge graph. It employs a compositional operation to obtain graph-aware node representations (i.e., external knowledge), which could conclude the structural information and the semantic information. The external knowledge is involved in the essay generation and helps select a proper decoding word.

Moreover, a discriminator is introduced for adversarial training. For alleviating the discrete output space problem of text, previous works adopt the adversarial training based on reinforcement learning (RL), which has the drawbacks of less-informative reward signals and high-variance gradients [1]. In contrast, this paper proposes to directly optimize the Wasserstein distance for the adversarial training, which avoids the problem of vanishing gradients and provides strong learning signals [5]. Based on the Wasserstein distance, the discriminator could operate on the continuous valued output instead of discrete text [18]. For aligning essays with the related topics, topics are combined with generated essays as generated samples and combined with targeted essays as real samples. By the minimax game, the discriminator would provide an informative learning signal guiding our model to generate high-quality essays.

In summary, our contributions are as follows:

– A topic-to-essay generation model is proposed based on the knowledge transfer between a teacher network and a student network. The teacher network could obtain informative features for the student network to learn, making the student network provide abundant information with only topics as the input source.
– A topic knowledge graph encoder is proposed to perform on the multi-hop knowledge graph extracted from the commonsense base. It helps our model exploit the structural and semantic information of the knowledge graph to facilitate essay generation. Moreover, a discriminator is introduced to improve generation quality by the adversarial training based on the Wasserstein distance.
– Experimental results on both automatic evaluation and human evaluation demonstrate that TEGKE could achieve better performance than the state-of-the-art methods.

2   Related Work

As a text generation task, TEG aims at generating high-quality and diverse paragraph-level text with given topics, which has drawn more attention. This task is first proposed by Feng et al. [4], and they utilize the coverage vector to

integrate topic information. For enriching the input source information, external commonsense knowledge has been introduced for TEG [15, 20]. Besides, Qiao et al. [15] inject the sentiment labels into a generator for controlling the sentiment of a generated essay. However, during essay generation, previous works [15, 20] only consider the nearest neighbors of topic nodes in the commonsense knowledge graph. This limitation blocks their models from generating high-quality essays. For better essay generation, this paper makes the first attempt to utilize both structural and semantic information from the multi-hop knowledge graph.

Poetry generation is similar to TEG, which could be regarded as a generation task based on topics. A memory-augmented neural model is proposed to generate poetry by balancing the requirements of linguistic accordance and aesthetic innovation [24]. The CVAE framework is adopted with adversarial training to generate diverse poetry [12]. Yang et al. [21] use hybrid decoders to generate Chinese poetry. RL algorithms are employed to improve the poetry diversity criteria [22] directly. Different from poetry generation showing obvious structured rules, the TEG task needs to generate a paragraph-level unstructured plain text, and such unstructured targeted output brings severe challenges for generation.

The RL-based adversarial training [6, 23] is used to improve essay quality in previous works [15, 20]. However, the noisy reward derived from the discriminator makes their models suffer from high-variance gradients. In contrast, our model directly optimizes the Wasserstein distance for the adversarial training without RL, achieving better generation performance.

3   Methodology

3.1   Task Formulation

Given a dataset including pairs of the topic words $\mathbf{x} = (x_1, ..., x_m)$ and the related essay $\mathbf{y} = (y_1, ..., y_n)$, for solving the TEG task, we want a θ-parameterized model to learn each pair from the dataset and generate a coherent essay under given topic words, where the number of essay words n is much larger than that of topic words m. Then, the task could be formulated as obtaining the optimal model with $\hat{\theta}$ which maximizes the conditional probability as follows:

$$\hat{\theta} = \arg\ \max_\theta P_\theta(\mathbf{y}|\mathbf{x}). \tag{1}$$

3.2   Model Description

Our TEGKE is based on the auto-encoder framework, utilizing both internal and external knowledge to enhance the generation performance. As shown in Fig. 2, the model mainly contains three encoders (i.e., a topic encoder, an essay encoder, and a topic knowledge graph encoder) and an essay decoder. A discriminator is introduced at the end of the essay decoder for adversarial training.

For internal knowledge enhancement, the topic encoder and the essay encoder encode the topic words and the targeted essay sequence as x_{enc} and y_{enc}, respectively. The latent features z_1 and z_2 are obtained from a teacher network
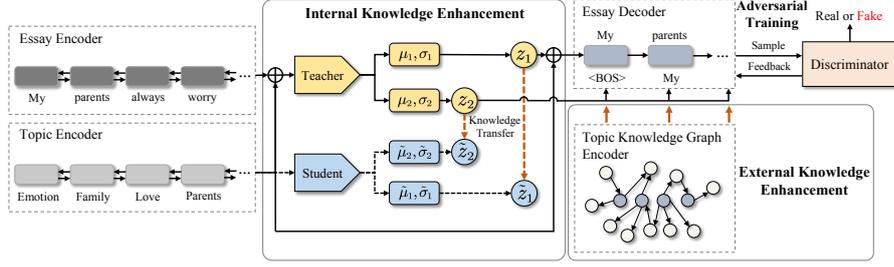
Fig. 2. Overview of the proposed model. Our model uses the teacher network for training (black solid arrows), and the student network for inference (black dotted arrows). The student network learns the latent features from the teacher network (red dotted arrows) for internal knowledge enhancement. The information from the topic knowledge graph encoder is integrated at each decoding step (red solid arrows) for external knowledge enhancement. During adversarial training, the generated essays are fed to the discriminator which provides learning signals as feedback.

taking both x_{enc} and y_{enc} as input. Then, a student network, which takes x_{enc} solely as input, produces $\tilde{z}_1$ and $\tilde{z}_2$ to learn from z_1 and z_2 as internal knowledge, respectively. The essay decoder would generate a topic-related essay by receiving the latent features from the teacher network during training or those from the student network during inference.

For external knowledge enhancement, the multi-hop topic knowledge graph is constructed from the commonsense knowledge base, ConceptNet. Then, the topic knowledge graph encoder could represent the topic-related structural and semantic information as external knowledge to enrich the source information. The extracted external knowledge is attentively involved in each decoding step of the essay decoder to help select proper words and boost generation performance.

Through the adversarial training based on the Wasserstein distance, the discriminator could make the generated essay more similar to the targeted essay, which improves essay quality.

Topic Encoder and Essay Encoder. The topic encoder employs a bidirectional gated recurrent unit (GRU) [2], which integrates the information of the topic sequence from both forward and backward directions. The topic encoder reads the embeddings of topic words $\mathbf{x}$ from both directions and obtains the hidden states for each topic word as follows:

$$\overrightarrow{h_i^x} = \overrightarrow{\mathrm{GRU}}(\overrightarrow{h_{i-1}^x}, e(x_i)), \ \overleftarrow{h_i^x} = \overleftarrow{\mathrm{GRU}}(\overleftarrow{h_{i+1}^x}, e(x_i)), \tag{2}$$

where $e(x_i)$ is the embedding of x_i. The representation of the i-th topic is obtained as $h_i^x = [\overrightarrow{h_i^x}; \overleftarrow{h_i^x}]$, and ";" denotes the vector concatenation. The mean-pooling operation is conducted on the representations of all topics to represent $\mathbf{x}$ as $x_{\mathrm{enc}} = \mathrm{mean}(h_1^x, ..., h_m^x)$. Similarly, another bidirectional GRU is adopted as the essay encoder. The representation of the essay $\mathbf{y}$ could be obtained in the same way as the topic encoder does, which is denoted as y_{enc}.

Internal Knowledge Enhancement. Although the auto-encoder framework has shown competitive performance in many text generation tasks, the limited source information of the TEG task cannot provide sufficient information for the decoder to reconstruct the targeted output essay. This paper notices that informative latent features produced by the encoder are essential for a better decoding procedure. Inspired by the CVAE framework taking both the source and the target to train a recognition network, a teacher network is proposed by taking both the topics and essay as source information to get informative latent features for the essay decoder. Since only topics could be accessed during inference, a student network taking topic words solely as input is designed to learn from the teacher network's latent features as internal knowledge. Different from CVAE that trains both the recognition network and the prior network to be close to each other in the latent space, the teacher network in our model maintains an independent training procedure following minimizing the reconstruction error. Because the teacher network is expected to provide strong supervision without being influenced by the student network. The student network would generate latent features which learn the information from the teacher network's latent features through knowledge transfer. That is, the student network is pushed to be close to the teacher network in the latent space.

The teacher network consists of two feed-forward networks, and each network takes x_{enc} and y_{enc} as input to produce the mean and the diagonal covariance by two matrix multiplications, respectively. The latent features z_1 and z_2 are sampled from two Gaussian distributions defined by the above two feed-forward networks, respectively. During training, z_1 is used as a part of the essay decoder's initial hidden state, and z_2 is used as a part of the essay decoder's input at each step to provide more source information. The decoder receives z_1 and z_2 to optimize the training objective. Similarly, there are two feed-forward networks in the student network, where each network takes x_{enc} solely as input to sample a latent feature. Then, the student network's latent features $\tilde{z}_1$ and $\tilde{z}_2$ could be obtained. During inference, the decoder decodes $\tilde{z}_1$ and $\tilde{z}_2$ into a essay. Hence, above latent features could be obtained as follows:

$$\begin{array}{l} z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2 \mathbf{I}) \\ z_2 \sim \mathcal{N}(\mu_2, \sigma_2^2 \mathbf{I}) \end{array}, \left(\begin{bmatrix} \mu_1 \\ \log\left(\sigma_1^2\right) \end{bmatrix}, \begin{bmatrix} \mu_2 \\ \log\left(\sigma_2^2\right) \end{bmatrix} \right) = \text{Teacher}(x_{\text{enc}}, y_{\text{enc}}), \quad (3)$$

$$\begin{array}{l} \tilde{z}_1 \sim \mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1^2 \mathbf{I}) \\ \tilde{z}_2 \sim \mathcal{N}(\tilde{\mu}_2, \tilde{\sigma}_2^2 \mathbf{I}) \end{array}, \left(\begin{bmatrix} \tilde{\mu}_1 \\ \log\left(\tilde{\sigma}_1^2\right) \end{bmatrix}, \begin{bmatrix} \tilde{\mu}_2 \\ \log\left(\tilde{\sigma}_2^2\right) \end{bmatrix} \right) = \text{Student}(x_{\text{enc}}), \quad (4)$$

where $\mathbf{I}$ is an identity matrix, and the reparametrization trick is used to sample the latent features. For enhancing the generation performance, the teacher network is trained to reconstruct the target, while the internal knowledge from the teacher network is transferred to the student network by minimizing the Kullback-Leibler (KL) divergence between the teacher's distributions and the student's distributions in the latent space as follows:

$$\mathcal{L}_{\text{trans}} = \text{KL}(\mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1^2 \mathbf{I}) || \mathcal{N}(\mu_1, \sigma_1^2 \mathbf{I})) + \text{KL}(\mathcal{N}(\tilde{\mu}_2, \tilde{\sigma}_2^2 \mathbf{I}) || \mathcal{N}(\mu_2, \sigma_2^2 \mathbf{I})). \quad (5)$$

Topic Knowledge Graph Encoder. Incorporating external commonsense knowledge is important to bridge the information gap between the source and the target. Unlike previous works only considering the nearest neighbor nodes of topics, this paper constructs a topic knowledge graph queried by the topic words over a few hops from ConceptNet to assist the generation procedure. Then, a topic knowledge graph $\mathbf{G} = (\mathbf{V}, \mathbf{R}, \mathbf{E})$ could be obtained, where $\mathbf{V}$ denotes the set of vertices, $\mathbf{R}$ is the set of relations, and $\mathbf{E}$ represents the set of edges. The topic knowledge graph encoder is designed to integrate the topic-related information from $\mathbf{G}$. By considering the topic knowledge graph, the objective of the TEG task could be modified as follows:

$$\hat{\theta} = \arg \max_\theta P_\theta(\mathbf{y}|\mathbf{x}, \mathbf{G}). \tag{6}$$

External Knowledge Enhancement. Appropriate usage of the structural and semantic information in the external knowledge graph plays a vital role in the TEG task. Each edge (u, r, v) in $\mathbf{G}$ means that the relation $r \in \mathbf{R}$ exists from a node u to a node v. This paper extends (u, r, v) with its reversed link (v, r_{rev}, u) to allow the information in a directed edge to flow along both directions [13], where r_{rev} denotes the reversed relation. For instance, given the edge $(worry, isa, emotion)$, the reversed edge $(emotion, isa\_r, worry)$ is added in $\mathbf{G}$. Our topic knowledge graph encoder is based on the graph neural network (GNN) framework, which could aggregate the graph-structured information of a node from its neighbors. Specifically, a graph convolution network (GCN) with L layers is employed. For jointly embedding both nodes and relations in the topic knowledge graph, this paper follows Vashishth et al. [19] to perform a non-parametric compositional operation ϕ for combining the neighbor node and the relation of a central node. As shown in Fig. 3, for a node $v \in \mathbf{V}$, its embedding would be updated at the $l+1$-th layer by aggregating information from its neighbors $N(v)$. The topic knowledge graph encoder treats incoming edges and outgoing edges differently to sufficiently encode structural information. Specifically, the related edges of the node v could be divided into the set of incoming edges and that of outgoing edges, denoted as $\mathbf{E}_{\text{in}(v)}$ and $\mathbf{E}_{\text{out}(v)}$, respectively. Then, the node embedding of v could be updated as follows:

$$o_v^l = \frac{1}{|N(v)|} \sum_{(u,r) \in N(v)} W_{\text{dir}(r)}^l \phi(h_u^l, h_r^l), \tag{7}$$

$$h_v^{l+1} = \text{ReLU}(o_v^l + W_{\text{loop}}^l h_v^l), \tag{8}$$

where h_v^0 is initialized by the original word embedding, and h_r^0 is initialized by the relation embedding. The weight matrix $W_{\text{dir}(r)}^l$ is a relation-direction specific parameter at the l-th layer as follows:

$$W_{\text{dir}(r)}^l = \begin{cases} W_{\text{in}}^l, & (u, r, v) \in \mathbf{E}_{\text{in}(v)} \\ W_{\text{out}}^l, & (v, r, u) \in \mathbf{E}_{\text{out}(v)} \end{cases}. \tag{9}$$

The compositional operation employs $\phi(h_u^l, h_r^l) = h_u^l + h_r^l$ when incoming edges are considered, and $\phi(h_u^l, h_r^l) = h_u^l - h_r^l$ when outgoing edges are considered [19].
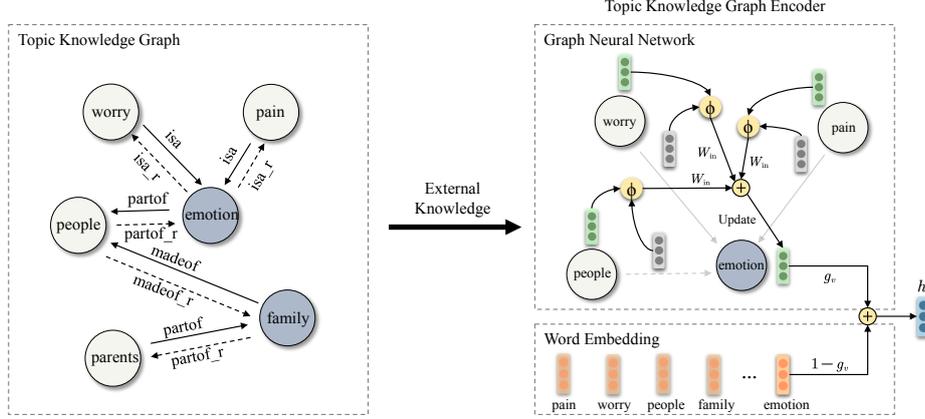
Fig. 3. Topic knowledge graph encoder. The graph neural network performs a compositional operation for a central node (e.g., emotion). Only incoming edges are considered in the diagram for clarity. The information from the topic knowledge graph is aggregated to update the embedding of the central node. Then, the final updated embedding is combined with the original word embedding to assist the essay decoding.

o_v^l is the aggregated information from the l-th layer, and the weight matrix W_{loop}^l is used to transform v's own information from the l-th layer. For the relation r, its embedding is updated as follows:

$$h_r^{l+1} = W_r^l h_r^l, \tag{10}$$

where W_r^l is a weight matrix. A gate mechanism is designed to combine h_v^L containing graph knowledge and h_v^0 containing original semantic knowledge by:

$$g_v = \text{Sigmoid}(W_{\text{gate}}[h_v^L; h_v^0]), \tag{11}$$

$$h_v = g_v \odot h_v^L + (1 - g_v) \odot h_v^0, \tag{12}$$

where W_{gate} is a weight matrix, and $\odot$ is the element-wise multiplication.

Finally, the node embedding h_v is obtained to encode both structural and semantic information of the knowledge graph as external knowledge, involved in each decoding step for better essay generation.

Essay Decoder. The essay decoder employs a single layer GRU. The initial hidden state s_0 is set with $s_0 = [x_{\text{enc}}; z_1]$ containing the topics' representation and the latent feature. Both internal and external knowledge should be involved in each decoding step. Specifically, the hidden state s_t of the decoder at time step t is obtained as follows:

$$s_t = \text{GRU}(s_{t-1}, [e(y_{t-1}); z_2; c_t^x; c_t^g]), \tag{13}$$

where $e(y_{t-1})$ is the embedding of the essay word y_{t-1} at the time step $t-1$, c_t^x is the topic context vector at the time step t, which integrates the output

representations from the topic encoder by the attention mechanism as follows:

$$e_{t,i}^x = (\tanh(W_x s_{t-1} + b_x))^T h_i^x, \ \alpha_{t,i}^x = \frac{\exp(e_{t,i}^x)}{\sum_{j=1}^m \exp(e_{t,j}^x)}, \ c_t^x = \sum_{i=1}^m \alpha_{t,i}^x h_i^x, \quad (14)$$

and c_t^g is the graph context vector, which integrates the representations of the graph nodes from the topic knowledge graph encoder as follows:

$$e_{t,v}^g = (\tanh(W_g s_{t-1} + b_g))^T h_v, \ \alpha_{t,v}^g = \frac{\exp(e_{t,v}^g)}{\sum_{u \in \mathbf{V}} \exp(e_{t,u}^g)}, \ c_t^g = \sum_{v \in \mathbf{V}} \alpha_{t,v}^g h_v. \quad (15)$$

The internal knowledge from the latent feature z_2, and the external knowledge from the graph context vector c_t^g would help the decoder select a proper word. Note that z_1 and z_z would be replaced with $\tilde{z}_1$ and $\tilde{z}_2$ during inference. Since our model takes both $\mathbf{x}$ and $\mathbf{y}$ as input when using the teacher network, the probability of obtaining an essay word for training is obtained by:

$$P_\theta(y_t | y_{<t}, \mathbf{x}, \mathbf{y}, \mathbf{G}) = \mathrm{Softmax}(W_o s_t + b_o). \quad (16)$$

Discriminator. A ψ-parameterized CNN-based discriminator [8] D_ψ is introduced in our model for adversarial training which would improve essay quality.

Adversarial Training. Due to the discrete output space problem of text generation, previous works heavily rely on the RL-based adversarial training which has less-informative reward signals and high-variance gradients. In contrast, this paper proposes the adversarial training through the Wasserstein distance for TEG. Based on the Wasserstein distance, the discriminator could operate on continuous valued output and provide strong learning signals by distinguishing between a real text sequence of one-hot vectors and a generated text sequence of probabilities. Specifically, the hidden state s_t of the essay decoder is employed to generate a probability output $y_t^\theta = \mathrm{Softmax}(W_o s_t + b_o)$. Then, a sequence of outputs $\mathbf{y}_\theta = (y_1^\theta, ..., y_n^\theta)$ could be regarded as the generated essay for adversarial training. For aligning the generated essay with the related topics, the pair of the topics $\mathbf{x}$ and the ground truth essay $\mathbf{y}$ is fed to D_ψ as the real sample, while the pair of $\mathbf{x}$ and $\mathbf{y}_\theta$ is treated as the generated sample. Then, the adversarial training objective based on the Wasserstein distance for D_ψ is formulated by:

$$\mathcal{L}_{D_\psi} = D_\psi(\mathbf{x}, \mathbf{y}_\theta) - D_\psi(\mathbf{x}, \mathbf{y}) + \lambda(||\nabla_{\hat{\mathbf{y}}} D_\psi(\mathbf{x}, \hat{\mathbf{y}})||_2 - 1)^2, \quad (17)$$

where the gradient penalty $(||\nabla_{\hat{\mathbf{y}}} D_\psi(\mathbf{x}, \hat{\mathbf{y}})||_2 - 1)^2$ weighted by λ is imposed on the discriminator to enforce the Lipschitz constraint, and $\hat{\mathbf{y}} = \alpha \mathbf{y} + (1 - \alpha)\mathbf{y}_\theta$ with $\alpha \sim \mathrm{Uniform}(0, 1)$. The auto-encoder framework in our model could act as a generator to minimize the following adversarial training objective as:

$$\mathcal{L}_{\mathrm{adv}} = -D_\psi(\mathbf{x}, \mathbf{y}_\theta) - \beta \log[P_\theta(\mathbf{y} | \mathbf{x}, \mathbf{y}, \mathbf{G})], \quad (18)$$

where the log-likelihood term $\log[P_\theta(\mathbf{y} | \mathbf{x}, \mathbf{y}, \mathbf{G})]$ weighted by β would help align the generated essay with the topics further and keep generation diversity. The generator and the discriminator D_ψ are alternately trained to play a minimax game, where D_ψ assists the generator to obtain high-quality essays.

3.3   Training and Inference

For the training procedure, the latent features for decoding an essay are computed by the teacher network. Two training stages are employed in TEGKE. At the first training stage, the negative log-likelihood is minimized to reconstruct the ground truth essay $\mathbf{y} = (y_1, ..., y_n)$ as follows:

$$\mathcal{L}_{\mathrm{rec}} = \sum_{t=1}^{n} -\log[P_\theta(y_t|y_{<t}, \mathbf{x}, \mathbf{y}, \mathbf{G})], \tag{19}$$

where all parameters except the student network's parameters are optimized in an end-to-end manner. For transferring internal knowledge from the teacher network to the student network, the KL divergence between the student's distributions and the teacher's distributions is minimized by $\mathcal{L}_{\mathrm{trans}}$ of Eq. (5) to optimize the student network's parameters.

At the second training stage, the auto-encoder framework in our model acts as a generator which is trained by $\mathcal{L}_{\mathrm{adv}}$ of Eq. (18). The discriminator is trained by $\mathcal{L}_{D_\psi}$ of Eq. (17) to provide a learning signal for the generator. Note that the student network is still optimized by $\mathcal{L}_{\mathrm{trans}}$ during the second stage. For the inference procedure, the latent features for decoding are computed by the student network. The input to our model is the topics $\mathbf{x}$ and the topic knowledge graph $\mathbf{G}$, and then the decoder would generate a related essay. The pseudo code of TEGKE is shown in the supplementary material: `https://arxiv.org/abs/2106.15142`.

4   Experiments

4.1   Datasets

Experiments are conducted on the ZHIHU corpus [4] consisting of real-world Chinese topic and essay pairs. The number of topic words is between 1 and 5. The length of an essay is between 50 and 100. For extracting external knowledge sufficiently, this paper constructs the topic knowledge graph from ConceptNet over 5 hops, and then 40 nodes are reserved per hop [7]. The constructed topic knowledge graph is a subgraph of ConceptNet. For this knowledge graph, the maximum number of nodes is 205, and the maximum number of edges is 912. The training set and the test set contain 27,000 samples and 2,500 samples, respectively. We set 10% of training samples as the validation set for hyper-parameters tuning. Besides, the experimental results on the ESSAY corpus [4] are shown in the supplementary material.

4.2   Settings

The essay decoder is a GRU with a hidden size of 1024. Both the topic encoder and the essay encoder are implemented as a bidirectional GRU with a hidden size of 512. The size of latent features is 512 in the teacher network and the student network. For the discriminator, the weight λ of the gradient penalty is set to 10.

The weight β is set to 10. The vocabulary size is 50,000, and the batch size is set to 32. The 200-dim pretrained word embeddings [16] are shared by topics, essays, and initial graph nodes. The 200-dim randomly initialized vectors are used as initial graph relation embeddings. Adam optimizer [9] is used to train the model with the learning rate 10^{-3} for the first training stage, and the learning rate 10^{-4} for the second training stage.

4.3 Baselines

TAV [4] encodes topic semantics as the average of the topic's embeddings and then uses an LSTM as a decoder to generate each word.
TAT [4] enhances the decoder of TAV with the attention mechanism to select the relevant topics at each step.
MTA [4] extends the attention mechanism of TAT with a topic coverage vector to guarantee that every single topic is expressed by the decoder.
CTEG [20] introduces commonsense knowledge into the generation procedure and employs adversarial training to improve generation performance.
SCTKG [15] extends CTEG with the topic graph attention and injects the sentiment labels to control the sentiment of the generated essay. The SCTKG model without sentiment information is considered as a baseline, since the original TEG task does not take the sentiment of ground truth essays as input.

4.4 Evaluation Metrics

In this paper, both automatic evaluation and human evaluation are adopted to evaluate the generated essays.

Automatic Evaluation. Following previous works [4, 15, 20], there are several automatic metrics considered to evaluate the model performance.
BLEU [14]: The BLUE score is widely used in text generation tasks (e.g., dialogue generation and machine translation). It could measure the generated essays' quality by computing the overlapping rate between the generated essays and the ground truth essays.
Dist-1, Dist-2 [11]: The Dist-1 and Dist-2 scores are the proportion of distinct unigrams and bigrams in the generated essays, respectively, which measure the diversity of the generated essays.
Novelty [20]: The novelty is calculated by the difference between the generated essay and the ground truth essays with similar topics in the training set. A higher score means more novel essays would be generated under similar topics.

Human Evaluation. Following previous works [15, 20], in order to evaluate the generated essays more comprehensively, 200 samples are collected from different models for human evaluation. Each sample contains the input topics and the generated essay. All 3 annotators are required to score the generated essays from 1 to 5 in terms of four criteria: **Topic-Consistency** (**T-Con.**), **Novelty**

Table 1. Automatic and human evaluation results. ↑ means higher is better. * indicates statistically significant improvements ($p < 0.001$) over the best baseline.

Method	Automatic Evaluation				Human Evaluation			
	BLEU(↑)	Novelty(↑)	Dist-1(↑)	Dist-2(↑)	T-Con.(↑)	Nov.(↑)	E-div.(↑)	Flu.(↑)
TAV	6.05	70.32	2.69	14.25	2.32	2.19	2.58	2.76
TAT	6.32	68.77	2.25	12.17	1.76	2.07	2.32	2.93
MTA	7.09	70.68	2.24	11.70	3.14	2.87	2.17	3.25
CTEG	9.72	75.71	5.19	20.49	3.74	3.34	3.08	3.59
SCTKG	9.97	78.32	**5.73**	23.16	3.89	3.35	3.90	3.71
TEGKE	**10.75***	**80.18***	5.58	**28.11***	**4.12***	**3.57***	**4.08***	**3.82***

(**Nov.**), **Essay-Diversity** (**E-div.**), and **Fluency** (**Flu.**). For novelty, the TF-IDF features of topic words are used to retrieve the 10 most similar training samples to provide references for the annotators. Finally, each model's score on a criterion is calculated by averaging the scores of three annotators.

4.5   Experimental Results

Automatic Evaluation Results. The automatic evaluation results over generated essays are shown in the left block of Table 1. Compared with TAV, TAT, and MTA, TEGKE consistently achieves better results on all metrics. This illustrates that, without introducing sufficient knowledge, their models obtain unsatisfactory performance due to the limited source information. CTEG and SCTKG consider the nearest neighbor nodes of topics from ConceptNet as external information. In contrast, the multi-hop topic knowledge graph provides more structural and semantic information which is extracted by our topic knowledge graph encoder. Hence, our model outperforms the best baseline by 0.78 on the BLEU score, demonstrating that the potential of our model to generate high-quality essays. Moreover, TEGKE could obtain competitive results on the Dist-1 scores, while greatly improving the Dist-2 and novelty scores by 4.95 and 1.86 over SCTKG, respectively. That is, the essays generated from our model would be more diverse and different from the essays in the training corpus. In general, by integrating various internal and external knowledge into generation, TEGKE could achieve better quality and diversity simultaneously.

Human Evaluation Results. The human evaluation results are shown in the right block of Table 1, and TEGKE could obtain the best performance on all metrics. The external knowledge incorporated by the topic knowledge graph encoder would help the decoder select topic-related words, and the adversarial training could further align generated essays with related topics. Thus, our model outperforms the best baseline by 0.23 on the topic-consistency score, showing that the generated essays are more closely related to the given topics. The improvement over the novelty, essay-diversity, and fluency scores demonstrates that TEGKE could obtain better samples in terms of quality and diversity. This conclusion is similar to that drawn from the automatic evaluation.

Table 2. Ablation study results.

Method	BLEU(↑)	Novelty(↑)	Dist-1(↑)	Dist-2(↑)
TEGKE	**10.75**	**80.18**	5.58	28.11
TEGKE w/o EX	10.18	78.67	5.38	21.16
TEGKE w/o AD	10.63	80.09	**5.65**	**28.33**
TEGKE w/o EX & AD	9.78	79.42	5.46	21.30

Ablation Study. To illustrate the effectiveness of our model's key parts, this paper performs an ablation study on three ablated variants: TEGKE without external knowledge enhancement (TEGKE w/o EX), TEGKE without adversarial training (TEGKE w/o AD), and TEGKE with only internal knowledge enhancement (TEGKE w/o EX & AD). The results are shown in Table 2.
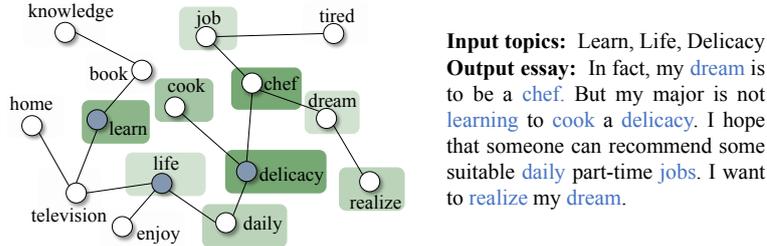
Internal Knowledge Enhancement. Based on only the internal knowledge from the teacher network, TEGKE w/o EX & AD achieves the worst results among variants. However, its performance is still comparable to CTEG adopting both adversarial training and commonsense knowledge, showing that the latent features produced by the TEGKE w/o EX & AD's topic encoder benefit essay generation. Specially, TEGKE w/o EX & AD increases Dist-1 by 0.27 and Dist-2 by 0.81 over CTEG. This improvement comes from the teacher and student networks' various outputs, because our decoder generates essays depending on two latent features sampled from different Gaussian distributions. The above results illustrate that utilizing a student to learn from a teacher makes our model learn the relation between topics and essays better, which enhances the model performance.

External Knowledge Enhancement. Compared with TEGKE, TEGKE w/o EX shows much inferior performance on all metrics. Specifically, TEGKE w/o EX drops 0.57 on the BLEU score, since the external knowledge would help the model select a topic-related word by exploring the topic words and their neighbors in the multi-hop topic knowledge graph. Besides, the diversity of generated essays from TEGKE w/o EX degrades, which is shown by the decline on the novelty, Dist-1, and Dist-2 scores. Specially, TEGKE w/o EX greatly drops 6.95 on Dist-2, due to lacking the commonsense knowledge to provide background information and enrich the input source. By utilizing external knowledge, TEGKE w/o AD still outperforms SCTKG on most metrics. That is, our graph encoder could extract more informative knowledge from the multi-hop knowledge graph.

Adversarial Training. Based on the adversarial training, TEGKE w/o EX boosts the BLEU score by 0.4 over TEGKE w/o EX & AD, and only slightly sacrifices the novelty, Dist-1, and Dist-2 scores due to the inherent mode collapse problem in adversarial training. It demonstrates that the proposed adversarial training could effectively improve the essay quality. Compared with TEGKE w/o AD, TEGKE increases the BLEU score by 0.12, illustrating that our adversarial training could cooperate with the external knowledge enhancement. Since the external knowledge greatly enriches the source information and boosts the model
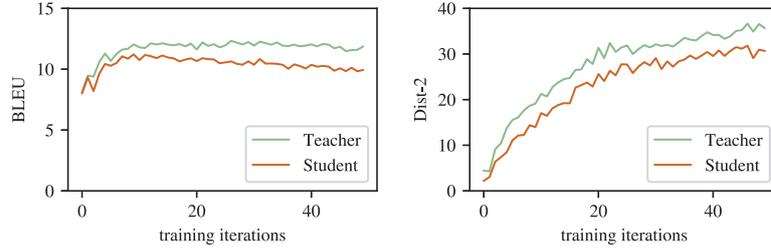
Fig. 4. Training curves. The BLEU score and the Dist-2 score are employed to measure quality and diversity, respectively. For both BLEU and Dist-2, the higher the better.

Input topics: Learn, Life, Delicacy
Output essay: In fact, my dream is to be a chef. But my major is not learning to cook a delicacy. I hope that someone can recommend some suitable daily part-time jobs. I want to realize my dream.

Fig. 5. Case study. The attention scores over the topic knowledge graph are shown on the left side. Deeper green indicates higher attention scores. The input topics and the generated essay are shown on the right side, where the selected words with higher attention scores are highlighted in blue. The original Chinese is translated into English.

performance, the improvement brought by the adversarial training is somewhat weakened when the topic knowledge graph is introduced.

4.6   Validity of Knowledge Transfer

To illustrate the validity of transferring knowledge from the teacher network to the student network, the performance of our model using the teacher network and that using the student network is shown in Fig. 4. The quality is measured by BLEU, and the diversity is measured by Dist-2. When our model uses the teacher network, the teacher network's latent features are fed to the decoder for generating essays. The model could maintain a stable training procedure and obtain excellent results since the ground truth essays are taken as input. The student network would learn from the teacher network's latent features. For the performance, the model using the student network closely follows that using the teacher network. Although the model using the student network performs slightly worse, the results on quality and diversity are still satisfactory.

4.7   Case Study

A case generated by our model is shown on the right side of Fig. 5. Under the given topics "learn", "life", and "delicacy", TEGKE obtains a high-quality essay

that mainly covers the semantics of input topics. The reason is that our model could integrate internal knowledge and abundant external knowledge into the generation procedure. By greatly enriching the source information, our model would generate novel and coherent essays.

To further illustrate the validity of our topic knowledge graph encoder, this paper visualizes the attention weights of Eq. (15) during the generation procedure on the left side of Fig. 5. Compared with the previous works only considering the 1-hop neighbors of topics, our model could use the information from the multi-hop topic knowledge graph. For instance, in the generated essay, "dream" is a 2-hop neighbor of the topic "delicacy", and "realize" is a 3-hop neighbor of the topic "delicacy". It is observed that all nodes from the path ("delicacy", "chef", "dream", and "realize") get higher attention scores during the generation procedure, indicating that the structural information of the graph is helpful. The generated essay is consistent with the topics' semantics since the topics "learn" and "delicacy" both obtain higher attention scores. Although the topic "life" does not appear in the generated essay, its 1-hop neighbor "daily" injects the corresponding semantic information about "life" into the generated essay.

5   Conclusion

This paper proposes a topic-to-essay generation model with comprehensive knowledge enhancement, named TEGKE. For internal knowledge enhancement, the teacher network is built by taking both topics and related essays as input to obtain informative features. The internal knowledge in these features is transferred to the student network for better essay generation. For external knowledge enhancement, the topic knowledge graph encoder is proposed to extract both the structural and semantic information from commonsense knowledge, which significantly enriches the source information. Moreover, the adversarial training based on the Wasserstein distance is introduced to improve generation quality further. Experimental results on real-world corpora demonstrate that TEGKE outperforms the state-of-the-art methods.

Acknowledgements

References

1. Caccia, M., Caccia, L., Fedus, W., Larochelle, H., Pineau, J., Charlin, L.: Language GANs falling short. In: ICLR (2020)
2. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP. pp. 1724–1734 (2014)

3. Fan, A., Lewis, M., Dauphin, Y.: Hierarchical neural story generation. In: ACL. pp. 889–898 (2018)
4. Feng, X., Liu, M., Liu, J., Qin, B., Sun, Y., Liu, T.: Topic-to-essay generation with neural networks. In: IJCAI. pp. 4078–4084 (2018)
5. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of Wasserstein GANs. In: NeurIPS. pp. 5769–5779 (2017)
6. Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Y., Wang, J.: Long text generation via adversarial training with leaked information. In: AAAI. pp. 5141–5148 (2018)
7. Ji, H., Ke, P., Huang, S., Wei, F., Zhu, X., Huang, M.: Language generation with multi-hop reasoning on commonsense knowledge graph. In: EMNLP. pp. 725–736 (2020)
8. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
10. Leppänen, L., Munezero, M., Granroth-Wilding, M., Toivonen, H.: Data-driven news generation for automated journalism. In: ICNLG. pp. 188–197 (2017)
11. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, W.B.: A diversity-promoting objective function for neural conversation models. In: NAACL. pp. 110–119 (2016)
12. Li, J., Song, Y., Zhang, H., Chen, D., Shi, S., Zhao, D., Yan, R.: Generating classical Chinese poems via conditional variational autoencoder and adversarial training. In: EMNLP. pp. 3890–3900 (2018)
13. Marcheggiani, D., Titov, I.: Encoding sentences with graph convolutional networks for semantic role labeling. In: EMNLP. pp. 1506–1515 (2017)
14. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: ACL. pp. 311–318 (2002)
15. Qiao, L., Yan, J., Meng, F., Yang, Z., Zhou, J.: A sentiment-controllable topic-to-essay generator with topic knowledge graph. In: EMNLP: Findings. pp. 3336–3344 (2020)
16. Song, Y., Shi, S., Li, J., Zhang, H.: Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In: NAACL. pp. 175–180 (2018)
17. Speer, R., Havasi, C.: Representing general relational knowledge in ConceptNet 5. In: LREC. pp. 3679–3686 (2012)
18. Subramanian, S., Rajeswar, S., Dutil, F., Pal, C., Courville, A.: Adversarial generation of natural language. In: the 2nd Workshop on Representation Learning for NLP. pp. 241–251 (2017)
19. Vashishth, S., Sanyal, S., Nitin, V., Talukdar, P.: Composition-based multi-relational graph convolutional networks. In: ICLR (2020)
20. Yang, P., Li, L., Luo, F., Liu, T., Sun, X.: Enhancing topic-to-essay generation with external commonsense knowledge. In: ACL. pp. 2002–2012 (2019)
21. Yang, X., Lin, X., Suo, S., Li, M.: Generating thematic Chinese poetry using conditional variational autoencoders with hybrid decoders. In: IJCAI. pp. 4539–4545 (2018)
22. Yi, X., Sun, M., Li, R., Li, W.: Automatic poetry generation with mutual reinforcement learning. In: EMNLP. pp. 3143–3153 (2018)
23. Yu, L., Zhang, W., Wang, J., Yu, Y.: SeqGAN: Sequence generative adversarial nets with policy gradient. In: AAAI. pp. 2852–2858 (2017)
24. Zhang, J., Feng, Y., Wang, D., Wang, Y., Abel, A., Zhang, S., Zhang, A.: Flexible and creative Chinese poetry generation using neural memory. In: ACL. pp. 1364–1373 (2017)