KTGAT: Improving the Robustness of Knowledge-enhanced Text Generation via Adversarial Training

HaiXiang Zhu
School of Computer Science
National University of Defense Technology
Changsha, China
zhuhaixiang_zhx@nudt.edu.cn

YiPing Song
College of Science
National University of Defense Technology
Changsha, China
songyiping@nudt.edu.cn

Bo Liu
School of Computer Science
National University of Defense Technology
Changsha, China
* Corresponding author: kyle.liu@nudt.edu.cn

2023 4th International Conference on Computer Engineering and Application (ICCEA) | 979-8-3503-4754-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCEA58433.2023.10135313

***Abstract*—The shortage of information in text generation has been a prominent area of research in Natural Language Processing (NLP). Current research endeavors aim to combine pre-trained models with rich open-world knowledge from external sources to increase the priori information and thereby enhance the informativeness of text generation. While recent studies suggest that integrating open-world and task-specific knowledge can improve text generation by addressing specific knowledge gaps in downstream tasks, the inherent semantic ambiguity in natural language remains a significant challenge that may impede knowledge acquisition and text generation. To overcome this challenge and improve the model's semantic comprehension and overall robustness, we propose a novel framework, the Knowledge Augmentation Text Generation model via Adversarial Training (KTGAT). Our method adds perturbations to the embedding layer, which are equivalent to constructing unstable samples. This approach improves the model's robustness to adversarial samples and the generalization performance of the original samples. Our experiments demonstrate that the proposed KTGAT framework outperforms the baseline model, thus proving its effectiveness in improving text generation. The generated text cases illustrate that our method enhances the model's semantic comprehension and enables it to search for knowledge items more effectively and accurately.**

Keywords-Text generation, Pre-trained models, Knowledge fusion, Adversarial training

I. INTRODUCTION

Natural Language Generation (NLG) is a challenging and crucial area in the field of Artificial Intelligence that has gained significant attention in recent years, mainly due to the growth of big data and computing power[41,42]. The primary goal of NLG is to enable models to generate coherent and meaningful human-like language text. Various techniques have been developed for text generation, including machine translation and question-answering systems. However, current generative models based on deep neural networks and pre-training[1] typically use end-to-end architectures and have limited ability to understand the complexities of human language due to the restricted knowledge contained in the input. In practical applications, the performance of text generation models needs improvement as the output generated is often mundane and nonsensical[2]. Despite being the latest advancements in NLP research[43], the generated text still lacks logic and information, leaving significant room for enhancing the performance of NLG models.

In light of this, recent studies in the domain of natural language generation have focused on the development of knowledge-based models that can incorporate external world knowledge beyond the input[3] to address the knowledge deficit in text generation. Typically, knowledge graphs and online encyclopedias are utilized as sources of knowledge. Incorporating such knowledge has demonstrated significant advancements in generating high-quality, coherent language output by enhancing the model's understanding of the input and its context. However, this approach is currently limited to scenarios with simple semantic relations, while human-generated discourse draws on a wealth of learned knowledge and contextual information. To improve NLG models' effectiveness and naturalness, exploring the integration of learned knowledge and context-specific information is essential. For instance, in conversations, humans often utilize knowledge from previous discussions to generate subsequent responses. Hence, incorporating learned knowledge and context-specific information could enable models to generate more accurate and coherent text.

Such work has been explored by researchers, ERNIE[37] solves the model's text generation knowledge deficit by pre-processing the entity features in the pre-trained corpus; SKT[39] models long time knowledge hidden variables, and information from previous conversations is used to generate the current conversation. CCM[40] combines knowledge graphs and graph neural networks to enhance text generation by understanding user discourse through graph attention mechanisms and searching for corresponding knowledge in the knowledge graph. TEGTOK[4] utilizes Wikipedia documents as source of open-world knowledge, and also incorporates task-specific knowledge related to the target of the current conversation to help the model produce more accurate and coherent output. PLATO-K[38] enhances text generation with extensive

knowledge from pre-trained corpus and knowledge from conversational contexts. The aforementioned conventional dialogue systems that incorporate knowledge often struggle to maintain coherent conversations due to their limited ability to understand the semantics of the input and the knowledge that they possess. This is caused by the linguistic ambiguity of the the linguistic ambiguity of the text, where the same word may mean different things in different contexts, which can mislead the model leading to disjointed and incoherent dialogue, where the system jumps from one topic to another without proper contextualization or continuity.

To overcome the challenge of maintaining coherence in dialog systems and to ensure engagement and relevance in responses. We propose the Knowledge-enhanced Text Generation via Adversarial Training (KTGAT) framework to overcome this challenge. Specifically, the ability of a language model to understand linguistic ambiguities and search for corresponding knowledge is linked to its semantic understanding. This understanding is conveyed through text embedding, which expresses word meanings and relationships. We use adversarial training in the embedding layer to enhance this process to generate controlled noise added to a continuous embedding vector. Compared to adding directly to discrete input text, this approach thus allows the model to understand the semantics better and search for knowledge to generate coherent outputs. In each training round, controlled noise is added to the embedding layer, followed by the backpropagation of losses. Adding noise to the embedding layer makes the model better understand the context and integrates external knowledge during the generation process. Thus, the proposed KTGAT framework effectively leverages task-specific and open-world knowledge to generate thematically consistent and contextually relevant high-quality text. This approach could improve the performance of text generation models, thereby improving the accuracy and coherence of natural language generation.

Our study presents the KTGAT framework, incorporating adversarial training to enhance knowledge-based text generation. Introducing controlled noise in the embedding layer improves the model's semantics comprehension and facilitates the generation of thematically coherent output. Experimental evaluations on two datasets demonstrate the superiority of our proposed method over the baseline model.

II. Related Works

A. Knowledge-enhanced text generation

Incorporating knowledge sources has been shown to improve the ability of models to understand context and reference entities in the output, and previous research has explored the role of knowledge bases in text generation[5]. Regarding unstructured knowledge sources, related studies explored retrieval-based enhanced text generation[6], which follows two phases: retrieval and generation. Subsequent studies have introduced retrieval and rewriting techniques[7] to improve further knowledge search and text generation performance in specific tasks. This paper investigates the two tasks of dialogue generation and question generation under text generation.

Dialogue systems aim to give appropriate responses from the system based on the user's queries and are usually treated as translation problems from input to output and use end-to-end models based on encoders and decoders. However, trainingmodels relying on just raw data can generate mundane and repetitive discourse. Various conditions, such as emotion[9] and persona[10], have been added to the dialogue to increase response diversity. Furthermore, to enhance the informativeness of dialogue generation, external knowledge has been incorporated into subsequent research[11].

This task bears similarities to reading comprehension, as it requires the model to generate a question based on the given document[12]. Previous studies have approached this task by extracting keywords based on task-specific characteristics and generating questions using these keywords[13]. More recent research has incorporated pre-trained models[14] to enhance the quality of the generated questions. Given the highly structured nature of the task, researchers have also introduced graph neural networks[15] to address this problem.

B. Adversarial Training

Adversarial training is a meaningful way to enhance the robustness of neural networks. In adversarial training, the samples are mixed with some tiny perturbations, and then the neural network is adapted to this change, making it robust to adversarial samples. In image processing, adversarial training may cause the model's performance to degrade on the correct sample, but in the field of NLP, there is a performance gain for the model[35].

FGSM[35] is to maximize the loss by making the direction of perturbation follow the direction of gradient boosting so that even minor perturbations are amplified to interfere with the training of the model, thereby improving the model's resistance and stability in complex situations. Miyato et al.[18] and Madry et al.[19] optimized FGSM by improving the normalization method and performing multiple perturbation addition that adds more perturbations to the model, They verified the effectiveness of adversarial training on simple text classification tasks.

III. Method

A. Model Architecture

Given an input text, the goal of text generation is to input an appropriate natural language discourse. Figure 1 illustrates our proposed text generation framework KTGAT, which contains the knowledge-enhanced text generation model and the **noise adder** for adversarial training. The text generation model contains a **retriever** and a **generator**, where the retriever is used to retrieve the open-world knowledge and task-specific knowledge based on the input text, while the generator generates based on the retrieved knowledge and the original input text. The noise adder adds noise to the

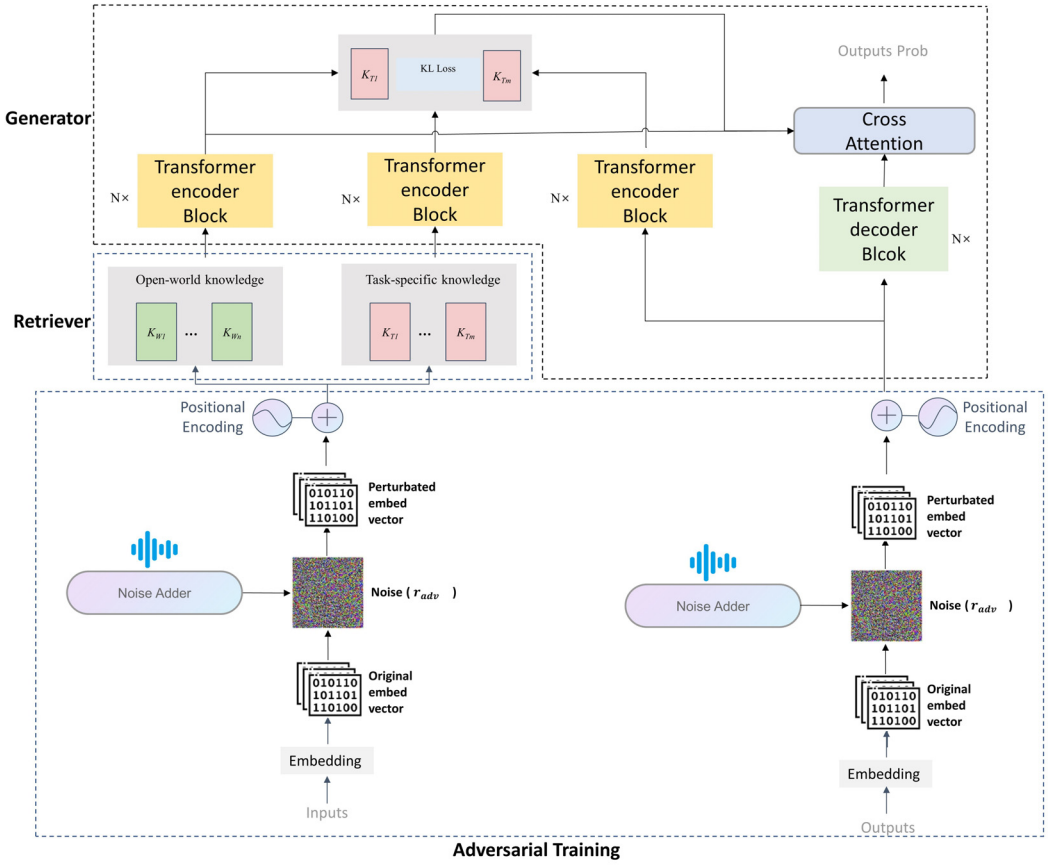

Figure 1. The Overview of our proposed KTGAT model which consist a knowledge-enhanced language model and noise adder for adversarial training.

embedding vector of the input text in the embedding layer of the model for adversarial training.

1) Retriever. The retriever first indexes all entries to retrieve knowledge entries relevant to the input and context during text generation. Given that the knowledge is divided into open-world and task-specific categories, two separate retrievers are designed: **Open-World** and **Task-specific.**

The retrieval process during text generation involves indexing all relevant knowledge entries and computing the vector similarity between the input text and each entry using a similarity metric function. The fast maximum inner-product search (MIPS)[36] technique is employed to improve retrieval effectiveness and efficiency. Pre-trained models encode both input and knowledge text as low-dimensional and continuous vectors. The Dense Passage Retriever (DPR) model[16] is adopted as the open-world knowledge retriever due to its effectiveness in retrieving information from Wikipedia, which serves as the open-world knowledge base with over 21 million articles. In addition to open-world knowledge, task-specific knowledge is incorporated into the model to enhance generative performance. A comparative learning-based approach is employed to train the task-specific knowledge retriever using a training example consisting of input text, matching knowledge, and a negative sample randomly selected from the knowledge base.

2) Generator. We utilize pre-trained BART model weights as initialization parameters for each module of the generator to integrate input and selected knowledge. The generator contains two encoders: open-world knowledge and input text encoder, and task-specific knowledge encoder. For the retrieved input text and open world knowledge, the tokens of both are first connected to obtain a new sequence, which is then fed to the corresponding encoder. For the encoder of specific knowledge, it is important to rank and select the most critical task knowledge entries since the model cannot see the future information at the time of testing. The retrieved task-specific knowledge entries are organized into a set of tokens and fed to the corresponding encoders. To determine the similarity between the output posterior representation and the input text and each task knowledge entry. In the training process, the posterior representation of the output with respect to the input is first computed using a fully connected neural network, then the similarity distributions of the input and output to each task-specific knowledge entry are computed separately, and the distance between these two distributions is measured by introducing a KL loss so that the model can

select the most suitable task-specific knowledge, and finally a BOW loss is added to optimize the training process.

In summary, the training objective of the model is formalized as, given a text pair denoted as (s,t), s is the input text and t is the output text. Considering the open-world knowledge variable z_1 and the task-specific knowledge variable z_2, the learning objective of the model is:

$$p(t \mid s) = \sum_{z_1, z_2} p_1(z_1 \mid s) p_2(z_2 \mid s) p_\theta(t \mid s, z_1, z_2) \quad (1)$$

where $z_1 \in \text{top} - m(p_1(\cdot \mid s))$ and $z_2 \in \text{top} - m(p_2(\cdot \mid s))$, $p_1(\cdot \mid s)$ and $p_2(\cdot \mid s)$ are modeled as a retrieval probability, which modeled by $sim(s, k_i^\alpha) = E_S^\alpha(s)^\top \cdot E_K^\alpha(k_i^\alpha), i \in \{1,2,\ldots\}$, $sim(,)$ is the similarity function, $E_S^\alpha(s)$ and $E_K^\alpha(k_i^\alpha)$ are input text and knowledge entry's representation respectively.

.

B. Adversarial Training

Semantic understanding is a critical factor in knowledge search and text generation. However, a word usually expresses different meanings in different contexts, and this linguistic ambiguity makes the process of knowledge search by the model unstable. We can use adversarial training to improve the model's understanding of text semantics to improve robustness. This regularization technique enhances robustness to small or worst-case perturbations by introducing perturbations r_{adv} in the input x. The model is trained to acquire the ability to maintain the correct target y of the output under perturbation. The goal of resistance training usually involves improving the model's ability to handle such perturbations:

$$-\log p(y \mid x + r_{adv}; \theta)$$
$$\text{where } r_{adv} = \arg\min_{r, \|r\| \le \epsilon} \log p(y \mid x + r; \hat{\theta}) \quad (2)$$

where θ is model parameter and $\hat{\theta}$ is the a constant copy of the model parameters.

FGM and PGD are two gradient information-based attack methods that have excellent performance and reliability in text classification tasks, and are highly interpretable and reproducible. In text generation, controlling the size of perturbations is crucial to generate high-quality text, and FGM and PGD can generate adversarial perturbations in a controlled range, so we choose them as our adversarial training methods in this paper.

1) FGM. The FGM (Fast Gradient Method)[18] is an adversarial training technique involving perturbing attacks and model defense. Perturbing attacks add noise to the embedding layer to create adversarial instances and model defense when the model propagates forward under the disturbance and calculates the loss. The objective of our training is to minimize the objective function of adversarial training. It is important to note that we use a copy of the constant $\hat{\theta}$ in the objective function instead of the original θ because the backpropagation algorithm does not propagate the gradient while creating the adversarial instances.

In each round of training, we determine the perturbation according to Equation (3) and then train the model to combat this perturbation relative to θ. However, it is challenging for neural networks to compute this relative value exactly. Therefore, we employ a linearization of the function $\log p(y \mid x; \hat{\theta})$ around a given point to approximate this value, which can then be computed by backpropagation in the neural network. After adding the linear approx-imation and the L2 parametrization, the resulting adversarial perturbation is obtained:

$$r_{adv} = -\epsilon g / \| g \|_2$$
$$\text{where } g = \nabla_x \log p(y \mid x; \hat{\theta}) \quad (3)$$

In addition to basic adversarial training, the FGM method incorporates virtual adversarial training[18] as a regularization technique. Virtual Adversarial Training aims to solve the task under semi-supervised learning by replacing the ground truth y in the original adversarial learning with the model's output. And then making the model adaptively regularized toward minimizing the loss. This method introduces additional training costs, which are as follows:

$$\text{KL}[p(\cdot \mid x; \hat{\theta}) \| p(\cdot \mid x + r_{v-adv}; \theta]$$
$$\text{where } r_{v-adv} = \arg\max_{r, \|r\| \le \epsilon} \text{KL}[p(\cdot \mid x; \hat{\theta}) \| p(\cdot \mid x + r; \hat{\theta})] \quad (4)$$

In virtual adversarial training, the KL divergence between distributions is denoted as $\text{KL}[p \| q]$, and additional training costs are introduced by minimizing Equation (4). This approach enables the model to resist perturbations most sensitive to the current $p(y \mid x; \hat{\theta})$. It is worth noting that to minimize Equation (3), only the input x is required, not the target y. As calculating the loss of virtual adversarial training analytically by hand can be challenging, backpropagation is used to approximate Equation (4) in the neural network.

The above describes the algorithmic of FGM, specifically in our task, the implementation of FGM is, we add a perturbation attack at the embedding layer, given the embedding vector $E = [e_1^s, e_2^s, \ldots, e_{l_s}^s]$ of the input text s, where e_i^s is the embedding of the $i-\text{th}$ word in the input text and l_s is the length of input, denote the conditional probability of the target text t given E as $p(t \mid s; \theta)$, where θ is model parameter, and then add the adversarial perturbation r_{adv}:

$$r_{adv} = -\epsilon g / \| g \|_2$$
$$\text{where } g = \nabla_s \log p(t \mid s; \hat{\theta}) \quad (5)$$

To improve the robustness of the model against the adversarial perturbations r_{adv} defined in the equation (5), the adversarial loss is as follows:

$$L_{adv}(\theta) = -\prod_{n=1}^{N} \log p(t_n \mid s_{1:n} + r_{adv}, t_{1:n-1} + r_{adv}; \theta) \quad (6)$$

where n is current training step. In our experiments, adversarial training refers to minimizing the negative log-likelihood plus L_{adv} with stochastic gradient descent.

During virtual adversarial training, the following virtual adversarial perturbations are calculated for each training round:

$$r_{v-adv} = \epsilon g / \| g \|_2$$
$$\text{where } g = \nabla_{s+d} \mathrm{KL}[p(\cdot \mid s; \hat{\theta}) \| p(\cdot \mid s + d; \hat{\theta})] \quad (7)$$

where d is a TD-dimensional small random vector. Then, the virtual adversarial loss is defined as

$$L_{v-adv}(\theta) = \prod_{n=1}^{N} \mathrm{KL}[p(\cdot \mid s_{1:n}, t_{1:n}; \hat{\theta}) \| p(\cdot \mid s_{1:n} + r_{v-adv}, t_{1:n} + r_{v-adv}; \theta)] \quad (8)$$

2) PGD. Project Gradient Descent (PGD)[19] is a more sophisticated approach than the single iteration of FGM. PGD attacks involve multiple iterations during training, each projecting perturbations to a specific range. The perturbation input for each iteration is derived based on the input and gradient of the previous iteration. If the derived perturbation is outside a specific range, it is "pulled back" to a fixed range. Formally, at each step of the iteration, the following steps are taken:

$$r_{adv}^{n+1} = \prod_{\|r_{adv}\|_F \le \epsilon} (r_{adv}^{t} + \alpha g(r_{adv}^{n}) / \| g(r_{adv}^{n}) \|_2)$$
$$\text{where } g(r) = \nabla_{r_{adv}} L(f_\theta(s + r_{adv}^{n}), t) \quad (9)$$

Here, $\| r_{adv} \|_F \le \epsilon$ represents the constrained space of the perturbation, $\prod_{\|r_{adv}\|_F \le \epsilon}$ is the projection function that projects the perturbation r_{adv} beyond the boundary ϵ into the $\epsilon - \text{ball}$ and $g(r)$ is the gradient of last step, as illustrated in the Figure 2. The adversarial sample obtained by using the first-order gradient in PGD is referred to as a "first-order adversarial".

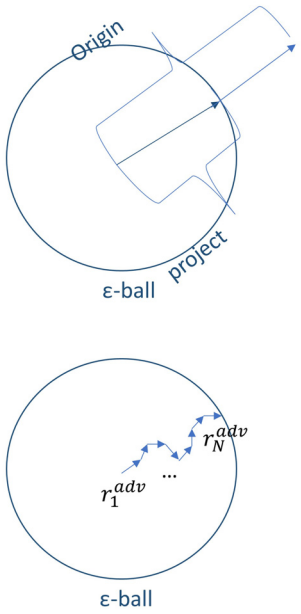

Figure 2. Schematic diagram of multiple projections of PGD, where ϵ-ball is the restricted range of projection.

IV. Experimental Settings

A. Datasets.

1) Dialogue generation. We use the Reddit dataset[20] as a conversation corpus. This dataset consists of posts and comments from the social forum Reddit, a well-known social news forum capable of reaching billions of users. The site is updated with a large number of posts and daily replies. The Reddit corpus spans the period from November 2016 to August 2018. In this paper, three million random data extracted from the dataset are used to build the task-specific knowledge base, and the rest is used for training and testing.

2) Question Generation. SQuAD[21] was first proposed for reading comprehension, where the model answers a question given a long document and a question related to the document's content. It was used as a question generation task because it enhanced the model's natural language understanding. The dataset is a document and a question, and the model is trained to generate questions about the document. In this task, forty-five thousand pieces of data were randomly selected. Then, as with Reddit, the dataset was divided into two parts, one for building the task-specific database and one for model training and testing.

3) Open-World Knowledge. Wikipedia is a multilingual encyclopedic knowledge base for the Internet that contains up to 55 million knowledge entries as of 2021. The free and open policy it adopts allows browsing users to edit most entries at will, while professional website editors are also available to do the content review. This paper uses the Wikipedia dump provided by Lee et al.[22] as the open-world knowledge base.

B. Evaluation Metrics

We employed a range of automated evaluation methods to assess the performance of our model. Specifically, we utilized the BLEU[23] metric to determine the extent of word overlap between the generated text and the reference text. We employed the METEOR[24] metric to measure the exact word-to-word matching rate. Additionally, we employed the ROUGE-L[25] metric to evaluate the longest subsequence match between the generated and reference texts. To further assess text similarity, we introduced word vector evaluation metrics that quantify the degree of similarity between two texts based on word similarity[26]. The metrics employed for this purpose included Embedding Average, Greedy Matching, and Vector Extrema.

Table I. Performance of our method and baselines on the Reddit dataset. The Best results in each group are highlight with bold. FGM and PGD represent different Adversarial training methods.

Method	BLEU-1	BLEU-2	ROUGE-L	Average	Greedy	Extrema
CVAE	7.45	2.85	9.68	0.6642	2.0853	0.3357
Transformer	7.97	3.14	10.51	0.6693	2.0703	0.3334
GPT-2	8.43	3.04	10.65	0.6484	2.0601	0.3303
DialoGPT	7.58	3.02	10.82	0.5976	2.0774	0.3185
BART	9.24	3.38	10.93	**0.6611**	2.0986	**0.3355**
TEGTOK	9.35	3.52	11.17	0.6433	2.1590	0.3329
KTGAT-FGM	**9.55**	**3.61**	**11.35**	0.6486	**2.1760**	0.3343
KTGAT-PGD	9.38	3.52	11.25	0.6497	2.1730	0.3371

Table II**.** Performance of our method and baselines on the SQuAD dataset. The Best results in each group are highlight with **bold**.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
RNN	31.34	13.79	7.36	4.26	29.75	9.88
DirectIn	31.71	21.18	15.11	11.20	22.47	14.95
H&S	38.50	22.80	15.52	11.18	30.98	15.95
NQG	43.09	25.96	17.50	12.28	39.75	16.62
TEGTOK	**46.52**	30.61	22.24	16.71	43.79	**20.36**
KTGAT-FGM	46.51	**30.71**	**22.40**	**16.98**	**43.80**	20.34
KTGAT-PGD	46.39	30.60	22.32	16.92	43.75	20.34

C. Baselines

(1)**RNN**[8], A sequence-to-sequence model containing an encoder and decoder, with temporal dependencies constructed by an LSTM; (2) DirectIn by simply taking the longest sentence in the sentence as the prediction of the model; (3) H&S[33] generates multiple questions based on the rules of (subject, relation, object), and then sorts the generated questions; (4) NQG[34] introduces an attention mechanism and a global attention mechanism to help model generation; (5)

CVAE[27], a variational self-encoder that models text distribution by learning latent variables and can specify labels at regeneration time; (6) Transformer[28], an encoder-decoder model based entirely on a self-attentive mechanism; (7) GPT-2[29], a model using the Transformer decoder architecture entirely, achieves robust performance in most natural language tasks by pre-training on a large corpus; (8) DialoGPT[30] uses the architecture of the GPT family, and differs from GPT-2 in that the pre-trained corpus is large-scale conversational data; (9) BART[17] combines the features of auto-regressive and auto-encode models by using random noise to disrupt the text structure for training, which simultaneously takes into account contextual information and makes the model suitable for text generation tasks

D. Implementation details

We utilized the pre-trained BART model as the initialization parameter of our model, leveraging the parameters and model interface provided by the esteemed Hugging Face library[31]. The encoder and decoder share a common word embedding, while the batch size is set at 64. We employed the AdamW[32] optimizer and employed a beam size of 3 during the decoding stage. Our training process involved 15 epochs, and we conducted it on a single RTX 3090.

V. Experiment Results

A. Overall Results

Tables I and II show the performance results of our proposed method and the baseline models on the two datasets for both tasks. The results show that our proposed method outperforms other methods on almost all evaluation metrics for both datasets. In particular, on the Reddit dataset, our method KTGAT-FGM improves BLEU-1 by 12% over GPT-2, DialoGPT by 25%, and BART by 3.3%, respectively, which demonstrates the effectiveness of fusing external knowledge. The improvement over TEGTOK was 2.1%. In addition, we observed that KTGAT_FGM and KTGAT-PGD improved by 1.6% and 0.7%, respectively, in ROUGE-L over the previous best method, TEGTOK, demonstrating the effectiveness of the

adversarial training incorporation. In the SQuAD task, KTGAT-FGM significantly outperformed the baseline model NQG, and BLEU-2 and BLEU-3 improved by 18.2% and 28%, respectively, further confirming the effectiveness of knowledge-enhanced text generation. In addition, our method improves 0.3% and 0.7% in BLEU-2 and BLEU-3,

Table III. The results of the two cases were generated from the Reddit and SQuAD datasets, respectively. Words of the same color are semantically related.

Reddit
Query: oh am sure all those people in fph are only there because they are really worried about the fat peoples health ! they 're all there to bully them into being healthier right ?
Open-world Knowledge: fact every body, can be happy, well and desirable. \"EVERYbody is a good body,\" was a PPPO catchphrase. PPPO celebrated the fat body with pride and politicized it too by refusing diets and denial.
Task-specific Knowledge: i 've found that the only people who have a problem with fph are fat asses or chubby chasers.
Ground_Truth: nope . we 're there to laugh at the collective delusion of the obese community .
TEGTOK: i'm sure they're all there because they want to make sure they don't die.
KTGAT-FGM: i'm sure they're all there to bully fat people into being healthier. i'm just saying that it's not a good thing to be a fat person.
KTGAT-PGD: i'm sure they're there to bully fat people into being healthy.
SQuAD
Passage: around 2 pm pdt -lrb- 21:00 utc -rrb- , the torch resurfaced about 3 km -lrb- 1.9 mi -rrb- away from the stadium along van ness avenue , a heavily trafficked thoroughfare that was not on official route plans .
Open-world Knowledge: NoMa includes the neighborhoods of Truxton Circle, Sursum Corda, Eckington, and Near Northeast and includes a section historically known as Swampoodle. Its southern tip is four blocks from Capitol Hill.
Task-specific Knowledge: when did the second line of metrobus open?
Ground_Truth: after the bearer of the torch disappeared into a warehouse , what road did the relay reappear on ?
TEGTOK: when did the torch resurface?
KTGAT-FGM: how far away from the stadium did the torch resurface?
KTGAT-PGD: how far away from the stadium did the torch resurface?

respectively, compared to the previous best method TEGTOK. KTGAT-PGD has a certain degree of performance degradation compared to TEGTOK, which we speculate is due to the multiple perturbation steps of PGD, resulting in a model that does not understand the semantics well and is challenging to train.

In addition, we compared two adversarial training methods , FGM and PGD. the results showed that both adversarial training methods exhibited stronger performance than the original method on most metrics, which indicates the effectiveness of adversarial training. Specifically, FGM showed better results than PGD. This may be due to the fact that PGD's multiple perturbations semantic understanding becomes more complex and therefore harder to converge.

B. Case Study

We presented case studies in Table III that compared our proposed method with TEGTOK on two datasets. For the Reddit dataset, our approach demonstrated the ability to retrieve relevant information from two knowledge bases, given a query, and effectively incorporate this information to enhance the generated text. For instance, if the query contains information about "fat" and "health," our method leverages these keywords to retrieve related entries from the knowledge base, seamlessly integrated into the generated text.

On the SQuAD dataset, the ground truth question asks about the torch resurfaced road, which strongly correlates with the concept of location. Our method was able to locate the entity "3 km -lrb- 1.9 mi -rrb- " based on this central theme of location in the passage and used this entity to search for relevant information in the knowledge base. Consequently, our approach accurately generated the location-related question, "how far from."

Overall, these case studies demonstrate the effectiveness of our proposed method in improving text generation performance by enhancing semantic understanding and leveraging relevant knowledge.

VI. CONCLUSION

This paper studied text generation tasks that use knowledge to improve model performance. We used two types of knowledge: open-world knowledge and task-specific knowledge. These were integrated with the encoding and decoding phases of text generation models to enhance the informativeness of text generation. Further, we propose to use adversarial training to enhance the semantic understanding of the model to help the model use knowledge better. The results of two experiments on dialogue and question generation show that our approach outperforms previous approaches in performance.

REFERENCES

[1] Radford, Alec and Karthik Narasimhan. “Improving Language Understanding by Generative Pre-Training.” (2018).

[2] Zhou, Hao et al. “Commonsense Knowledge Aware Conversation Generation with Graph Attention.” International Joint Conference on Artificial Intelligence (2018).

[3] Ji, Shaoxiong et al. “A Survey on Knowledge Graphs: Representation, Acquisition, and Applications.” IEEE Transactions on Neural Networks and Learning Systems 33 (2020): 494-514.

[4] Tan, Chaohong et al. “TegTok: Augmenting Text Generation via Task-specific and Open-world Knowledge.” ArXiv abs/2203.08517 (2022): n. pag.

[5] Liu, Ye et al. “KG-BART: Knowledge Graph-Augmented BART for Generative Commonsense Reasoning.” ArXiv abs/2009.12677 (2020): n. pag.

[6] Lewis, Patrick et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” ArXiv abs/2005.11401 (2020): n. pag.

[7] Ren, Shuo et al. “A Retrieve-and-Rewrite Initialization Method for Unsupervised Machine Translation.” Annual Meeting of the Association for Computational Linguistics (2020).

[8] Sutskever, Ilya et al. “Sequence to Sequence Learning with Neural Networks.” NIPS (2014).

[9] Song, Zhenqiao et al. “Generating Responses with a Specific Emotion in Dialog.” Annual Meeting of the Association for Computational Linguistics (2019).

[10] Zhang, Saizheng et al. “Personalizing Dialogue Agents: I have a dog, do you have pets too?” Annual Meeting of the Association for Computational Linguistics (2018).

[11] Moon, Seungwhan et al. “OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs.” Annual Meeting of the Association for Computational Linguistics (2019).

[12] Du, X. et al. “Learning to Ask: Neural Question Generation for Reading Comprehension.” Annual Meeting of the Association for Computational Linguistics (2017).

[13] Cho, Jaemin et al. “Mixture Content Selection for Diverse Sequence Generation.” ArXiv abs/1909.01953 (2019): n. pag.

[14] Chan, Ying-Hong and Yao-Chung Fan. “A Recurrent BERT-based Model for Question Generation.” Conference on Empirical Methods in Natural Language Processing (2019).

[15] Kipf, Thomas and Max Welling. “Semi-Supervised Classification with Graph Convolutional Networks.” ArXiv abs/1609.02907 (2016): n. pag.

[16] Karpukhin, Vladimir et al. “Dense Passage Retrieval for Open-Domain Question Answering.” ArXiv abs/2004.04906 (2020): n. pag.

[17] Lewis, Mike et al. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.” Annual Meeting of the Association for Computational Linguistics (2019).

[18] Miyato, Takeru et al. “Adversarial Training Methods for Semi-Supervised Text Classification.” arXiv: Machine Learning (2016): n. pag.

[19] Madry, Aleksander et al. “Towards Deep Learning Models Resistant to Adversarial Attacks.” ArXiv abs/1706.06083 (2017): n. pag.

[20] Dziri, Nouha et al. “Augmenting Neural Response Generation with Context-Aware Topical Attention.” ArXiv abs/1811.01063 (2018): n. pag.

[21] Rajpurkar, Pranav et al. “SQuAD: 100,000+ Questions for Machine Comprehension of Text.” Conference on Empirical Methods in Natural Language Processing (2016).

[22] Lee, Kenton et al. “Latent Retrieval for Weakly Supervised Open Domain Question Answering.” ArXiv abs/1906.00300 (2019): n. pag.

[23] Papineni, Kishore et al. “Bleu: a Method for Automatic Evaluation of Machine Translation.” Annual Meeting of the Association for Computational Linguistics (2002).

[24] Banerjee, Satanjeev and Alon Lavie. “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.” IEEvaluation@ACL (2005).

[25] Lin, Chin-Yew. “ROUGE: A Package for Automatic Evaluation of Summaries.” Annual Meeting of the Association for Computational Linguistics (2004).

[26] Forgues, Gabriel, Pineau, Joelle, Larchevêque, Jean-Marie and Tremblay, Réal. "Bootstrapping dialog systems with word embeddings." Paper presented at the meeting of the Nips, modern machine learning and natural language processing workshop, 2014.

[27] Zhao, Tiancheng et al. “Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders.” ArXiv abs/1703.10960 (2017): n. pag.

[28] Vaswani, A. , et al. "Attention Is All You Need." arXiv (2017).

[29] Radford, Alec et al. “Language Models are Unsupervised Multitask Learners.” (2019).

[30] Zhang, Yizhe et al. “DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation.” Annual Meeting of the Association for Computational Linguistics (2019).

[31] Wolf, Thomas et al. “HuggingFace's Transformers: State-of-the-art Natural Language Processing.” ArXiv abs/1910.03771 (2019): n. pag.

[32] Loshchilov, Ilya and Frank Hutter. “Decoupled Weight Decay Regularization.” International Conference on Learning Representations (2017).

[33] Serban, Iulian et al. “Generating Factoid Questions With Recurrent Neural Networks: The 30M Factoid Question-Answer Corpus.” ArXiv abs/1603.06807 (2016): n. pag.

[34] Du, X. et al. “Learning to Ask: Neural Question Generation for Reading Comprehension.” Annual Meeting of the Association for Computational Linguistics (2017).

[35] Goodfellow, Ian J. et al. “Explaining and Harnessing Adversarial Examples.” CoRR abs/1412.6572 (2014): n. pag.

[36] Shrivastava, Anshumali and Ping Li. “Asymmetric LSH (ALSH) for Sublinear Time Maximum Inner Product Search (MIPS).” ArXiv abs/1405.5869 (2014): n. pag.

[37] Zhang, Zhengyan et al. “ERNIE: Enhanced Language Representation with Informative Entities.” Annual Meeting of the Association for Computational Linguistics (2019).

[38] Bao, Siqi et al. “PLATO-K: Internal and External Knowledge Enhanced Dialogue Generation.” ArXiv abs/2211.00910 (2022): n. pag.

[39] Kim, Byeongchang et al. “Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue.” ArXiv abs/2002.07510 (2020): n. pag.

[40] Zhou, Hao et al. “Commonsense Knowledge Aware Conversation Generation with Graph Attention.” International Joint Conference on Artificial Intelligence (2018).

[41] Chen, Wangqun et al. “Jointly Learning Sentimental Clues and Context Incongruity for Sarcasm Detection.” IEEE Access PP (2022): 1-1.

[42] Dong, Diwen et al. “Sentiment-Aware Fake News Detection on Social Media with Hypergraph Attention Networks.” 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (2022): 2174-2180.

[43] Chen, Wangqun et al. “Commonsense-Aware Sarcasm Detection with Heterogeneous Graph Attention Network.” 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (2022): 2181-2188.