A Scene-Aware Prompt Template of Chinese Text Data Augmentation Method Based on Large Language Model

Lingfei Kong
School of Data Science and Applications
Inner Mongolia University of Technology
Hohhot, China
18709879027@163.com

Gang Wang*
School of Data Science and Applications
Inner Mongolia University of Technology
Hohhot, China
wg@imut.edu.cn
*Corresponding author

Jingheng He
School of Data Science and Applications
Inner Mongolia University of Technology
Hohhot, China
jingheng.he@outlook.com

Ying Wang
School of Data Science and Applications
Inner Mongolia University of Technology
Hohhot, China
20131800794@imut.edu.com

Yunpeng Gao
School of Data Science and Applications
Inner Mongolia University of Technology
Hohhot, China
1756583539@qq.com

2024 IEEE 5th International Conference on Pattern Recognition and Machine Learning (PRML) | 979-8-3503-5592-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/PRML62565.2024.10779819

***Abstract*—In Natural Language Processing (NLP) tasks, the target domain data is usually limited and the quality is poor. Text data augmentation is one of the effective methods to solve the problem of insufficient sample size in many NLP tasks. At present, on the one hand, the text data augmentation method is difficult to achieve the diversity and authenticity of the generated data text structure, on the other hand, it is also a challenge for the screening and elimination of low-quality samples in the generated data. Therefore, this paper proposes a scene-aware template of chinese text data augmentation method (Scene-Aware Prompt Template, SAPT) based on large language model. The SAPT method first uses prompt learning to construct a Scene-Aware Prompt (SAP)template. Secondly, combined with the text understanding and generation ability of the large language model, the SAP is improved by analyzing the semantic and contextual features of the original data to generate scene prompt words. Then, the original data is re-expressed into multiple data samples with different text structures but similar semantics by using the perfect SAP and large language model. Further, the quality of generated data samples is evaluated by fusing Rouge score and cosine similarity score, and low-quality data samples are filtered. The final generated data will be used for training the target domain task model. The experimental results on the Chinese news dataset THUCNews show that the overall performance of the classification model trained by the data generated by the SAPTDA method proposed in this paper reaches 89.89 %, which is 2.79 % higher than the baseline model, and has superior performance compared with the existing text data augmentation methods.**

***Keywords*—Natural Language Processing，Data Augmentation，Prompt Learning，Large Language Model**

I. INTRODUCTION

The results of Natural Language Processing (NLP) tasks depend largely on the quality and quantity of data used for training. However, due to the limitation of privacy protection or annotation cost, the data in the target domain is usually more scarce and of low quality. Therefore, in the target domain, the training data often cannot meet the needs of researchers, which is a common problem in the practice of the target domain. A natural and widely used strategy to alleviate such problems is to perform Data Augmentation (DA) [1]. How to use a small amount of raw data to amplify a large number of available samples through DA method is challenging.

Large Language Model (LLM)shows natural adaptability to various natural language processing tasks by training on diverse and complex data sets [2]. In addition, the original trained LLM introduces a large number of manual annotation samples in the fine-tuning stage, making the generated language more in line with human expression habits. Therefore, LLM can understand human text representation more efficiently and accurately, thus providing technical support for users in text generation and data augmentation. Prompt learning is a kind of learning method that changes the downstream task into a text generation task by adding ' prompt information ' to the input without changing or significantly changing the structure and parameters of the pre-training language model [3]. It is suggested that learning can make the model perform well in low-resource scenarios by transforming downstream tasks, and fill the boundary between pre-training and downstream tasks to a certain extent.

The outstanding performance of LLM and the excellent generalization of prompt learning make the generative data augmentation methods using large language models and prompt learning gradually accepted by the academic community.

Autonomous Region University Network Security and Education Management Information Engineering Research Center.

2022 The central government supports the local- ' Inner Mongolia Autonomous Region University Network Security and Education Management Information Engineering Research Center ' project.

However, the imperfection of the prompt template and the existence of low-quality samples in the generated data are still the main factors restricting the effectiveness of such methods.

II. Related Work

A. Data Augmentation

Traditional text-level data augmentation methods rely on direct operation of existing sample libraries. Jason Wei proposed a simple text classification data augmentation technology-EDA (Easy Data Augmentation), including the following four methods : synonym replacement (SR, using a synonym list to replace the words in the sentence), random exchange (RI, randomly exchanging two words of the sentence, changing the word order), random insertion (RS, randomly inserting synonyms of a word in the original sentence) and random deletion (RD, randomly deleting words in the sentence). These methods have shown improved performance in many text classification tasks [4]. Xie showed how to apply the supervised DA method to unsupervised data through the consistency training of (x, DA (x)) pairs[5]. Sosuke Kobayashi proposed a method to generate enhanced examples by replacing words with other words randomly selected based on the distribution of the current context 's recurrent language model[6].Steven Y. Feng proposed a task called semantic text exchange (STE), which involves adjusting the overall semantics of the text to adapt to the context of the inserted new words or phrases, that is, replacing entities (RE). They achieve this goal by using a system called SMERTI and a masked language model (LM) approach[7].

The oversampling method is the most intuitive method to increase the sample size. The commonly used oversampling methods include random oversampling and Synthetic Minority Over-Sampling Technique (SMOTE). The SMOTE method is an interpolation method that generates new synthetic samples between the sample and its neighboring samples. As a classical oversampling method, SMOTE has been improved and optimized by many researchers[8].

The generative method generates new samples by learning the distribution of data. The commonly used generative methods include variational Bayesian autoencoder[9], diffusion model generation adversarial network[10], large language model method, etc. Mehrotra et al. applied GAN to small sample learning and proposed a Generative Adversarial Residual Pairwise Network to solve the single sample learning problem. Seq2seq and language models have also been used for data augmentation [11]. Steven Y. Feng et al. proposed a task called semantic text exchange (STE), which involves adjusting the overall semantics of the text to adapt to the context of the inserted new words or phrases, called substitution entity (RE). This goal is achieved by using a system called SMERTI and the masked LM method, and the generated data samples can be directly used for data augmentation[12].Ateret Anaby-Tavoret al. learned the label condition generator by fine-tuning GPT-2 on the training data, and used it to generate candidate examples for each class. Then, the classifier trained on the original training set is used to select the first k candidate examples. The experimental results show that these examples do belong to the corresponding categories, which proves that the method can be used for data augmentation[13]. Haixing Dai et al. proposed a text data augmentation method based on ChatGPT (AugGPT) to solve the problem that the current text data augmentation methods either cannot ensure the correct labeling of the generated data, or cannot ensure sufficient diversity of the generated data, or the coexistence of the two. AugGPT reformulates each sentence in the training sample into multiple samples with similar concepts but different semantics[14]. Bohan Li et al. proposed a MIXPRO augmentation method, which enhances common input text and templates through token-level, sentence-level and era-level Mixup strategies[15].

In addition to the above general data augmentation methods, different data also have their own specific sample expansion methods. The idea of sampling method is simple and clear, but it fails to make full use of the information of data itself. Direct manipulation of the content of the original text may change the semantics of the original text because the synonyms have different meanings in different contexts or introduce too much noise ; using large language models and prompt learning to construct templates for data augmentation will not accurately measure the effectiveness of the generated data because of simple prompt templates and unreasonable evaluation indicators. Therefore, the limited templates and text used in the learning with a small amount of hints still leave a huge space for performance improvement.

B. Prompt Learning

The essence of prompt learning is to design a template that is highly compatible with the upstream pre-training task. Through the template design, the potential of the upstream pre-training model is fully tapped, so that the upstream pre-training model can better complete the downstream task without relying on the labeled data as much as possible. The key steps of hint learning include the following three points : 1) design the task of pre-training the language model, 2) design the style of the input template, and 3) design the label style and the mapping method of the model output to the label.

Prompt engineering is the process of creating a prompt template that aims to obtain optimal performance in downstream tasks. In many previous studies, this process involves human engineers or algorithms to find the best template for each task that the model needs to perform. The first step of the prompt engineering is to consider the form of the prompt, and then decide whether to use manual or automatic methods to create the required type of prompt. There are two main types of prompts : cloze prompts and prefix prompts. The cloze hint is used to fill in the gaps in the text string (e.g., ' this book is about [Z] '), while the prefix hint is used to continue the prefix of the string (e.g., ' what is this book about ? [Z] '). Which prompt form to choose will depend on the specific task and the model used to solve the task. In general, prefix hints are more beneficial for generating class tasks or tasks solved using a standard autoregressive language model, because they are consistent with the nature of the model generating text from left to right.

The most direct way to create a prompt template is to manually create an intuitive template based on human intuition. For example, Petroni et al. pioneered the creation of a manually created cloze template on the LAMA dataset to obtain knowledge in the language model[16]. Brown et al.created manually-made prefix prompts to handle various tasks,

including question answering, translation, and common sense reasoning detection tasks[17]. Schick et al. used predefined templates in a small number of learning settings for text classification and conditional text generation tasks[18].

Although the strategy of manually making templates is very intuitive and can indeed solve various tasks to a certain extent, this method also has drawbacks. First, creating and experimenting with these hints takes time and experience, especially for some complex tasks such as semantic parsing. Second, even experienced prompt designers may not be able to manually find the best prompts. In order to solve these problems, many methods have been proposed to automate the template design process. Jiang et al. proposed a mining-based Mine method that can automatically find templates with a given set of training inputs[19]. Wallace et al. performed a gradient-based search on the actual markers to find a short sequence that can trigger the underlying pre-trained LM to generate the required target prediction[20]. Gao et al. introduced the seq2seq pre-trained language model T5 into the template search process and used reinforcement learning to generate prompts that control the text generation process[21].

C. *Large Language Model*

Large-scale pre-trained language model is a mainstream language model developed on the basis of neural network language model. In recent years, large-scale pre-trained language models have developed rapidly[22]. Foreign products such as generative pre-trained transformer (GPT) series (GPT1, GPT2, GPT3[23], Chat GPT / Instruct GPT[24], GPT4 [25]) are rapidly iteratively updated, driving the rapid development of the entire industry while setting off a wave of research and development of large language models. In this context, March 2021. The super-large-scale intelligent model ' Wudao 1.0 ' was released. Among them, ' Wudao · Wenyuan ' is a pre-trained language model with Chinese as the core, and the model parameters reach 2.6 billion. 'Wudao · Wenhui ' is a super-large-scale pre-training model for cognition, and the number of model parameters reaches 11.3 billion. In June 2021, the super-large-scale intelligent model ' Wudao 2.0 ' was released, and the number of parameters reached 1.75 trillion. Subsequently, the ' Wen Xin Yi Yan ' model, the ' Tong Yi Qian Wen ' model, and the Tsinghua GLM series of large models have also emerged. These large language models have strong text understanding and generation capabilities [26].

Text comprehension and generation ability refers to the ability of language models to generate coherent and semantically reasonable text based on given context information. Taking GPT-3 as an example, it has a very large scale of 175 billion parameters and can generate high-quality text in various contexts, including but not limited to articles, dialogues, and codes. Through unsupervised learning, GPT-3 can automatically learn language structure, grammatical rules and semantic relations, so as to realize the wide application of diversified tasks.

At present, there is no universally accepted concept definition for large language models. N Carlini et al. pointed out that the large language model is composed of neural networks with a large number of parameters (usually more than billions), and uses self-supervised or semi-supervised learning to train on a large number of unlabeled texts. It has universal capabilities and can perform a wide range of natural language processing tasks, including text summarization, translation sentiment analysis, etc. [27]. Wayne Xin hao et al. pointed out in their review that large models refer to language models that are trained on massive text corpora and contain at least billions of level parameters, such as GPT-3[28], PaLM [29], LLaMA[30]], etc. These mainstream large-scale language models have made significant progress in generating capabilities. Through deep learning techniques and large-scale training data, they can accurately understand the context and generate text that conforms to semantic logic, providing strong support for various applications in the field of natural language processing.

III. METHOD

A. *Overall Framework*

In this chapter, we introduce two important components of our proposed scene-aware prompt template text data augmentation method based on large language model, including i) scene-aware prompt template design based on large language model ; ii) Low-quality sample generation filter. The overall framework of the SAPT (Scene-Aware Prompt Template) method is shown in Figure 1, and the algorithm of the SAPT method is shown in Algorithm 1.

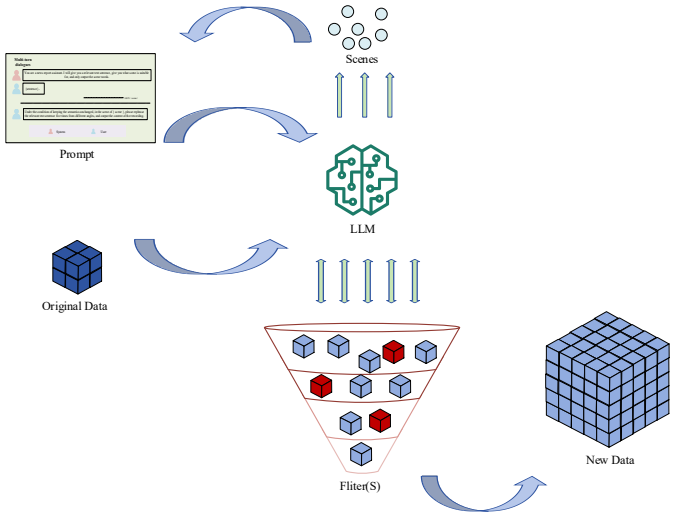

Fig. 1. The framework of SAPT

B. *Scene-Aware Prompt Template Design*

Devlin et al. ' s research shows that the general method of applying the pre-trained language model to the target domain is fine-tuning. However, in the target field, it usually faces the challenge of lack of data[31]. Therefore, how to enhance low-resource data has become a topic of concern for researchers. Dai et al. used prompt learning templates to solve the problem of lack of data in small sample learning scenarios[14], but using simple prompt learning templates cannot guarantee the quality of the generated samples. Our goal is to maintain the same authenticity of the samples generated by the prompt template as the source data, but the text structure is different and can learn new knowledge for the target model to adapt to downstream tasks. Therefore, this paper proposes a construction method of scene-aware prompt template based on large language model, which automatically adds scene information to the prompt

template, so that the large language model can better understand and process text data by using scene information, and generate sample data that meets the needs of researchers.

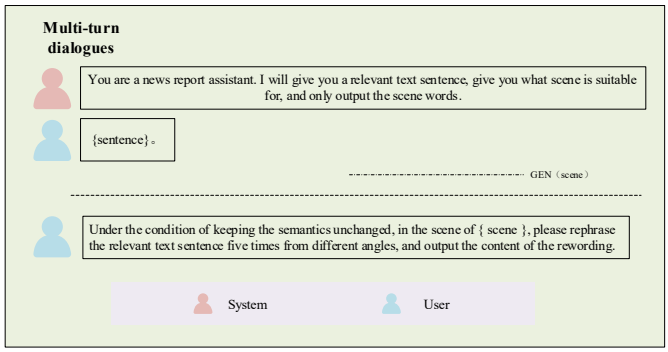

Fig. 2. Scene-Aware Prompt Template for multi-round dialogues

In order to clarify the purpose of constructing the prompt template, this paper designs a scene-aware prompt template for multiple rounds of dialogue. The hint template is shown in Figure 2. The THUCNews dataset uses a multi-round dialogue scene-aware prompt template for data augmentation. We first use the data $X_{\text{data}} = (x_1, x_2, x_3 ... x_n)$ in the original dataset D_{b} and the prompt template X_{prompt} to generate the scene prompt word X_{scene} ,where xi represents the vector representation of the ith word, and the input encoding process can be expressed as:

$$H = Encoder(X_{data}) \tag{1}$$

where, H is the representation of context information.

Based on the context information H of the data X_{data} , the probability distribution of the generated scene words is predicted to be :

$$P(scene \mid H, X_{prompt}) = Soft\max(W_0 [h_t; X_{prompt}] + b_0) \tag{2}$$

Here, *scene* represents the scene words to be generated, W_O and B_O are the parameters of the model, and h_t is a representation of the context information H .

Secondly, the original data X_{data} and the generated scene word X_{scene} are added to the prompt template X_{prompt} , and the updated prompt template X^*_{prompt} is obtained :

$$X^*_{prompt} = Concat(X_{promot}, X_{data}, X_{scene}) \tag{3}$$

In the process of updating the prompt template X_{prompt} , we retain the information of the original data X_{data} and add additional scene words to enrich the context knowledge, instead of completely replacing the original data. The updated prompt template X^*_{prompt} will be used as the model input X_{input} .

Then Xinput is input into the large language model, and a new data sample D_{x} is generated using the large language model :

$$D_x = Gen(LLM, X_{input} \mid X^*_{prompt}) \tag{4}$$

This semi-automatic template construction method simplifies the complexity of the input templates of different target domains and reduces the amount of data that the model needs to process. The correspondence from X_{data} to X_{input} is completed by using the template X_{prompt} , which realizes the mapping from the original data to the model input, so as to ensure the accurate transmission and effective use of information. In the template, we set the placement position of each original data X_{data} in the prompt template, as well as the generation and use position of the scene prompt word X_{scene} . The scene prompt word X_{scene} as a class label provides domain information for the template. By using the domain information provided by the scene prompt word, the model is guided to focus on the information related to the specific scene, so that the model can generate text content that is more in line with the specific scene and context, thereby improving the quality of the generated samples.

C. *Low Quality Sample Generation Filter*

After generating data D_{x} , we find that there are low-quality (meaningless) samples in D_{x} . In order to filter low-quality generated data samples and further improve the quality of generated data, we design a low-quality generated sample filter (S) based on text structure and semantics to filter and eliminate meaningless generated data samples. As shown in Figure 3.

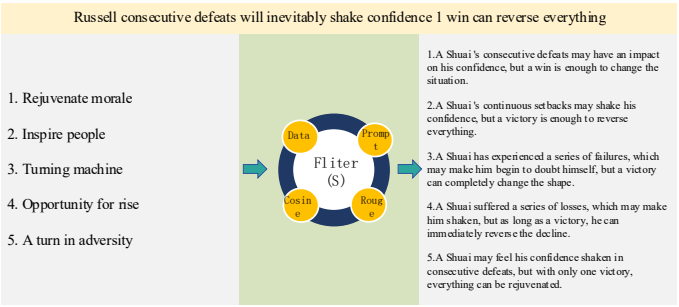

Fig. 3. Filter (S) filter meaningless low quality data example

LLM usually uses standard metrics or indicators and evaluation tools to evaluate the performance of the model, such as BLEU score, ROUGE-N index score and BERT score that can measure accuracy. For example, ROUGE score is used to quantify the similarity and quality between the text generated by the language model and the reference text. At present, most of the existing evaluation work adopts this evaluation protocol because of its subjectivity, automatic calculation and simplicity. Different from previous studies, the filter S proposed in this paper combines Rouge-N score and cosine similarity score to evaluate the text structure consistency (i.e., the overlap between the generated data samples and the original data in terms of vocabulary expression) and semantic consistency (i.e., whether the generated data samples are consistent with the original data in terms of semantics) of each generated sample D_{x} .

1) Text structure consistency assessment: Rouge-N mainly counts the recall rate on N-gram. The similarity between the

generated sample and the original data is concerned from the literal overlap :

$$Rouge-N=\frac{\sum_{gram_n\in S} Count_{match}(gram_N)}{\sum_{gram_n\in S} Count_{match}(gram_N)} \tag{5}$$

When N = 1, the matching number of single words in the generated data sample and the original data sample is counted. When N = 2, the matching number of binary phrases in the generated data samples and the original data samples is counted.

Compared with the matching of single words (N = 1), the selection of binary phrases with N = 2 can better capture the semantic information and the relevance between words. The semantics of Chinese texts are often expressed by phrases or sentences composed of multiple words. Therefore, combining words into binary phrases can more accurately reflect the semantic content of the text. When more common subsequences are selected (e.g., N > 2), although the range of matching can be further expanded, more noise and ambiguity will be introduced. In Chinese text, longer subsequences will lead to more word order changes and unnecessary matching, thus reducing the accuracy of similarity evaluation. In addition, with the increase of N, the computational complexity will also increase, which will also lead to performance degradation.

Therefore, the ROUGE-2 index with N = 2 is selected in filter S :

$$Rouge-2=\frac{\sum_{gram_2\in S} Count_{match}(gram_2)}{\sum_{gram_2\in S} Count_{match}(gram_2)} \tag{6}$$

Compared with other ROUGE-N indicators, it can better balance the accuracy and computational efficiency of similarity. By capturing the matching number of binary phrases, we can more comprehensively evaluate the similarity and structural consistency of texts and reduce the impact of word order changes. The lower ROUGE-2 index score indicates that the generated samples use different expressions than the original data, and the generated samples have obvious differences compared with the original data text structure, so as to screen out low-quality generated data samples and improve the quality and availability of generated data.

Semantic consistency assessment: Some of the most common semantic similarity measures include Euclidean distance, cosine similarity and dot product similarity. Cosine similarity performs better when dealing with high-dimensional sparse data. In natural language processing, text is often represented as a high-dimensional sparse vector, where each dimension corresponds to a word or a feature of a word. In this case, cosine similarity can better measure the directional similarity between texts without being affected by the number of dimensions. In contrast, the performance of Euclidean distance in high-dimensional space will be affected by the number of dimensions, while the dot product similarity emphasizes the length difference between vectors, which may not accurately reflect the semantic similarity. Cosine similarity can better deal with the length change and relative proportion of the text. In Chinese text, the length of different sentences may vary greatly, but there may still be semantic similarities between them. Cosine similarity can better deal with this situation by considering the angle between vectors rather than the length, so as to capture semantic similarity more accurately. In addition, the calculation of cosine similarity is simple and efficient, and does not require vector normalization or additional distance conversion, so it is easier to implement and calculate in practical applications.

Therefore, in the filter S, we choose Cosine Similarity to capture the distance relationship in the potential space :

$$\mathrm{Cos}(\theta)=\frac{A\cdot B}{\|A\|\times\|B\|}=\frac{\sum_{i=1}^{n} A_i\times B_i}{\sqrt{\sum_{i-1}^{n}(A_i)^2}\times\sqrt{\sum_{i-1}^{n}(B_i)^2}} \tag{7}$$

where *A* and *B* denote the two embedding vectors of comparison, respectively. The cosine similarity measures the cosine value of the angle between two vectors and is limited by the range between 0 and 1.

Algorithm 1 SAPT algorithm

Input：
Db: The original data set, Xprompt：Prompt template
Output:
Dn: New data set after data augmentation
Definition：
Xdata： The original data, Xscene： Scene prompt words, Tr： Text structure consistency threshold, Tc：Semantic consistency threshold. The initial values of Rouge and Cos (θ) are 0. Dx：Data generated by Xdata
Procedure：

```
1、 For Xdata in Db do
2、   Xscene ← GEN（LLM , Xdata , Xprompt）
3、   Xprompt ← UPDATE（Xprompt , Xscene）
4、   Dx ← GEN（LLM , Xprompt）
5、   For di in Dx do
6、     Rougei = Rouge（Xdata , di）
7、     Cos(θ)i = Cosine（Xdata , di）
8、     Rouge += Rougei
9、     Cos(θ) += Cos(θ)i
10、   End for
11、  AveRouge = Rouge / length.(Dx)
12、  AveCos(θ) = Cos(θ) / length.(Dx)
13、  if（AveRouge < Tr && AveCos(θ) > Tc）
14、    Dn.append(Dx)
15、  else
16、    to step 4
17、End for
```

For each generated sample D_x, we will use the designed filter (S) to evaluate its quality. Filter (S) calculates the Rouge-2 score and cosine similarity score of D_x and the original data, and compares them with their respective thresholds. The retained samples need to meet the following two conditions at the same time : i) The Rouge-2 score is lower than the text structure consistency threshold : it indicates that the generated samples and the original data have a low degree of matching in two consecutive phrases, that is, there are differences in the text structure. ii) The cosine similarity score is higher than the semantic consistency threshold : it indicates that the generated sample has a high semantic similarity with the original data, that is, it is semantically consistent. Otherwise, we will use the same scene-aware template to regenerate the data samples. That is :

$$D_x = \begin{cases} Gen_{D_x}(LLM, X_{input} \mid X_{prompt}, X_{data}, X_{scene}) & r(D_x) < T_r \wedge c(D_x) > T_c \\ \mathrm{Re}\, Gen_{D_x} & others \end{cases} \quad (8)$$

D_n represents the generated sample set that meets the conditions, D_x represents the new sample generated by each original data, $r(D_x)$ represents the Rouge-2 score of sample D_i, $c(D_x)$ represents the Rouge-2 score of sample D_i, T_r represents the text structure consistency threshold, which is set to 0.30 in this paper, T_c represents the semantic consistency threshold, which is set to 0.90 in this paper.

IV. Experiments

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

THUCNews DataSet		
	financia	ING says it will buy British savers' Icelandic deposits. ING, the Netherlands' largest financial services group, said yesterday... British savers have recently continued to transfer their deposits into seemingly... "reliable" major Icelandic banks.
	household	Land supply for protected housing up 36.3 per cent in H1 The reporter learnt from the Ministry of Land and Resources... Affordable housing land supply increased... A year-on-year decline of 4.5 per cent. I want to comment
	education	Newsflash: Chinese student named UK International Student Star... Recently, Li Tieqiang, a Chinese student from Zhanjiang, Guangdong Province, was awarded the 2009 British Council... International students from different regions received regional awards.
	⋮	⋮
	technology	Digital China Takes Over Shenzhen Local Tax Core Levy Management System... By Han Ping: China's largest ... Digital China will increase its market expansion in Shenzhen.
	politics	U.S. warship visiting Poland opens fire on Polish coast 3 times (photo) By Liang Xu Globe Reporter The U.S. Aegis... The USS Ramech conducted a joint military exercise with the Polish Navy with a total of 250 people on board, including five women.
	⋮	⋮
	society	Poll: What are your suggestions for changes to the Consumer Protection Act. The Consumer Protection Law, which has been in force for 15 years, will see its first... Opinion China (www.minyi.net.cn) Published
	recreation	Serenity reveals that she once gave up her green card: I'm stubborn!... Serenity did go to live in the US for a long time... The habits are also Eastern, she should be able to adapt." ...
	sport	Hailey Fung senior: Wang Shouye was taken away is not a surprise, the football circle is so chaotic... When asked about the situation of Wang Shouye, Zhao Da Reel always said that he did not know... (Sina Sports)

Fig. 4. Some candidate classification categories and examples from the THUCNews dataset

A. Dataset Preprocessing

1) THUCNews Dataset: This paper uses the THUCNews dataset for experiments. The THUCNews dataset is a large-scale public dataset for Chinese text classification tasks. The dataset is generated by Tsinghua University based on the historical data of Sina News RSS subscription channel from 2005 to 2011. It contains 740,000 news documents (2.19 GB), all in UTF-8 plain text format, including 14 candidate classification categories, such as finance, lottery, real estate, etc. The specific details of some candidate classification categories are shown in Figure 4.

TABLE I. Details of THUCNews Data Subset After Processing

category	example
sport	Raptors will land in Europe to sign La Liga giants. He was on the bench in Houston.
financia	ING said it would buy British depositors Icelandic deposits
education	Art examination is not easy to test the professional course culture class should pay attention to
technology	China Mobile radio and television cooperation further : CMMB into the TD data card
politics	The Ministry of Health launched the first batch of 59 public demonstration sites for hospital affairs.
stock	U.S. stocks were first low and then slightly higher. Hong Kong market struggled for stability.
game	The selection of commentators for Jianwang 2 series Wulin Competition will be launched soon
constellation	The year of Scorpio, the lucky prophet (picture)
household	After the summer promotion hot discount in the solid wood furniture market, there are frequent problems in the products.
society	The man was mistakenly arrested twice because of the same identity card number as the fugitive.
lottery	Sina expert zhibo double chromosphere 09007 : mantissa and value recommendation 27-33
recreation	Zhou Huimin promoted Ni Tai to work actively in the New Year to do good public welfare (picture)
fashion	Fashion Street Photos Japan MM with a clever trick
realestate	The four years of soaring house prices in Shenzhen have attracted the attention of private real estate developers.

2) Data Pre-processing: The THUCNews dataset is re-integrated and divided manually on the basis of the original Sina news classification system. We extracted a subset of 500 samples from each news category and processed them as our initial data set. Table 1 shows the specific details of the processed data set. The THUCNews dataset is reintegrated and divided manually on the basis of the original Sina news classification system. We extracted a subset of 500 samples from each news category and processed them as our initial data set. Table I shows the specific details of the processed dataset.

B. Experiment Results

1) Classification Experiment: This section verifies whether the scene-aware template text data augmentation method based on large language model can effectively generate data samples. In the experiment, we use BERT as the basic model. Firstly, we extract part of the original data set as our test data set ; then, the model is trained on the basic data set and the generated data set to generate the classification model Basebert and SAPTbert ; finally, the performance of Basebert and SAPTbert is evaluated on the test set.The hardware configuration used in this chapter is GPU : 12G NVIDIA GeForce RTX 3060, CPU : 16G 12th Gen Intel (R) Core (TM) i7-12700 2.10GHz. The experimental tools and software environment are python-3.10.9, pytorch-2.0.1, Cuda-11.8, and the large language model is Chat GPT3.5. In the experiment, the output dimension of the basic Bert model is set to 768, the dropout is set to 0.5, the maximum sequence length is set to 35, and the learning rate is set to 4e-5. The experimental results are shown in Table II.

TABLE II. NEWS REPORT CLASSIFICATION RESULTS

category	*Basebert*			*SAPTbert*		
	R	***P***	***F1***	***R***	***P***	***F1***
financial	86.29%	89.61%	87.92%	91.14%	89.86%	90.50%
lottery	96.89%	92.37%	94.56%	96.29%	94.66%	95.47%
realestate	93.71%	75.06%	83.35%	92.29%	88.25%	90.22%
stock	67.43%	90.42%	77.25%	86.57%	95.28%	90.72%
household	86.29%	90.69%	88.43%	82.00%	92.28%	86.84%
education	88.00%	94.77%	91.26%	94.29%	93.48%	93.88%
technology	83.14%	76.38%	79.73%	90.29%	88.76%	89.52%
society	80.86%	94.35%	87.08%	88.86%	87.61%	88.23%
fashion	88.00%	95.06%	90.99%	90.29%	86.58%	88.39%
politics	84.86%	85.59%	85.22%	82.29%	90.57%	86.23%
sport	89.14%	93.69%	91.50%	88.00%	95.06%	91.39%
constellation	99.71%	96.68%	98.17%	99.43%	93.30%	96.27%
game	92.57%	72.79%	81.82%	89.14%	86.43%	87.76%
recreation	82.29%	82.05%	82.17%	85.43%	80.81%	83.06%

Table II shows in detail the performance of the two models for classifying different news stories, and the experimental results show that the classification model trained using the generated data has a significant improvement in a number of classifications such as Stocks (+13.4%), Technology (+9.79%), Games (+5.94%), and only has a slight impact on classifying news stories on Home, Fashion, Sports, and Horoscopes, and comparing it to Basebert, the SAPTbert shows a 2.79% improvement in overall performance compared to Basebert. These results show that text data augmentation using the SAPT method can effectively improve the performance of deep learning models in applications, validating the effectiveness of the present data augmentation method.

2) Contrast Experiment: After verifying the effectiveness of the proposed SAPT method, this paper conducts experiments on some existing data augmentation methods on the THUCNews dataset. The experimental results are shown in Table III.

Compared with the existing text data augmentation methods, although the samples generated by the traditional data augmentation methods BackTranslationAug and SwapWordAug are semantically consistent with the initial data, they also have a high degree of overlap in the text structure. The SAPT method generates samples without changing the data semantics and reduces the overlap of the text structure. The overall classification performance of the model is 0.79 % and 1.14 % higher than that of the BackTranslationAug and SwapWordAug methods, respectively. The SAPT proposed in this paper and the AugGPT proposed by Dai et al. are based on the idea of prompt learning. However, SAPT uses prompt learning to construct a scene-aware prompt template. Adding scene prompt words in the template can ensure that the large language model captures and reflects the scene features of the original data when generating samples, and improves the semantic consistency between the generated samples and the original data, so that the enhanced text data is more in line with the text samples required by the original data scene. The low-quality generated sample filter is composed of rouge score and cosine similarity, which can quantitatively measure the text structure consistency and semantic consistency between the generated sample and the original data, remove the low-quality generated sample, and improve the effect of text data augmentation. The overall classification performance of the model is 1.29 % higher than that of AugGPT.

TABLE III. COMPARATIVE EXPERIMENTAL RESULTS OF DIFFERENT DATA AUGMENTATION METHODS

Data Augmentation	***R***	***P***	***F1***
Raw	87.08%	87.83%	87.10%
SwapWordAug	88.82%	89.17%	88.75%
BackTranslationAug	89.04%	89.46%	89.10%
AugGPT	88.61%	88.86%	88.60%
SAPT(Ours)	89.74%	90.33%	89.89%

*3) Ablation Experiment:*This section conducts ablation experiments on the scene-aware prompt template design part and the low-quality generated sample filter part of the SAPTDA method mentioned in Chapter 3. The experimental results are shown in Tables IV and V.

TABLE IV. ABLATION EXPERIMENTAL RESULTS OF EACH PART

Number	**Scene**	**Fliter**	***R***	***P***	***F1***
1			87.08%	87.83%	87.10%
2	√		88.67%	89.09%	88.67%
3		√	88.95%	89.17%	88.96%
4	√	√	89.74%	90.33%	89.89%

Table IV shows the influence of the two parts of the method on the overall classification performance of the model. The results show that : 1 When only the scene-aware prompt template part is added to the prompt template, R increases by 1.59 %, P increases by 1.26 %, and F1 value increases by 1.57 %. This shows that the introduction of scene awareness in the prompt template can make the generated data samples maintain the same scene characteristics as the original data, and improve the semantic consistency between the generated data samples and the original data. When only the filter part is added to the prompt template, R increases by 1.87 %, P increases by 1.34 %, and F1 increases by 1.86 %. This shows that the filter proposed in this paper can screen out the low-quality samples in the generated data, while ensuring the authenticity of the generated data and improving the text structure diversity of the generated data samples.When adding the scene-aware prompt template part and the filter part at the same time, R, P and F1 are significantly improved (+ 2.66 %, + 2.50 %, + 2.79 %), and

compared with only adding the scene-aware prompt template part or the filter part, there are still improvements (+ 1.07 %, + 1.24 %, + 1.22 %), (+ 0.79 %, + 1.16 %, + 0.93 %), which indicates that the combination of the two parts produces synergies and improves the overall quality of the generated data and the classification performance of the model.

TABLE V. ABLATION EXPERIMENTAL RESULTS OF DIFFERENT NEWS REPORT TYPES

category	*base*			*Prompt + scene*			*Prompt + S*			*Prompt + scene + S(ours)*		
	R	***P***	***F1***	***R***	***P***	***F1***	***R***	***P***	***F1***	***R***	***P***	***F1***
financial	86.29%	89.61%	87.92%	84.00%	91.02%	87.37%	87.71%	90.56%	89.11%	91.14%	89.86%	90.50%
lottery	96.89%	92.37%	94.56%	95.14%	97.37%	96.24%	97.43%	96.06%	96.74%	96.29%	94.66%	95.47%
realestate	93.71%	75.06%	83.35%	81.43%	93.44%	87.02%	86.00%	88.53%	87.25%	92.29%	88.25%	90.22%
stock	67.43%	90.42%	77.25%	93.43%	84.06%	88.50%	90.00%	84.22%	87.02%	86.57%	95.28%	90.72%
household	86.29%	90.69%	88.43%	85.71%	84.03%	84.87%	82.57%	87.05%	84.75%	82.00%	92.28%	86.84%
education	88.00%	94.77%	91.26%	94.86%	90.96%	92.87%	92.29%	94.44%	93.35%	94.29%	93.48%	93.88%
technology	83.14%	76.38%	79.73%	95.14%	75.00%	83.88%	93.43%	80.15%	86.28%	90.29%	88.76%	89.52%
society	80.86%	94.35%	87.08%	88.29%	86.55%	87.41%	82.29%	89.16%	85.59%	88.86%	87.61%	88.23%
fashion	88.00%	95.06%	90.99%	83.43%	92.41%	87.69%	86.57%	84.87%	85.71%	90.29%	86.58%	88.39%
politics	84.86%	85.59%	85.22%	82.86%	88.41%	85.55%	82.29%	93.20%	87.41%	82.29%	90.57%	86.23%
sport	89.14%	93.69%	91.50%	96.00%	91.30%	93.59%	95.14%	90.49%	92.76%	88.00%	95.06%	91.39%
constellation	99.71%	96.68%	98.17%	98.57%	96.37%	97.46%	98.29%	96.36%	97.31%	99.43%	93.30%	96.27%
game	92.57%	72.79%	81.82%	77.71%	93.47%	84.87%	84.57%	93.08%	88.62%	89.14%	86.43%	87.76%
recreation	82.29%	82.05%	82.17%	84.86%	83.19%	84.02%	86.86%	80.42%	83.52%	85.43%	80.81%	83.06%

Table V shows in detail the influence of the two-part generated data samples on the model 's recognition of different news reports. It can be seen that : 1 When only the scene-aware prompt template part is added to the prompt template, the classification ability of 10 kinds of news reports has been improved. The five news reports of lottery (+ 3.87 %), realestate (+ 3.67 %), stock (+ 11.25 %), technology (+ 4.15 %) and game (+ 3.05 %) reached the highest value of the ablation experiment. When only the filter part is added to the prompt template, the financial news report is added to the improved news report classification, which has a slight impact on the social news. When adding the scene perception prompt template part and the filter part at the same time, the classification ability of 10 kinds of news reports has been improved, including 1 peak stock (+ 13.4 %), technology (+ 9.79 %), game (+ 5.94 %), realestate (+ 6.87 %), and the classification of financial (+ 2.58 %) and education (+ 2.62 %) news reports has also been significantly improved. This shows that the data samples generated by adding the scene-aware prompt template part and the filter part at the same time can significantly improve the classification ability of the model to news reports, and only have a slight impact on the classification of some news reports, which is in line with the expected effect.

V. CONCLUSION

In this paper, aiming at the problem that natural language processing tasks often lack target domain data sets, we propose a scene-aware template of chinese text data augmentation method based on large language model (SAPT). SAPT uses semi-automatic prompt template to enhance text data. Different from other methods, we introduce scene prompt words into the prompt template, so that the large language model captures and reflects the scene features of the original data when generating samples, and improves the semantic consistency between the generated samples and the original data. In addition, the proposed low-quality generated sample filter filters low-quality samples in the generated data by fusing Rouge score and cosine similarity score as indicators to evaluate the quality of generated samples. Experimental results show that SAPT shows excellent performance compared with most existing text data augmentation methods.

In the future work, we will further study how to optimize the scene perception template to adapt to various downstream tasks in different target domains, so as to improve the generalization of the template and build a more realistic and effective target domain dataset. Future research will also focus on improving the template generation process, exploring innovative technologies to capture subtle contextual clues, optimizing the filter S to improve the efficiency of the SAPT method, and enhancing the authenticity and diversity of the generated data samples.

REFERENCES

[1] L. Bonhan, H. Yutai, C. Wangxiang, Data Augmentation Approaches in Natural Language Processing: A Survey[J]. 2021.DOI:10.48550/arXiv.2110.01852.

[2] W. Yaozu, L. Qing, D. Zhangjie, X. Yue, Current status and trends in large language modeling research[J]. Chinese Journal of Engineering. DOI: 10.13374/j.issn2095-9389.2023.10.09.003

[3] L. Pengfei, Yuan. Weizhe, Fu. Jinlan, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Comput Surv, 55(9): 195

[4] J. Wei and K. Zou, 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

[5] X. Qizhe, D. Zhihang, H. Eduard, T. Luong,and L. Quoc, 2020. Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems, 33.

[6] S. Kobayashi, 2018. Contextual augmentation:Data augmentation by words with paradigmatic relations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 452–457,New Orleans, Louisiana. Association for Computational Linguistics.

[7] F. Steven Y, L. Aaron W, and H. Jesse, 2019.Keep calm and switch on! Preserving sentiment and fluency in semantic text exchange. In Proceeding of the 2019 Conference on Empirical Methods inNatural Language Processing and the 9th Interna-tional Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2701–2711, Hong Kong, China. Association for Computational Linguistics.

[8] N. Chawla, K. Bowyer, L. Hall,et al.SMOTE: Synthetic Minority Over-sampling Technique[J].Journal of Artificial Intelligence Research, 2002, 16(1):321-357.DOI:10.1613/jair.953.

[9] T. Yuhta, S. Takashi, L. WeiHsiang, et al.SQ-VAE: Variational Bayes on Discrete Representation with Self-annealed Stochastic Quantization[J]. 2022.DOI:10.48550/arXiv.2205.07547.

[10] N. Hojjat, M. Parisa Fard, N. Mohammad, et al.Generative Adversarial Networks (GANs) in Networking: A Comprehensive Survey & Evaluation[J]. 2021.DOI:10.1016/j.comnet.2021.108149.

[11] M. Akshay, D. Ambedkar, Generative Adversarial Residual Pairwise Networks for One Shot Learning[J]. 2017.DOI:10.48550/arXiv.1703.08033.

[12] F. Steven Y, L. Aaron W, and H. Jesse, 2019.Keep calm and switch on! Preserving sentiment and fluency in semantic text exchange. In Proceeding of the 2019 Conference on Empirical Methods inNatural Language Processing and the 9th Interna-tional Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2701–2711, Hong Kong, China. Association for Computational Linguistics.

[13] A. Anaby-Tavor, B. Carmeli, E. Goldbraich,A. Kantor, G.e Kour, S. Shlomov, N. Tepper, and N. Zwerdling. 2020. Do not have enough data? Deep learning to the rescue! In Proceedings ofAAAI, pages 7383–7390.

[14] D. Haixing, L. Zhengliang , L Wenxiong , et al. Chataug: Leveraging chatgpt for text data augmentation[J]. arxiv preprint arxiv:2302.13007, 2023.

[15] L. Bohan, D. Longxu., H. Yutai, F. Yunlong, M. Honglin, & C. Wanxian ,(2023). MixPro: Simple yet Effective Data Augmentation for Prompt-based Learning. ArXiv, abs/2304.09402.

[16] F. Petroni, T. Rockt¨aschel, S. Riedel, P. Lewis, A. Bakhtin, W. Yuxiang, and M. Alexander. 2019. Language models as knowledge bases? In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

[17] T. Brown, B. Mann, N. Ryder,et al.Language Models are Few-Shot Learners[J]. 2020.DOI:10.48550/arXiv.2005.14165.

[18] T. Schick, H Schütze, It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners[J]. 2020.DOI:10.48550/arXiv.2009.07118.

[19] Z. Jiang, X. Frank F, J. Araki, and G. Neubig, 2020c. How can we know what language models know? Transactions ofthe Association for Computational Linguistics, 8:423–438.

[20] W. Eric, F. Shi, K. Nikhill, G. Matt, and S. Sameer, 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

[21] W. Xiaozhi , G. Tianyu , Z. Zhaocheng ,et al.KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation[J].Transactions of the Association for Computational Linguistics, 2021.DOI:10.1162/TACL_A_00360.

[22] L. Tianyang, W. Yuxin, L. Xiaoyang, etal.A survey of transformers [J].AI Open ,2022,3:111 132.

[23] T. Brown, B. Mann, N.Ryder, M.Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. Mc Candlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” in Ad vances in Neural Information Processing Systems 33: An nual Conference on Neural Information Processing Sys tems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[24] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wain wright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, “Training language models to follow instructions with human feedback,” CoRR, vol. abs/2203.02155, 2022.

[25] OpenAI, “Gpt-4 technical report,” OpenAI, 2023.

[26] Z. Du, Y. Qian, X. Liu,et al.GLM: General Language Model Pretraining with Autoregressive Blank Infilling[J]. 2021.DOI:10.18653/v1/2022.acl-long.26.

[27] N. Carlini, F. Tramer, E. Wallace, et al. Extracting training data from large language models[C //30th USENIX Security Symposium (USENIX Security 21). 2021: 2633-2650.

[28] X. Wayne Xin, Z. Kun, L. Junyi, et al. A survey of large language models[OL . arXiv preprint arXiv:2303.18223, 2023.

[29] A. Chowdhery, S. Narang, J. Devlin,et al.PaLM: Scaling Language Modeling with Pathways[J]. 2022.DOI:10.48550/arXiv.2204.02311.

[30] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozi` ere, N. Goyal, E. Ham bro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “Llama: Open and efficient foundation language models,” CoRR, 2023.

[31] J. Devlin, C.Ming-Wei, L.Kenton , and T.Kristina 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171‑4186, Minneapolis, Minnesota. Association for Computational Linguistics.