PC-SAN: Pretraining-Based Contextual Self-Attention Model for Topic Essay Generation

Fuqiang Lin[1], Xingkong Ma[1], Yaofeng Chen[1], Jiajun Zhou[1] and Bo Liu[1*]
[1] College of Computer, National University of Defense Technology
Changsha, 410073, China
[e-mail: linfuqiang13@nudt.edu.cn]
*Corresponding author: Bo Liu
[e-mail: kyle.liu@nudt.edu.cn]

Received January 31, 2020; revised May 5, 2020; accepted July 21, 2020; published August 31, 2020

Abstract

Automatic topic essay generation (TEG) is a controllable text generation task that aims to generate informative, diverse, and topic-consistent essays based on multiple topics. To make the generated essays of high quality, a reasonable method should consider both diversity and topic-consistency. Another essential issue is the intrinsic link of the topics, which contributes to making the essays closely surround the semantics of provided topics. However, it remains challenging for TEG to fill the semantic gap between source topic words and target output, and a more powerful model is needed to capture the semantics of given topics. To this end, we propose a pretraining-based contextual self-attention (PC-SAN) model that is built upon the seq2seq framework. For the encoder of our model, we employ a dynamic weight sum of layers from BERT to fully utilize the semantics of topics, which is of great help to fill the gap and improve the quality of the generated essays. In the decoding phase, we also transform the target-side contextual history information into the query layers to alleviate the lack of context in typical self-attention networks (SANs). Experimental results on large-scale paragraph-level Chinese corpora verify that our model is capable of generating diverse, topic-consistent text and essentially makes improvements as compare to strong baselines. Furthermore, extensive analysis validates the effectiveness of contextual embeddings from BERT and contextual history information in SANs.

Keywords: Natural language generation; Essay generation; Pretraining-based method; Self-attention network; Deep learning

This research was funded by the National Natural Science Foundation of China under grant No. 61572513.

http://doi.org/10.3837/tiis.2020.08.001 **ISSN : 1976-7277**

1. Introduction

Automatic topic essay generation (TEG) is a challenging task that aims at generating smooth, logical, and topic-related paragraph-level text conditioned on multiple topics. **Fig. 1** shows a simple illustration of TEG, which takes a set of topic words as input and outputs an essay closely surrounded by the theme of topics. TEG is playing an increasingly important role in many practice applications to improves the efficiency of manual writing and reduces the workload of human writers, such as topic-based news writing and story generation. Meanwhile, it involves the abilities of natural language understanding and inference, to some extent, represents the development of natural language processing.

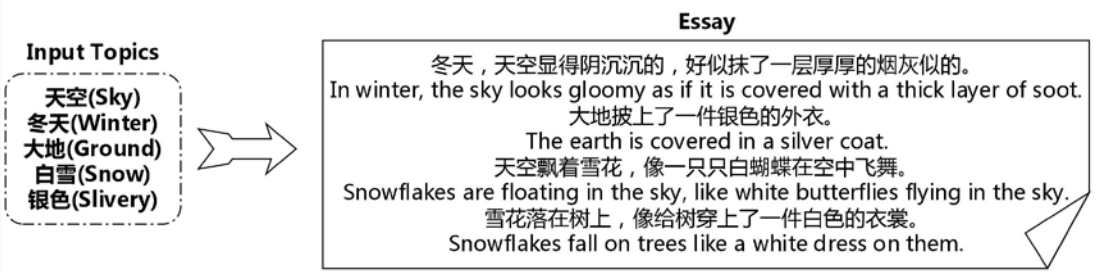


Fig. 1. Chinese example of TEG task.

Despite the broad applications of TEG, there are still several challenges to be addressed. Firstly, the semantic information gap between source topic words and target sequences is unavoidable. As shown in **Fig. 1**, the semantics information in the topic words is much less than that contained in the target sequence for the TEG task, making the generated essays uninformative and monotonous. Thus, the performance of TEG is still unsatisfactory compared with other NLP tasks. Secondly, to write an essay, the model needs to consider not only the integrity but also the relevance of given topics. However, there still lacks a mature solution to make the generator aware of the multi-topic information and control the semantics of the generated text under the theme of topics.

Current methods of topic-to-sequence learning tasks can be roughly divided into two categories: template-based approaches and neural networks. Early works mainly adopt corpora-based methods [1,2] that learn templates and rules to construct essays, which have been widely used in news writing and match reports. Such methods enjoy good interpretability and controllability, but also have several shortcomings. For example, the quality of template extraction depends on artificial feature engineering or rule intervention, and the generated essays are often unsatisfactory in terms of diversity and fluency. When involving in neural networks, researchers have attempted to study TEG using the seq2seq framework and attention networks. The core idea of the TA-Seq2Seq model [3] is to get the topic words of essays through the LDA model, then apply the topic information to the encoder-decoder structure for the guidance of generation. To lessen the duplication problem in the generated essays, Feng et al. [4] proposed the multi-topic-aware LSTM (MTA-LSTM) model, which captures the topic distribution and relatedness by incorporating attention and coverage mechanism. However, the generated essays are still unsatisfactory, with the lack of diversity and poor topic-consistency.

The approaches mentioned above provide efficient ways to generate essays. However, the challenges of TEG have not been well resolved. On the one hand, the semantics of topic words are not explicitly enriched, and the semantic gap greatly reduces the performance of TEG. On the other hand, the attention networks adopted by previous approaches still have defects in the TEG task. Firstly, due to the separate processing of topic words, models can hardly capture the overall semantics of multiple topics, making it difficult to generate essays with a unified theme. Secondly, the attention mechanism tends to ignore past historical information, which often leads to the repetition of some topic words while the neglect of others in the generated text.

To address the issues mentioned above and improve the performance of TEG, we propose a pretraining-based contextual self-attention (PC-SAN) model that builds upon the seq2seq framework. With the purpose of enriching the semantics of topic words and capturing the relationships between them, we adopt the pre-trained contextual language models BERT [5] to obtain the topics' contextual embeddings. Due to the pretraining stage on large-scale unlabeled corpora, BERT has shown strong abilities in capturing the dependence of tokens and can provide linguistic knowledge for TEG. As applied to our model, such external knowledge provides a more comprehensive understanding of multiple topics, which has a positive effect on improving the quality of the generated text, especially in topic-integrity and diversity. Moreover, we employ a dynamic weighted sum of BERT layers output for each layer of the generator, with locked BERT parameters. Compared with canonical ways of appending a thin layer after the BERT structure, our method can take advantage of the linguistic contained in BERT more fine-grained. In the decoding stage, since self-attention networks (SANs) have shown their effectiveness in capturing semantic dependencies and the flexibility in modeling sequential data, we employ a multi-head attention decoder rather than LSTM as the generator. Further, we incorporate the target-side context to the query layers in SANs to overcome the shortcomings of attention networks in the TEG task. The contextual information is dynamically updated at each time step, making the decoder aware of the history information and conditioned on the generated target-side context. Therefore, the target-side contextual history mechanism contributes to alleviating the topic drift problem and enhancing topic-consistency between the input topics and the generated essay.

The contributions of our work are summarized as follows:

1. We incorporate the pre-trained language model BERT into the encoder as contextual embeddings to alleviate the information gap. By fusing the hidden states across encoder layers, the semantics of the topic words are enriched, and the relationship between multiple topics is considered. Moreover, a dynamic weighted sum is employed and leads a better performance in semantic gap alleviation.
2. We design a target-side contextual history mechanism in self-attention networks to guide generation. With the help of the context-aware generator, the quality of the generated text is improved.
3. We conduct experiments on the benchmark datasets ESSAY and ZhiHu corpora. The experimental results show that our proposed method achieves 8.75% and 9.91% relative improvement in terms of BLEU and diversity compared with the best baseline. The quality of the generated essays gets improved, especially on topic-consistent and topic-integrity.

2. Related Work

In related NLG tasks, the most similar work to TEG is poetry generation, since they employ the same inputs and generate text conditioned on given topics. Existing methods can be

roughly classified into two categories: template-based methods and neural networks. Traditional poetry generation approaches mainly adopt template-ruled and pattern-based methods [2,6], which learn templates from poetries and generate new poetry based on lexical and prosodic constraints. When involving neural networks, many models have shown impressive results [7,8]. Wang Q et al. [9] and Wang Z et al. [10] adopted the seq2seq structure [11,12] and attention networks [13], which realized poetry generation with the theme of multiple topics. Zhang et al. [14] further utilized a memory-augmented method to balance the linguistic accordance and aesthetic innovation. To improve the topic-consistency of the generated poem, Yang et al. [15] presented a conditional variational autoencoder with hybrid decoders to learn topic information. Despite their success, poetry generation is the creation of strictly structured texts based on themes and rhythmical constraints. For example, the quatrain consists of four lines and each line has five or seven characters, which follows these specific rules and can be formally described. However, the TEG task aims to generate long unstructured text without specific rule constraints, which poses extra challenges. Therefore, existing methods of poetry generation are difficult to be applied directly to essay generation.

Along a similar direction, the common practices of TEG adopt the seq2seq structure with attention networks. Xing et al. [3] developed a topic-aware approach incorporating topic information into a seq2seq framework to generate informative and interesting responses for chatbots. To improve the thematic integrity and readability of the generated essay, Feng et al. [4] introduced a coverage mechanism [16] by maintaining a coverage vector to utilize topic distributed information. These works provide effective methods to implement the TEG task and achieve some experimental results. Inspired by their works, we adopt the encoder-decoder structure and train our model in an end-to-end way. However, there are still several problems have not received sufficient attention in previous works. Firstly, the information gap between the generated essays and topics is prone to make the generated essays of low quality. To solve the problem, we employ BERT to enrich the semantic information of the input topic words in our encoder. Secondly, the lack of contexts in origin attention networks makes the generator unable to closely surround input topics and involve the overall semantics of input topics. To address similar issues in response generation tasks, Dziri et al. [17] proposed a hierarchical recurrent encoder-decoder (THRED) model, which incorporates topic information and the context into the seq2seq model by a joint attention mechanism. Wu et al. [18] designed a context-aware editing model with an editing vector as the context to generate more grammatical and informative text. The contextual vectors in their works come from source-side sequences, which indeed contributes to improving the performance in response generation. When migrating to the TEG task, since the semantics contained in the source input is extremely limited, we incorporate the contextual information from target-side output rather than the source-side.

3. Approach

3.1 Overview

In this section, we present a pretraining-based contextual self-attention (PC-SAN) model for topic essay generation, which is formulated as follows. Given a set of k topic words $T = \{t_1, t_2, \ldots, t_k\}$, the task aims to generate an essay (or a paragraph) $Y = \{y_1, y_2, \ldots, y_t\}$ under the given topics T. The generated essays require to be informative, diverse and logical, as well as closely related to the semantics of multiple topics. To this end, the core idea of the proposed model is to fill the information gap between source topics and target sequence,

meanwhile, to make the generator aware of overall topic information. **Fig. 2** illustrates the architecture of our model PC-SAN, which consists of a pretraining-based topic encoder and a target-side contextual history attention decoder.

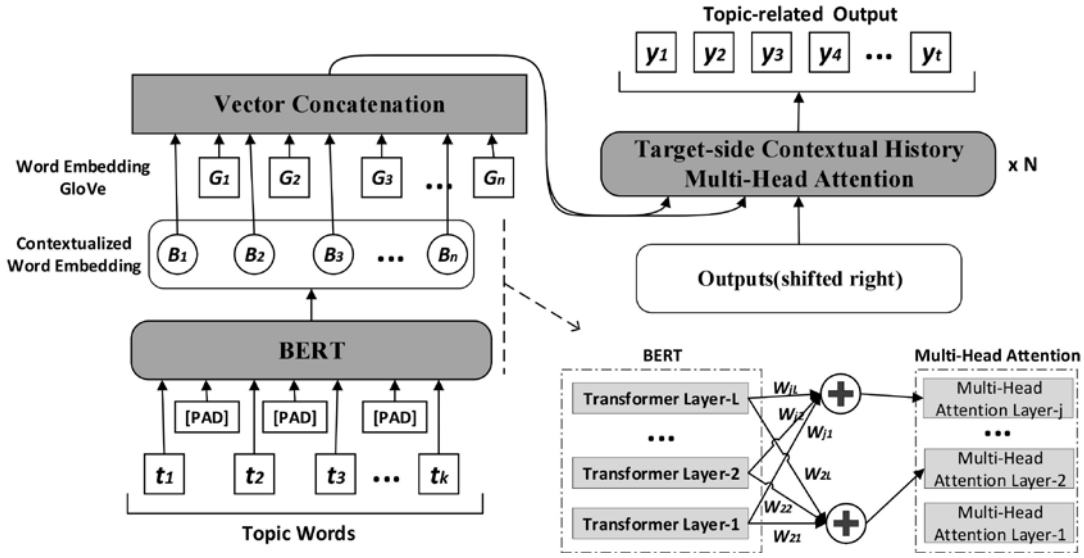


Fig. 2. The architecture of our model PC-SAN

In the PC-SAN model, we adopt the encoder-decoder structure. Topic words are encoded into representation vectors in the encoding stage, containing both word embeddings $\{G_1, G_2, \ldots, G_n\}$ and contextual embeddings $\{B_1, B_2, \ldots, B_n\}$. External linguistics knowledge can enrich the semantics of multiple topics, which indeed helps to alleviate the semantics gap. Therefore, we utilize pre-trained language models to capture the topic information fully. In practice, we employ BERT to obtain contextual embeddings of multiple topics. Due to the bidirectional language structure and the pre-training stage on large-scale unlabeled data, the relationship of topics and the context information is essentially exploited. Further, to take full advantage of the linguistics knowledge, we use a dynamic weighted sum of hidden states across all layers of BERT.

As for the decoding stage, our goal is to generate topic-consistent essays on the basis of the representation vectors provided by the encoder. Since self-attention networks have shown the powerful ability on capturing the dependencies of sequences, we employ the multi-head self-attention networks [19] as the decoder. However, a potential problem of original SANs is the repetition of some topic words while neglecting others in the generated essays, which is caused by the lack of context and the neglect of past alignment information in attention models. To reduce the over-generation and under-generation problems in the output, we apply a novel target-side contextual history mechanism to self-attention networks (SANs) [20] at each time step. The motivation is to synthesize the target-side context of the previous time step as a coverage vector to continuously adjust the topic distribution along with the generating words during decoding.

Compared with previous methods, our model has made corresponding improvements in response to the challenges of the TEG task. Firstly, we make full use of the rich semantics contained in the pre-trained language models to alleviate the information gap. We innovatively adopt a dynamic weighted sum of BERT layers output to generate contextual topic representations for each layer of the decoder, which makes our model generate more

informational and diverse essays. Secondly, the target-side contextual history mechanism makes the generator aware of the context when decoding. It is noteworthy that origin attention models lack context and tend to ignore past alignment information in the TEG task, which indeed degrades the performance of TEG. Our method combines the target-side historical context in SANs to guide generation, which contributes to improving the quality of the generated text in terms of topic-integrity and eliminate the duplication problem in the output.

3.2 Pretraining-Based Topics Encoder

The information gap between given topics and generated essays is innate in the TEG task, which directly causes the performance of TEG lags behind other NLG tasks by a large margin. Toward alleviating the gap, the semantics of source input requires to be enriched. Inspired by the great success of pre-trained language models [21,22], topic words I are encoded into representation vectors $H = \{h_1, h_2, \ldots, h_m\}$ in the encoding stage, including word embeddings and contextual embeddings. In detail, we use GloVe [23] to get the word embeddings of topics. In consideration of BERT's wide usage and great success, we employ BERT as contextual embeddings for the input sequence. Unlike previous works using only the last layer output of BERT or fine-tuning BERT structure with one additional output layer, we use the linear combination of embeddings across all layers of BERT for each layer of the decoder. The intention is that relevant studies have shown that different layers of BERT capture different linguistic information [21], where higher layers capture semantic information while lower layers tend to learn surface and lexical information [24]. Therefore, integrating multi-layer output takes better advantage of the linguistic information contained in the BERT model. Based on the consideration that different layers of the decoder are also concerned about different linguistic information, we maintain varied weight vectors for different layers of the decoder and use the weighted sum of the output from each layer in BERT to obtain contextual embeddings.

Instead of gaining the semantic representations of each topic separately, we splice together the set of topic words I with the special token '[PAD]' and feed them into BERT as a whole, which has a positive effect on improving the topic-correlation and topic-integrity. Then BERT outputs L layer hidden states for the input sequence, represented by the following formula:

$$H_{bert} = \mathrm{BERT}(t_1, t_2, \ldots, t_k) \tag{1}$$

where $H_{bert} = \{h^1_{bert}, h^2_{bert}, \ldots, h^L_{bert}\}$ denotes the hidden states of the BERT model that consists L layers, and $\mathrm{BERT}(\cdot)$ is the BERT model.

We maintain a weight set $\{\alpha_{j1}, \ldots, \alpha_{jL}\}$ for the decoder, where the weights vary from different layers and j indicts the j-th layer. The contextual embeddings $BERT_{jT}$ of the input sequence I for the j-th layer in the decoder is a weighted sum of BERT layer hidden sates.

$$BERT_{jT} = \sum_{l=1}^{L} \alpha_{jl} h^l_{bert} \tag{2}$$

The final representation vectors H_{jT} of the input topics I contain the word embeddings and contextual embeddings, which are calculated as follow:

$$H_{jT} = [Glove_T, BERT_{jT}] \tag{3}$$

$$Glove_T = \mathrm{GloVe}(t_1, t_2, \ldots, t_k) \tag{4}$$

where $Glove_T$ indicates the word embeddings of input topic words and $\mathrm{GloVe}(\cdot)$ is the GloVe model. $[\,,\,]$ denotes the vector concatenation.

It is intuitively plausible that the pretraining-based encoder enriches the semantics of source topics by taking advantage of the linguistics learned by the pre-trained language model, which is of great help to improve the quality of the generated essays. Besides, processing multiple topics as a whole makes the encoder better understand multiple topics, which ensures that the model is aware of the whole meaning of given topics and improves the quality of the generated text in terms of topic-integrity and topic-relevance.

3.3 Target-side Contextual History Attention Decoder

Due to the long length of paragraph-level essays in the TEG task, the decoder should capture both short-term and long-term dependencies. As a variant of attention models, SANs have shown the effectiveness of modeling dependencies and outperform LSTM in parallel computation. Therefore, we employ the multi-head attention model [19] as the decoder, which is composed of stacked self-attention, encoder-decoder attention, and point-wise fully connected layers. SANs calculate the attention weight between each pair of tokens in the sequence, capturing global and local connections more flexibly than RNN, especially the dependencies of words in long sequences. The multi-head attention decoder is a combination of multiple SANs structures. Each head learns features in different representation subspaces, which extends the ability of the model to focus on different locations.

Given an input sequence $H = \{h_1, h_2, \ldots, h_n\} \in \mathbb{R}^{n \times d}$, SANs first transformer it into query layers $Q \in \mathbb{R}^{n \times d}$, key layers $K \in \mathbb{R}^{n \times d}$, and value layers $V \in \mathbb{R}^{n \times d}$:

$$Q, K, V = HW^Q, HW^K, HW^V \tag{5}$$

where $\{W^Q, W^K, W^V\} \in \mathbb{R}^{d \times d}$ are trainable parameters matrices and d indicates the dimensionality of hidden states. The three types of representations transform H into different subspaces. Multi-head attention networks set up several groups $\{W^Q, W^K, W^V\}$, allowing the model to learn relevant information in different representation subspaces. The output layers $O \in \mathbb{R}^{n \times d}$ are calculated as a weighted sum of the value layers V, where the weight assigned to each value is computed by a compatibility function of the query layers Q with the corresponding key layers K:

$$O = \mathrm{ATT}(Q, K, V) \tag{6}$$

where $\mathrm{ATT}(\cdot)$ indicts the attention function [13] that is a mapping that a query and a set of key-value pairs to the output layers. In this work, we use scaled dot-product attention to calculate the similarity between the query and each key, learn the word dependency of sentences, and capture the internal structure of sequences:

$$\begin{aligned} \mathrm{ATT}(Q, K, V) &= \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \\ &= \mathrm{softmax}\left(\frac{HW^Q W^K H}{\sqrt{d}}\right)V \end{aligned} \tag{7}$$

where $\frac{1}{\sqrt{d}}$ is the scaling factor.

Furthermore, recent researches indicate that the calculation of the compatibility between the queries Q and the keys K in the SANs does not make good use of the context [25]. On the one hand, SANs separately calculate the similarity of word pairs without considering contextual semantic information, resulting in origin SANs not perfectly adapting to our tasks, especially during decoding. On the other hand, scaled dot-product attention is used to calculate the similarity between items as Equation (7) in SANs. It makes the generator tend to generate

words or phrases that frequently appear before, which often leads to under-generation and over-generation problems in the generated text.

To the end, we present an efficient way to exploit contextual historical information to the essay generation model. Since topic words entered the TEG task are limited, and the global semantics has been captured at the encoding stage by incorporating contextual embeddings, which will be proven enough to model the source-side context in section 5.2.2. Conversely, because of the long length of generated essays, the problem of lacking context in the decoder is more prominent, which easily leads to the repetition of some topic words and the neglect of others in the generated text. Thus, we focus on applying the contextual information from the target-side outputs to the decoder for the guidance of generation.

Specifically, to contextualize the transformations of the SANs, we first calculate the contextual vectors that summary the representation of the target-side context. In this work, we follow Cho et al. [12] that applies the mean operation to the input layers to obtain the context information. For each SAN layer with the input sequence $H = \{h_1, \ldots, h_i, \ldots, h_n\} \in \mathbb{R}^{n \times d_m}$, when calculating the output at the position i, the self-attention layers in the decoder allow attending to the current position and all positions before it, but not the ones behind it. Therefore, it is necessary to shield the left-forward information flow in the decoder to maintain the autoregressive property of the model. Inspired by the work of Vaswani et al. [19], we calculate the context by masking out (setting to all zero) all values in the input that corresponds to illegal connections. Finally, the context vector is calculated as follows:

$$C = \overline{\mathrm{MASK}(H)}U \tag{8}$$

$$\mathrm{MASK}(H) = MH \tag{9}$$

where $U \in \mathbb{R}^{d_m \times d}$ is trainable parameter matrices and $\mathrm{MASK}(\cdot)$ denotes mask operation. The concrete approach is to employ the upper triangular matrix $M \in \mathbb{R}^{n \times n}$, where the values of the upper triangle are all 1 while the values of the lower triangle and the diagonal are all 0, to each layer. Note that $C \in \mathbb{R}^{n \times d}$ varies from position to position, which takes advantage of the target-side useful context at each hidden state, meanwhile maintain the flexibility on the parallel computation of the SANs.

To make the generator condition on the target-side context at each time step, we propose a simple and effective way to import the context vectors into the SANs, which transforms contextual information into queries Q as follows:

$$\hat{Q} = (1 - \lambda)Q + \lambda C \tag{10}$$

$$\lambda = \sigma(QV^Q + CV^C) \tag{11}$$

where C is the target-side context matrix and $\{V^Q, V^C\} \in \mathbb{R}^{d \times}$ are trainable parameters. The contextual information is propagated into query layers Q through addition and λ is the adjustment parameter, which indicts the weight of the contextual information.

Correspondingly, the calculation formula (6) of self-attention layer is changed based on the target-side context representation:

$$\begin{aligned} O &= \mathrm{ATT}(\hat{Q}, K, V) \\ &= \mathrm{softmax}\left(\frac{\hat{Q}K^T}{\sqrt{d}}\right)V \end{aligned} \tag{12}$$

The proposed method essentially takes advantage of the target-side contextual information for the guidance of generation and makes the SANs aware of the historical attention information. Besides, since the same token is integrated with different context representations at different positions, it can essentially reduce the duplication problem of the generated text in the TEG task.

3.4 Training

In the training phase, we adopt an end-to-end learning approach that is commonly used in the seq2seq structure. The training goal of our model is to maximize the log-likelihood of the ground-truth essay represented as sequences of tokens $\mathbf{y} = (y_1, y_2, \ldots, y_{t'})$ given a set of topic words $T = \{t_1, t_2, \ldots, t_k\}$ as follows:

$$\arg\max_{\Theta} \frac{1}{N} \sum_{n=1}^{N} \log\left(P_{\Theta}\left(\mathbf{y}^{\mathbf{n}} \mid T^{\mathbf{n}}\right)\right) \tag{13}$$

To alleviate the over-fitting problem easily caused by the one-hot label, we follow Pereyra et al. [26] that smooths the label with the setting of $\epsilon_{ls} = 0.1$. Furthermore, we adopt dropout [27] with a rate of $\epsilon_{drop} = 0.1$ as the regularization during training, which potentially improves the performance.

4. Experiment

In this section, we introduce experimental datasets, evaluation metrics, baselines, and parameter settings in detail.

4.1 Dataset Description and Evaluation Metrics

In this paper, we utilize two corpora, named ESSAY and ZhiHu [4], as the experimental datasets, which consist of topic-essay pairs crawled from the Internet. After filtering, each article in the dataset is between 50 and 120 words in length and corresponds to 5 topic words. The description of the dataset is shown in **Table 1**. It contains 305000 paragraph-level topics-essay pairs in the ESSAY dataset, where we randomly select 300000 pairs as the training set and 5000 pairs as the test set. While for ZhiHu, 55000 topics-essay pairs are chosen in total, and sizes of the training set and the test set are 50000 and 5000.

Table 1. Dataset statistics

Dataset	Training	Test	Total
ESSAY	300000	5000	305000
ZhiHu	50000	5000	55000

Like most NLG tasks, we present both automatic evaluation and manual evaluation to evaluate the performance of approaches more scientifically. In this paper, we adopt the BLEU score [28] for the automatic evaluation of TEG. The BLEU metric uses N-gram matching rules to calculate the proportion of similar n-phrases between candidate sequence and reference sequences, which is suitable for the evaluation of our work. Moreover, here we propose the **Diversity** indicator to intuitively evaluate the diversity of the essays, which is measured by the lexical diversity. In simple terms, we calculated the proportion of unique words in the generated text to assess the diversity of the generated text as follows:

$$Diver = \frac{len(uni(\boldsymbol{Y}))}{len(\boldsymbol{Y})} \tag{14}$$

where $\boldsymbol{Y}$ indicts the tokens contained in the generated essay, while $uni(\boldsymbol{Y})$ is unique tokens, which clears redundant tokens in $\boldsymbol{Y}$. And $len(\cdot)$ donates the number of the tokens.

As for manual evaluation, we absorb four indicators described in the work of Feng et al.: **Topic-Integrity**, **Topic-Relevance**, **Coherence**, and **Fluency**. The score of each indicator ranges from 1 to 5. Specifically, we randomly select 200 generated samples for each model. Then three experienced Chinese experts are invited to rate these chosen samples based on these four indicators. Additionally, the value of kappa is calculated in manual evaluation to characterize the labeling consistency of multiple experts.

4.2 Parameter Settings

In the PC-SAN model, we use 768-dim GloVe to obtain the word embeddings of topics and adopt the simplified Chinese version of BERT to obtain contextual embeddings. The decoder is composed of a stack of 6 multi-head attention layers with 8 heads, and the number of hidden units per layer is 768. Dropout is used with the rate set to 0.1. We adapt Adam optimizer with the setting of $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-9}$, where the batch size is set to be 32. At test time, the beam size of the beam search is 5, which improves the naturalness of the generated text.

4.3 Baselines

For the topic-to-essay generation, we compare our model with the following baseline models.

- **PPG** [10]: PPG is a planning-based poetry generating method by using a modified RNN encoder-decoder framework, which aims to generate poems that are coherent and consistent with the intent of users.
- **SC-LSTM** [29]: SC-LSTM is a statistical language generator based on a semantically controlled long short-term memory (LSTM). The core idea of the SC-LSTM system lies in introducing the DA (dialogue act) mechanism into the LSTM network, which adds keywords in the form of one-hot vectors to guide generating.
- **MTA-LSTM** [4]: MTA-LSTM is a multi-topic-aware long short-term memory network, which generates multi-topic related and expression-coherent essays by incorporating attention and coverage mechanism.
- **Transformer** [19]: Transformer relies entirely on attention networks and has shown promising empirical results in various NLP tasks, such as machine translation and dialogue generation [30,31]. Both the encoder and decoder of Transformer are composed of a stack of 6 identical layers. Among them, encoder layers consist of a multi-head self-attention sub-layers and a fully connected feed-forward sub-layer, while decoder layers have an additional encoder-decoder attention sub-layer.

5. Results and Discussion

In this section, we describe experimental results and conduct extensive analysis. To verify the effectiveness of our method, we compare our model with the following existing approaches: PPG, SC-LSTM, MTA-LSTM, and Transformer. Further, two ablation studies are designed to illustrate the effect of key components.

5.1 Performance Comparison

To evaluate our model and baselines more comprehensively and intuitively, **Table 2** presents the human evaluation results. The evaluation scores include four criteria: topic-integrity, topic-relevance, fluency, and coherence. Besides, the score of kappa is 0.69 on four dimensions, which indicts that the scoring result of three experts has a good consistency.

Table 2. Averaged ratings of manual evaluation for different methods

Method	Topic-Integrity	Topic-Relevance	Fluency	Coherence
PPG	2.66	3.06	3.27	2.92
SC-LSTM	3.14	3.44	3.55	3.44
MTA-LSTM	3.31	3.74	3.83	3.64
Transformer	2.95	3.29	3.84	3.31
PC-SAN(our)	3.41	3.96	4.12	3.74

PPG has been proved to be effective to generate Chinese poetry that is coherent and semantical consistent with the intent of users, to some extent, the quality of the generated poetries could be closed to which of humans. However, there could not achieve comparable performance in the TEG task. The main reason is that the approach was designed to generate poetry with the consideration of special features, such as structure, rhythmic and tonal patterns, which are unique to poetry and can be easy to model. When applied to essay generation, the lack of these constraints leads to performance degradation, especially in terms of coherence, topic-integrity, and topic-relevance. In a manner, it also confirmed our view that essay generation is a more challenging task, and the methods for poetry generation cannot be directly applied to the TEG task. By introducing the DA mechanism in the form of one-hot topic vectors as the role of sentence planning, which took multi-topic distribution into account, SC-LSTM did bring notable improvements in all four dimensions. The gap of performance between SC-LSTM and PPG confirmed the attention mechanism is useful in modeling the association between topic words and generated text. Afterward, MTA-LSTM employed a multi-topic coverage vector rather than the one-hot feature representation to indict the weight of each topic and was sequentially updated during decoding. It brought a slight performance improvement because of the inherent relation of topic words considered. Although the Transformer performed worse than SC-LSTM and MTA-LSTM because there is no explicit modeling topic word distribution, the Transformer slightly outperformed other baselines in terms of fluency. The improvement may benefit from the powerful strength of direct capturing dependencies of tokens by the SANs.

As shown in **Table 2**, it is obvious that our method consistently outperformed all baseline methods in all four indicators, especially in terms of topic-integrity, topic-relevance, and coherence. Both our method and strong baseline MTA-LSTM construct the semantics of topics to guide the generation of essays better. However, the performance varied depending on how the semantics was captured and utilized. Firstly, our model introduced BERT to gain contextual embeddings, which provide extensive knowledge for essay generation to fill the semantics gap. To take full advantage of linguistic information, we develop the dynamic weights to synthesize hidden states across all layers of BERT. It is more flexible to fully utilize different levels of linguistic information captured by BERT, allowing the encoder to capture the semantics of multiple topics with a finer granularity. Secondly, due to the powerful ability of SANs to capture language dependencies, we employed the multi-head attention decoder instead of LSTM. Moreover, the target-side contextual history attention, which transformed the contextual information into the query layers, indeed alleviated the lack of contexts in SANs and brought an additional improvement.

Table 3. Automatic evaluations of different methods on the ESSAY and ZhiHu datasets

Method	ESSAY		ZhiHu		Average	
	BLEU	Diver	BLEU	Diver	BLEU	Diver
PPG	2.46	2.59	1.08	2.07	1.77	2.33
SC-LSTM	3.67	3.69	1.32	2.28	2.495	2.985
MTA-LSTM	4.37	5.45	1.91	3.43	3.14	4.44
Transformer	3.09	3.56	1.19	2.16	2.14	2.86
Gold(corpus)	-	12.36	-	12.46	-	12.41
PC-SAN(our)	**4.68**	**5.83**	**2.15**	**3.93**	**3.415**	**4.88**
Improve	7.09%	6.97%	12.56%	14.57%	8.75%	9.91%

To support the objective evaluation, **Table 3** reports the evaluation results in terms of the BLEU score and diversity (Diver) on the ESSAY and ZhiHu datasets. The best performances are shown in bold, and the averages on two datasets are calculated. We can see that the BLEU scores and diversity are highly consistent with the results of subjective evaluation. Compared with the strong baseline MTA-LSTM, our model gained an average increase of 8.75% and 9.91% in terms of BLEU and diversity, respectively. These results verified that our method is capable of generating more diverse, smooth, and topic-related essays. Besides, all methods performed better on the ESSAY dataset than ZhiHu, which might be caused by the different writing styles and various presentations of ZhiHu website users. An interesting point is the diversity of ZhiHu is superior to which of ESSAY in the corpus, but all methods have achieved better results on ESSAY rather than ZhiHu. This may be because the higher diversity of ZhiHu comes from proper nouns, such as names of people, and places, that are less common in the ESSAY corpus. Since these proper nouns rarely used in the whole corpus, less attention is paid to these words at the decoding stage, and they contribute little to improving diversity in the output. Overall, it is easy to find that the BLEU score of our method is still at a general level, compared with other NLG tasks, e.g., machine translation with a score of 26.4 [19], which supported the hypothesis that TEG is a more challenging task in NLP.

5.2 Ablation Study

The experimental results mentioned above show that our proposed model outperforms the baselines in the TEG task. Next, we design an ablation study to understand the effects of key components in our method.

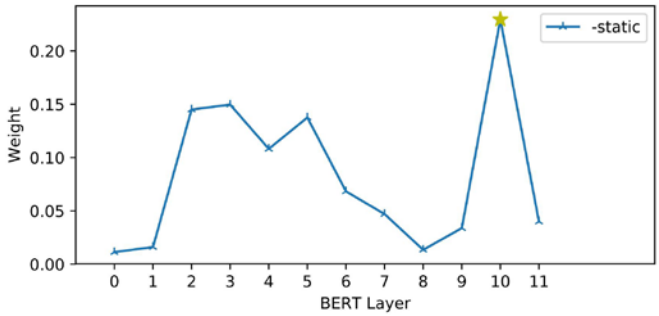
5.2.1 Effects of Dynamic Weight Contextual Embedding

We first design experiments to evaluate the effects of the contextual embeddings of topic words, which are obtained with the dynamic weighted sum of the hidden states from BERT. We propose several variant models to conduct ablation studies on the effectiveness and display the results in **Table 4**. As seen, the variant models include the model without BERT contextual embeddings (**-None**), the model using only the last layer output of BERT (**-Last**), and the model incorporating a static weighted sum of the per-layer output from BERT (**-Static**). It can be observed that all models with contextual embeddings of the topic words from BERT, no matter what form, outperformed the **-None** model, which shows the contextual information captured by BERT is helpful to the TEG task. The **-Static** model is superior to the **-Last** one, which approves that different linguistic properties of the sentence are captured by different layers and aggregate the output of each layer is useful in downstream tasks. Further, our proposed dynamic weight mechanism boosts the performance by 17.15% and 16.05% in

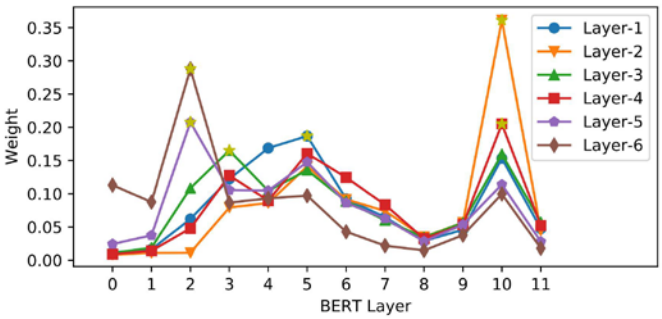
terms of BLEU and diversity, compared with the static weight one. It shows that the semantic information needed for each layer of the decoder varies, and the setting of dynamic weight better fuses semantic information for the topic-to-essay learning.

Table 4. Automatic evaluations of variant methods on the effect of dynamic weight

Method	ESSAY		ZhiHu		Average	
	BLEU	Diver	BLEU	Diver	BLEU	Diver
PC-SAN (Full Model)	4.68	5.83	2.15	3.93	3.415	4.88
-Static	4.02	4.81	1.81	3.60	2.915	4.205
-Last	3.62	4.23	1.71	3.16	2.665	3.695
-None	3.29	3.77	1.57	2.22	2.43	2.995



(a) **-Static**: static weight



(b) **PC-SAN**: dynamic weight

Fig. 3. The visualization of weight distribution. (a) **-Static** model; (b) **PC-SAN** model

To visually distinguish the importance of each layer output of BERT, **Fig. 3** shows the weight distribution. Concerning the -Static model, the varied weights are all greater than zero, which means that different layers of BERT learn different levels of linguistics, and all layers contribute to constructing contextual information of topic words. A minor exception is that the penultimate layer, rather than the last one, gains the highest weight. This reason may be the last layer is too closed to the target functions (masked language model and next sentence prediction) of the pre-training stage. Further, both weight distribution and dominant layers (marked with golden stars in **Fig. 3**) vary for layers of the decoder in the PC-SAN model, which indicts that different layers of the decoder focus on diverse semantics information. The improvements in performance and the varied weights jointly validate our claim that the use of dynamic weights can better capture the richness of context and bring improvement to TEG.

5.2.2 Effects of Contextual History Attention Network

In our model, we incorporate the target-side contextual history by transforming the context of currently generated text into the query layers of the decoder. We present controlled experiments on the effects of contextual history and propose two variants. The first one (**-None-context**) is the native SANs without any additional context information, and the other one (**-Source-context**) is the SANs with global information from source-side context as Equations (12) for the guidance of decoding. The results of this group of experiments are shown in **Table 5**. As seen, our strategy is superior to others with improvements of the BLEU score by about 8.75% and the diversity by 2.95%, which shows the validity of contextual information from the target-side in the essay generation task. Unexpectedly, the source-side context is proven beneficial for machine translation [32-34], however, the performance of the **-Source-context** approach is only comparable to **-None-context** one in our experiment. We attribute the reason to the fact that the semantics of source-side sequences are simple with the limited topic words in the TEG task. The decoder of our model exploits the dynamic weighted sum of encoder layers representations, which has already embedded enough useful contextual information of the source sequences.

Table 5. Automatic evaluations of variant methods on the effect of target-side contextual history

Method	ESSAY		ZhiHu		Average	
	BLEU	Diver	BLEU	Diver	BLEU	Diver
PC-SAN (Full Model)	4.68	5.83	2.15	3.93	3.415	4.88
-None-context	4.56	5.61	1.72	3.87	3.14	4.74
-Source-context	3.58	4.23	1.71	3.43	3.145	3.83

5.3 Case Study

Table 6. Examples of essay generated by our model and baselines

Input Topics	蜜蜂，燕子，大雁，花丛，青蛙 (Bee, Swallow, Geese, Flower, Frog)
PPG	小燕子从南方飞回来了，花丛中，青蛙在呱呱地叫着，蜜蜂在跳舞，大雁飞回南方，飞回南方。 (Swallows fly back from the south, frogs croak in the flowers, bees dance, and wild geese fly back to the south. Wild geese fly back to the south.)
SC-LSTM	小燕子飞到南方去了，花丛中有许多蜜蜂在采蜜。青蛙在唱歌，小燕子在跳舞。蝴蝶在跳舞。小燕子在唱歌。花丛中有许多小鸟在唱歌。 (Swallows fly back to the south. There are many bees collecting honey in the flowers. Frogs are singing and swallows are dancing. Butterflies are dancing. Swallows are singing. Many birds are singing in the flowers.)
MTA-LSTM	春天，小燕子从南方飞回来了，大雁从南方飞回来了，小燕子在花丛中飞来飞去，青蛙在荷叶上呱呱地叫着，蜜蜂在荷叶上嗡嗡地叫着，好像在说：春天真美啊！小燕子也叽叽喳喳地叫着，好像在说：春天真美啊！ (In spring, swallows fly back from the south and geese fly back from the south. Swallows fly among the flowers, frogs croaked on the lotus leaf. Bees buzz on the lotus leaf as if to say: spring is really beautiful! Swallows also chirped, as if to say: spring is beautiful!)

Transformer	小燕子从南方飞回来了，大雁从南方飞回来了，小燕子在花丛中飞来飞去，青蛙在荷叶上呱呱地叫着，蜜蜂在花丛中忙碌着，蝴蝶在花丛中捉蝴蝶。(Swallows fly back from the south, and wild geese fly back from the south. Swallows fly among the flowers. Frogs croak on the lotus leaves, and bees are busy in the flowers, and the butterflies catch butterflies in the flowers.)
PC-SAN(our)	春天，万物复苏，大雁从南方飞回来了，小蜜蜂在花丛中飞来飞去采蜜，小青蛙在荷叶上呱呱地叫着，好像在说：春天来了！小燕子在空中飞来飞去，小蝴蝶翩翩起舞。 (In spring, everything revives, and geese fly back from the south. Bees fly around in the flowers to collect honey. Frogs croak on the lotus leaves as if to say: spring is coming! Swallows fly in the air, and butterflies dance.)

Table 6 shows the generated essays of our model and baselines with "*bee*", "*swallow*", "*geese*", "*flower*", and "*frog*" as input topics. As seen, we have translated the original Chinese output into English. The quality of generated essays of benchmark models is unsatisfactory. For the output of PPG, Transformer, and SC-LSTM, sentences are mostly semantical coherent separately, but the overall relevance of the essay is weak and the context is not cohesive. Moreover, there are massive duplicate phrases and self-contradiction issues in the generated text. MTA-LSTM further learns topic semantics information and topic distribution by utilizing the coverage vector, which indeed helps to infer that the whole essay revolves around the key theme "*spring*" according to given topics. However, due to insufficient linguistic information, the semantics gap makes the generated text uninformative and weak topic-consistency. By contrast, our model improves the quality of the generated essays. With the great help of the pre-trained language model, the proposed model can generate informative, diverse, and topic-consistent essays. Besides, the target-side contextual history SANs further eliminate the duplication and self-contradiction issues that occur in baselines.

6. Conclusion and Future Work

In this paper, we develop a pretraining-based contextual self-attention model (PC-SAN) for topic essay generation. Our approach focuses on the relationship and integrity of multiple topics, as well as takes into account the target-side contextual information. We apply BERT to the encoder to obtain the contextual embeddings of topic words and introduce the dynamic linear combination of all layers of BERT to enrich the semantics of topics. In addition, our model leverages the self-attention networks (SANs) to replace LSTM structures and transform the target-side contextual history information into the query layers of SANs to alleviate the lack of context in origin SANs. In terms of generation performance, the proposed PC-SAN consistently outperformed strong baselines on two large-scale datasets: ESSAY and ZhiHu. Extensive controlled experiments are conducted to verify the effects of our method. The experimental results prove that our model essentially improves performance compared to previous approaches and can generate informative, diverse, and topic-related essays.

Although this paper focuses on the topic essay generation, our approach is universal and can be applied to several natural language processing tasks. Further, the ideas of pretraining-based method and context-aware attention networks would also contribute to some multimodal works, such as joint semantic-visual space construction [35] and visual semantics representation [36,37].

In the future, we plan to adopt the text segmentation technique and integrate traditional template-based methods and neural networks to improve performance.

References

[1] Tosa, N., Obara, H., Minoh, M., "Hitch haiku : An interactive supporting system for composing haiku poem," in *Proc. of International Conference on Entertainment Computing. Springer*, pp. 209-216, 2008. Article (CrossRef Link)

[2] Yan, R., Jiang, H., Lapata, M., Lin, S.D., Lv, X., Li, X., "i, poet : automatic Chinese poetry composition through a generative summarization framework under constrained optimization," in *Proc. of the Twenty-Third international joint conference on Artificial Intelligence. AAAI Press*, pp. 2197-2203, 2013.

[3] Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., Ma, W.Y., "Topic aware neural response generation," in *Proc. of Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[4] Feng, X., Liu, M., Liu, J., Qin, B., Sun, Y., Liu, T., "Topic-to-essay generation with neural networks," in *Proc. of the 27th International Joint Conference on Artificial Intelligence. AAAI Press*, pp. 4078-4084, 2018. Article (CrossRef Link)

[5] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171-4186, 2019.

[6] He, J., Zhou, M., Jiang, L, "Generating chinese classical poems with statistical machine translation models," in *Proc. of Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[7] Yi, X., Sun, M., Li, R., Yang, Z., "Chinese poetry generation with a working memory model," in *Proc. of the 27th International Joint Conference on Artificial Intelligence. AAAI Press*, pp. 4553-4559, 2018. Article (CrossRef Link)

[8] Yi, X., Sun, M., Li, R., Li, W., "Automatic poetry generation with mutual reinforcement learning," in *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3143-3153, 2018. Article (CrossRef Link)

[9] Wang, Q., Luo, T., Wang, D., Xing, C., "Chinese song iambics generation with neural attention-based model," in *Proc. of the Twenty-Fifth International Joint Conference on Artificial Intelligence. AAAI Press*, pp. 2943-2949, 2016.

[10] Wang, Z., He, W., Wu, H., Wu, H., Li, W., Wang, H., Chen, E., "Chinese Poetry Generation with Planning based Neural Network," in *Proc. of COLING the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1051-1060, 2016.

[11] Sutskever, I., Vinyals, O., Le, Q.V., "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, pp. 3104-3112, 2014.

[12] Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., Bengio, Y., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 1724–1734, 2014. Article (CrossRef Link)

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

[14] Zhang, J., Feng, Y., Wang, D., Wang, Y., Abel, A., Zhang, S., Zhang, A., "Flexible and creative chinese poetry generation using neural memory," in *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1364–1373, 2017. Article (CrossRef Link)

[15] Yang, X., Lin, X., Suo, S., Li, M., "Generating thematic Chinese poetry using conditional variational autoencoders with hybrid decoders," in *Proc. of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 4539-4545, 2017. Article (CrossRef Link)

[16] Tu, Z., Lu, Z., Liu, Y., Liu, X., Li, H., "Modeling Coverage for Neural Machine Translation," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 76-85, 2016. Article (CrossRef Link)
[17] Dziri, N., Kamalloo, E., Mathewson, K.W., Zaiane, O., "Augmenting Neural Response Generation with Context-Aware Topical Attention," in *Proc. of the First Workshop on NLP for Conversational AI*, pp. 18–31, 2018. Article (CrossRef Link)
[18] Wu, Y., Wei, F., Huang, S., Wang, Y., Li, Z., Zhou, M., "Response generation by context-aware prototype editing," in *Proc. of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 7281-7288, 2019. Article (CrossRef Link)
[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., "Attention is all you need," *Advances in neural information processing systems*, pp. 5998-6008, 2017.
[20] Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y., "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.
[21] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., "Deep contextualized word representations," in *Proc. of NAACL-HLT*, pp. 2227-2237, 2018. Article (CrossRef Link)
[22] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., "Improving language understanding by generative pre-training," *Technical report, OpenAI*, 2018.
[23] Pennington, J., Socher, R., Manning, C. Glove., "Global vectors for word representation," in *Proc. of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543, 2014. Article (CrossRef Link)
[24] Raganato, A., Tiedemann, J., "An analysis of encoder representations in transformer-based machine translation," in *Proc. of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 287-297, 2018. Article (CrossRef Link)
[25] Yang, B., Li, J., Wong, D.F., Chao, L.S., Wang, X., Tu, Z., "Context-aware self-attention networks," in *Proc. of the AAAI Conference on Artificial Intelligence*, 2019. Article (CrossRef Link)
[26] Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G., "Regularizing neural networks by penalizing confident output distributions," *arXiv preprint arXiv:1701.06548*, 2017.
[27] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. Dropout, "a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, 15, 1929-1958, 2014.
[28] Papineni, K., Roukos, S., Ward, T., Zhu, W.J., "BLEU: a method for automatic evaluation of machine translation," in *Proc. of the 40th annual meeting on association for computational linguistics*, *Association for Computational Linguistics*, pp. 311-318, 2002. Article (CrossRef Link)
[29] Wen, T.H., Gasic, M., Mrkši'c, N., Su, P.H., Vandyke, D., Young, S., "Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems," in *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1711-1721, 2015. Article (CrossRef Link)
[30] Mazare, P.E., Humeau, S., Raison, M., Bordes, A., "Training Millions of Personalized Dialogue Agents," in *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2775-2779, 2018. Article (CrossRef Link)
[31] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J., "Personalizing Dialogue Agents: I have a dog, do you have pets too?," in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204-2213, 2018. Article (CrossRef Link)
[32] Meng, F., Lu, Z., Wang, M., Li, H., Jiang, W., Liu, Q., "Encoding Source Language with Convolutional Neural Network for Machine Translation," in *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 20-30, 2015. Article (CrossRef Link)

[33] Zhang, J., Zhang, D., Hao, J., "Local translation prediction with global sentence representation," in *Proc. of Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
[34] Wang, X., Tu, Z., Wang, L., Shi, S., "Exploiting Sentential Context for Neural Machine Translation," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6197–6203, 2019. Article (CrossRef Link)
[35] Hong, R., Li, L., Cai, J., Tao, D., Wang, M., Tian, Q., "Coherent semantic-visual indexing for large-scale image retrieval in the cloud," *IEEE Transactions on Image Processing*, 26, 4128-4138, 2017. Article (CrossRef Link)
[36] Shekhar, R., Takmaz, E., Fernández, R., Bernardi, R., "Evaluating the Representational Hub of Language and Vision Models," in *Proc. of the 13th International Conference on Computational Semantics - Long Papers*, pp.211–222, 2019. Article (CrossRef Link)
[37] Hong, R., Yang, Y., Wang, M., Hua, X.S., "Learning visual semantic relationships for efficient visual retrieval," *IEEE Transactions on Big Data*, 1, 152-161, 2015.

Fuqiang Lin received the bachelor's degree in Computer Science from National University of Defense Technology in 2017. Now he is studying for the Ph.D. degree, major in social network analysis and natural language processing.

Xingkong Ma received the B.S. degree in computer science and technology from the School of Computer, Shandong University,China, and the M.S., Ph.D. degree in Computer Science and Technology from the National University of Defense Technology, China. He is an associate professor in the College of Computer Science, National University of Defense Technology. His current research interests include the areas of data dissemination and publish/subscribe.

Yaofeng Chen received the B.S. degree in Software Engineering from Central South University in 2013 and received the M.S. degree in Computer Science from National University of Defense Technology in 2016. Since he has been researching in the field of Social Media Analysis and continues to pursue the Computer Science Ph.D. degree in National University of Defense Technology.

Jiajun Zhou is a doctoral candidate in the College of Computer, National University of Defense Technology, Changsha, China. His current research interests include a broad area of data mining, social network mining, personalized recommendation for location-based services.

Bo Liu is an professor in the College of Computer, National University of Defense Technology, Changsha, China. His research interests include social network mining, recommendation systems, network security and blockchain technology.