Multi-angle Prediction Based on Prompt Learning for Text Classification

Zhengyu Ju[1], Zhao Li[1,2(✉)], Shiwei Wu[3], Xiuhao Zhao[3], and Yiming Zhan[3]

[1] Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[2] Shandong Computer Science Center (National Supercomputer Center in Jinan), Jinan, China
liz@sdas.org
[3] Evay Info, Jinan, China
{wushw,zhanym}@sdas.org

Abstract. The assessment of Chinese essays with respect to text coherence using deep learning has been relatively understudied due to the lack of large-scale, high-quality discourse coherence evaluation data resources. Existing research predominantly focuses on characters, words, and sentences, neglecting automatic evaluation of Chinese essays based on articles' coherence. This paper aims to research automatic evaluation of Chinese essays based on articles' coherence by leveraging some data from LEssay, a Chinese essay coherence evaluation dataset jointly constructed by the CubeNLP laboratory of East China Normal University and Microsoft. The coherence of Chinese essays is primarily evaluated based on two big aspects: 1. The smoothness of logic (the appropriateness of using related words, and the appropriateness of logical relationship between contexts) 2. The reasonableness of sentence breaks (how well punctuation is used and how well the sentence is structured). Therefore, in this paper, we adopt prompt learning and cleverly design a multi-angle prediction prompt template that can realize the assessment of the coherence of Chinese essay from four angles. During the inference stage, the prediction of the coherence of Chinese essays is obtained through the multi-angle prediction template and voting mechanism. Notably, the proposed method demonstrates excellent results in the NLPCC2023 Shared-Task7 Track1.

Keywords: Prompt learning · Multi-angle prediction · Voting mechanism

1 Introduction

The difficulty of automatic coherence evaluation for Chinese essays using artificial intelligence (AI) is marked by a dearth of extensive labeled data specifically dedicated to evaluating the coherence of Chinese essays. Consequently, there exists a notable research gap in the classification of the coherence degree of Chinese essays within the AI-based Chinese essay assessment techniques.

F. Liu et al. (Eds.): NLPCC 2023, LNAI 14304, pp. 281–291, 2023.
https://doi.org/10.1007/978-3-031-44699-3_25

With rapid rise of large language models (LLMs) such as T5 [13] and GPT-3 [1], etc. Researchers have found that pre-trained language models (PLMs) yield remarkable outcomes in various downstream tasks encompassing text classification, question answering, and knowledge graph. Through in-depth research, it has been discerned that since LLMs are trained with a large amount of data in the pre-training stage, they have acquired rich knowledge [3,11].

For text classification problem, based on this finding, researchers have proposed the utilization of the fine-tuning approach to mine and make use of the knowledge of PLMs acquired in pre-training stage, wherein a classifier is appended to a PLM to enable the adaptation of the PLM to a downstream task. However, the effectiveness of fine-tuning in capturing the knowledge gained by PLMs during pre-training is limited in scenarios characterized by sparse training data, such as few-shot and zero-shot scenarios. Such limitations become particularly evident in real-world settings. For instance, when evaluating the coherence of Chinese essays, the practical challenge of the lack of discourse coherence evaluation resources has persistently hindered progress.

In light of the lack of massive labeled data, researchers have recently proposed the adoption of prompt learning to effectively mine and leverage the knowledge acquired during the pre-training phase of PLMs. Prompt learning has emerged as a promising approach in which the text classification problem can be converted into a cloze problem by using [MASK] token and prompt characters, when dealing with text classification task. The cloze problem format closely aligns with the pre-training task of PLMs, leading to enhanced stimulation of PLMs and obtaining the knowledge of PLMs acquired during pre-training better. The final prediction is achieved through the application of answer engineering which is another research content of prompt learning. For instance, in the case of predicting the coherence of an article, a prompt template is defined as follows: "<TEXT>这篇文章的连贯程度? <MASK>". Here, the <TEXT> placeholder is replaced by an article's text, resulting in a new input. It will be introduced into a PLM. Assuming the coherence category is labeled as "excellent coherence", the [MASK] token is most likely to be filled with words from the words set that represents "excellent coherence". The mapping from the words set representing categories to the corresponding classes is referred to as the verbalizer [6], serving as an effective mechanism to bridge the regression values of PLMs and the final prediction regression values which can identify which category the input belongs to directly.

In this paper, we aim to automate the classification of coherence of Chinese essays by leveraging prompt learning. A small amount of labeled data in LEssay is used in our study.

Merely employing the prompt template "<TEXT>这篇文章的连贯程度? <MASK>" and treating it as a conventional text classification problem fail to realize the particularity of assessing coherence of Chinese articles. Therefore, this simplistic approach is inadequate in achieving reliable coherence predictions.

The assessment of coherence of Chinese essays encompasses four crucial aspects: the use of related words, the logical relationship between contexts, the

use of punctuation, and the sentence structure. Based on such particularity, in this paper, we will design a prompt template that can predict the coherence of Chinese articles from the four angles respectively. By doing so, we move beyond the simplistic notion of treating coherence assessment as a mere text classification task, and we can obtain more comprehensive knowledge from PLMs.

In scenarios characterized by limited annotated data, such as the few-shot scenario, updating the parameters of randomly initialized model components becomes challenging. However, we contend that the extensive knowledge acquired by PLMs during the pre-training stage is often sufficient to address downstream task, without the need to introduce additional parameters or randomly initialized model components to help PLM make predictions. In this paper, we propose the utilization of prompt learning to mine and leverage the knowledge of PLMs. The primary objective of our proposed method is to construct a prompt template that serves as the vital link between PLMs and downstream task and enables effective and comprehensive mining of PLM knowledge through this template construction better. The construction of the prompt template constitutes the focal point of prompt learning research. Meanwhile, our proposed method does not introduce supplementary model components. We design a multi-angle prediction prompt template carefully to realize the comprehensive acquisition and utilization of the knowledge acquired by PLMs in the pre-training stage well. Furthermore, we craft the training and inference modes of the model in a thoughtful manner to ensure accurate predictions in coherence classification for essays.

In this paper, the procedure of our proposed method can be outlined as follows: (1) Given that Chinese articles are basically long text, it is feasible to gain the coherence of essays by analyzing a portion of the text, each Chinese essay's text is sliced and evenly segmented into two parts, effectively leveraging the semantic information within the text and augmenting the amount of trainable supervised data to a certain extent. (2) A multi-angle prediction prompt template is devised specifically, capable of generating coherence prediction from each assessment angle. (3) During the model training stage, the prompt template is integrated with the original input sequence. The wrapped input sequence is then introduced to a PLM, utilizing a mapping mechanism to obtain multiple [MASK] regression values, The loss values between these multiple regression values and the ground truth are strictly calculated, and model optimization is achieved by minimizing the sum of these loss values. (4) In the inference stage, multiple [MASK] regression values are obtained using the same approach employed during training. Subsequently, a voting mechanism is employed to facilitate the prediction of article coherence.

This paper makes several key contributions, which can be summarized as follows:

- Making use of the special properties of Chinese text, slice the article into two equal length text, which alleviates the problem of less supervised data available for training to a certain extent.
- Conduct prompt engineering for prompt learning, involving the design of a prompt template tailored for multi-angle prediction to facilitate comprehensive coherence assessment of Chinese essays.

– Devise a rigorous training mechanism that ensures strict model optimization, coupled with the astute utilization of a voting mechanism to realize the inference of the coherence of Chinese essays.
– In NLPCC2023 SharedTask7 Track1, the proposed method has demonstrated outstanding performance, substantiating the effectiveness of the proposed method.

2 Related Work

2.1 BERT

BERT [4], introduced by Google in 2018, is a pre-trained language model that employs a deep bidirectional Transformer [15] architecture as its core component. This model has attained state-of-the-art (SOTA) performance across various Natural Language Processing (NLP) tasks.

2.2 Prompt Learning

The advent of LLMs and the recognition that the extensive knowledge is acquired by PLMs during the pre-training phase have spurred a burgeoning development in prompt learning [9]. Prompt learning has emerged as a more effective approach than fine-tuning, particularly in few-shot scenario and zero-shot scenario. Promisingly, prompt learning has found application in a wide range of downstream tasks of NLP, including but not limited to Text Classification [14], Natural Language Understanding [10], Relationship Extraction [2,5]. Notably, this methodology has exhibited noteworthy predictive performance in these downstream tasks of NLP.

2.3 Prompt Engineering

Prompt engineering plays a pivotal role in prompt tuning. As evidenced in the preceding example, prompt characters and [MASK] token are used to construct a prompt template. We denoted the template as $f_{prompt}(\cdot)$. The original input sequence $x = (x_0, x_1, ...x_N)$ is integrated into the template to form a cloze input form $f_{prompt}(x)$, we can bridge the gap between PLMs and downstream tasks by this way.

Prompt engineering methods encompass two primary approaches: manual template design and automatic template construction. The former entails the expertise and experience of designers who possess specialized knowledge in the domain that the used dataset related to. Skillfully crafting a template through manual designing can yield excellent outcomes, particularly in zero-shot scenarios. On the other hand, automatic template construction methods, such as prompt mining [7], prompt paraphrasing, continuous prompt [12], mitigate the need for manual intervention. These approaches can train template using limited data to make accurate prediction.

3 Task Definition

The objective of this task is to construct a multi-angle prediction prompt template and use the template to comprehensively acquire and utilize the knowledge acquired by PLMs in the pre-training stage to accurately predict the coherence classification of Chinese essays in scenario of having limited annotated data available for model training and verification. The coherence classification results are categorized into three levels, namely 2 (excellent coherence), 1 (moderate coherence), and 0 (incoherence).

4 Method

This section will delineate the proposed approach, encompassing several key components: (1) supervised data preprocessing; (2) prompt characters design and template construction; (3) overview of model operation flow; (4) model training and inference.

4.1 Supervised Data Preprocessing Module

The raw input to the model is text which is a metadata of LEssay. Given that the majority of the input text is long text, it is observed that the coherence of an article can be assessed by analyzing a portion of the text, rather than having to consider the whole article. Therefore, we employ slicing operation, dividing the original input sequence $x = (x_1, x_2, ...x_N)$ into two input sequences $x_a = (x_1, ...x_{\frac{N}{2}})$,$x_b = (x_{\frac{N}{2}+1}, x_{\frac{N}{2}+2}, ...x_N)$ evenly. This approach partially mitigates the issue of having too little supervised data available for training.

4.2 Prompt Characters Design and Template Building Module

Prompt character design plays a pivotal role in the prompt engineering process. By carefully selecting prompt characters that are relevant to the specific task and dataset, we can elicit and utilize knowledge of PLMs better.

Given the unique nature of this task that we can evaluate the coherence of an essay from four specific assessment angles, we design four groups of prompt characters to represent four assessment angles, including "关联词使用恰当程度？", "上下文之间逻辑关系情况？", "标点符号使用情况？", "句子结构情况？", which is referred to as $T0, T1, T2, T3$.

A prompt template, designed to realize multi-angle prediction, has been devised cleverly by using four groups of prompt characters: "$< TEXT >< T0 >< MASK >< T1 >< MASK >< T2 >< MASK >< T3 >< MASK >$". This template enables the coherence prediction of Chinese essays base on four distinct assessment angles. Each <MASK> token will make prediction based on the coherence assessment angle in front of the <MASK> respectively.

4.3 Overview of Model Operation Flow

The PLM utilized in this study is denoted as M. After the slicing operation, the input sequence is divided into two parts. Let us assume that the input sequence after slicing is represented as $original\_input = (x_1, x_2, ..., x_n)$. As shown in Fig. 1, The model input sequence $original\_input$ is integrated with the multi-angle prediction prompt template defined in Sect. 4.2, yielding a new input sequence x_p, $x_p =< original\_input >< T0 >< MASK >< T1 >< MASK >< T2 >< MASK >< T3 >< MASK >$. The integrated input sequence x_p is then introduced into the pre-trained language model M. The regression value of MLM Head of M is obtained finally.

$$logits = M\left(x_p\right) \tag{1}$$

where $logits \in \mathbb{R}^{n \times vocab\_size}$, n is the max sequence length of input sequence. $vocab\_size$ is vocabulary length of M.

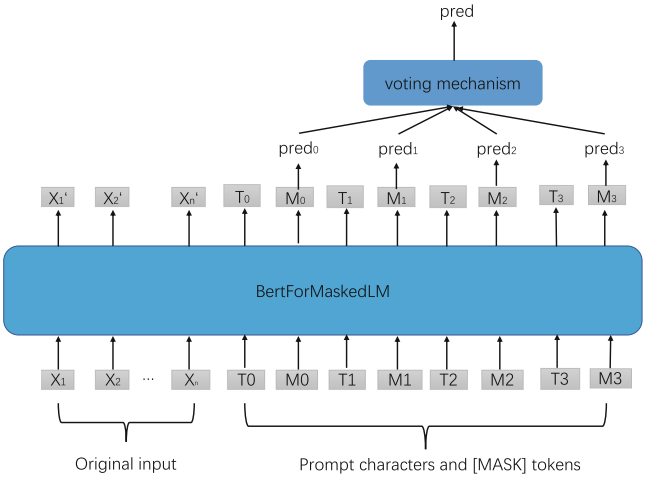

Fig. 1. Model architecture diagram

In order to realize the mapping between the PLM's vocabulary and the answer space, we design a mapping mechanism $f : \nu \longmapsto \gamma$, Where ν is the words set representing classes, γ is the set of classes. We define $\nu_2 = \{$"好"$\}$, $\nu_1 = \{$"一般"$\}$, $\nu_0 = \{$"差"$\}$, $\gamma = \{0, 1, 2\}$. $U_{y \in \gamma} \nu_y = \nu$. The probability that [MASK] is predicted as class y:

$$p\left(y \| x_p\right) = p\left([MASK] = a \| x_p\right), a \in \nu_y \tag{2}$$

The value of $p\left([MASK] = a \| x_p\right)$ is the $[tokenizer.encode(a)]$-th value of $logits$ in the [MASK], ν_k is the words set that represent class k.

For an input sequence, we can get four probability vectors because the template we designed has four [MASK] tokens, which is the reason why the template we design can realize multi-angle prediction.

4.4 Training and Inference

Training. Following the approach described in Sect. 4.3, the x of the sliced data {x: *original input*, *label*: y} is combined with the multi-angle prediction prompt template to generate the new input sequence x_p. Subsequently, x_p is introduced into M and the proposed mapping mechanism is employed to obtain the probability vector of [MASK] tokens respectively.

$$p_{pi} = [p_{pi}(y = 0|x_p), p_{pi}(y = 1|x_p), p_{pi}(y = 2|x_p)], i = 0, 1, 2, 3 \tag{3}$$

where p_{pi} represents the probability vector of i-th [MASK] when x_p is introduced into our proposed model, $p_{pi}(y = m|x_p)$ represents the probability that class m is predicted in the i-th [MASK] when x_p is introduced into our proposed model.

Hence, when the proposed model receives an input sequence x_p, it generates a vector at each [MASK] position, denoted as p_{p0}, p_{p1}, p_{p2}, p_{p3}, which represent the probability vectors based on four assessment angles respectively, then we can calculate the loss:

$$loss_0 = BCE\_Loss\,(p_{p0}, label) \tag{4}$$

$$loss_1 = BCE\_Loss\,(p_{p1}, label) \tag{5}$$

$$loss_2 = BCE\_Loss\,(p_{p2}, label) \tag{6}$$

$$loss_3 = BCE\_Loss\,(p_{p3}, label) \tag{7}$$

$$loss = loss_0 + loss_1 + loss_2 + loss_3 \tag{8}$$

Inference. For the raw input sequence to be predicted $x = (x_1, x_2, ...x_N)$, two input sequences can be obtained after slicing operation. We note sliced input sequence as $x_a = \left(x_1, \ldots x_{\frac{N}{2}}\right)$, $x_b = (x_{\frac{N}{2}+1}, x_{\frac{N}{2}+2}, ...x_N)$. Two input sequences are introduced into our proposed model respectively. And we will get $P_a = [p_{a0}, p_{a1}, p_{a2}, p_{a3}]^\top$, $P_b = [p_{b0}, p_{b1}, p_{b2}, p_{b3}]^\top$, P_m represents the combined tensor of the regression values of the sliced input sequence x_m at the four [MASK] positions, $P_m \in \mathbb{R}^{4\times3}$. p_{kj} represents the regression vector of the input sequence x_k at the j-th [MASK] token, $p_{kj} \in \mathbb{R}^{1\times3}$. As a result, For the original input sequence $x = (x_0, x_1, ...x_N)$, we can calculate the regression value.

$$P_x = [M_{x0}, M_{x1}, M_{x2}, M_{x3}]^\top = P_a + P_b \tag{9}$$

$$P_a + P_b = [p_{a0} + p_{b0}, p_{a1} + p_{b1}, p_{a2} + p_{b2}, p_{a3} + p_{b3}]^\top \tag{10}$$

where M_{xj} is the regression vector of input sequence x in j-th [MASK] token.

Compute the index of the maximum value of each [MASK] regression vector's elements, they are the prediction results of the coherence classification relative to the four coherence classification assessment angles respectively.

$$pred_0 = argmax\,(M_{x0}) \tag{11}$$

$$pred_1 = argmax\,(M_{x1}) \tag{12}$$

$$pred_2 = argmax\left(M_{x2}\right) \tag{13}$$

$$pred_3 = argmax\left(M_{x3}\right) \tag{14}$$

By employing a voting mechanism, the final prediction for the coherence level of a Chinese article is determined by selecting the index that was predicted the most times among the four assessment angles.

$$num_i = num\_of\left(pred_j == i\right), i = 0, 1, 2; j = 0, 1, 2, 3 \tag{15}$$

$$pred = argmax\left(num_i\right) \tag{16}$$

where i represents the index of class and $pred_j$ represents the predicted value of the j-th criterion. The final predicted value $pred$ is the index that the category was predicted the most times.

5 Experiment

In this section, we will demonstrate the effectiveness of the proposed method based on some of data from LEssay.

5.1 Dataset

To evaluate the effectiveness of the proposed method, we conducted experiments using some of data from the LEssay, which is utilized in the NLPCC2023 Shared-Task7 Track1. It is specifically designed for evaluating the coherence of Chinese articles. Our experimental data consists of 50 Chinese essays for training, 10 Chinese essays for validation, and 5,000 Chinese essays for testing. The coherence of the Chinese essays can be classified into three categories: 2 (excellent coherence), 1 (moderate coherence), and 0 (incoherence).

5.2 Baseline

Fine-Tuning. Fine-tuning adds a [CLS] token at the beginning of the original input sequence and then feeding it into a PLM. To predict the classification results, a classifier composed of a linear layer is added to the last layer of the PLM to predict the classification results.

P-Tuning. P-tuning [10] is an automatic method for constructing prompt templates to facilitate downstream task prediction. This approach employs custom prompt embedding to obtain preliminary prompts and uses MLP and LSTM to further process prompts, resulting in the final prompts for a template. P-tuning has demonstrated excellent performance in both few-shot and fully-supervised settings.

5.3 Implementation Details

For all our experiments, we employ $bert-base-chinese$ as our PLM. The epoch is 50. We use AdamW [8] as the model optimizer and the learning rate is set to 2e-5. Loss values are calculated by BCE loss function. During the training phase, we set the batch size to 2, while for verification and testing, the batch size is set to 8.

5.4 Main Results

During the testing phase, we employed precision (P), recall (R), and Macro-F1 (F1) to evaluate the effectiveness of our proposed model. The specific results of our proposed method and baseline are shown in Table 1.

Table 1. Experiment results on NLPCC2023 SharedTask7 Track1 dataset

Model	P	R	F1
Fine-tuning	34.78	34.48	29.15
P-tuning	34.12	33.36	33.36
Ours	**36.26**	**37.10**	**35.77**

In both Fine-tuning and P-tuning, randomly initialized parameters are introduced to assist in model prediction. In few-shot setting, due to the scarcity of adequate labeled data, it becomes challenging to optimize these randomly initialized parameters effectively. From Table 1, it can also be observed that the proposed method is better than Fine-tuning and P-tuning in terms of precision, recall and Macro-F1. Furthermore, the Macro-F1 is 6.62% higher than Fine-tuning and 2.41% higher than P-tuning, respectively.

5.5 Analyze

To further assess the effectiveness of our proposed method, we conducted a series of ablation experiments as follows: 1. Randomly initialize 4 prompt tokens to replace the prompt characters used in our proposed method, it means that we use $< TEXT >< T0 >< MASK >< T1 >< MASK >< T2 >< MASK >< T3 >< MASK >$ as prompt template, where $< T0 >, < T1 >, < T2 >, < T3 >$ are initialized randomly. 2. Remove four groups of prompt characters, which means that it uses $< TEXT >< MASK >< MASK >< MASK >< MASK >$ as prompt template. 3. The prompt characters still use four essay coherence evaluation criteria, but it only uses one $< MASK >$. Consequently, the prompt template used is $< TEXT >< T0 >< T1 >< T2 >< T3 >< MASK >$, indicating the absence of multi-angle prediction. 4. Use one $< MASK >$ and no prompt character is added, which means that it uses $< TEXT >< MASK >$ as prompt template. 5. The slicing operation is not used. The specific results are shown in Table 2.

Table 2. Ablation experiment results

Model	P	R	F1
Ours	**36.26**	37.10	**35.77**
Initialize the prompt tokens at random	35.44	**38.29**	33.77
Without prompt characters	33.70	34.29	24.27
With prompt characters and one [MASK]	33.15	32.92	28.47
Without prompt characters and with one [MASK]	36.20	36.78	34.56
Without slicing operation	35.13	35.44	33.71

In comparison to our proposed method, the ablation experiments involving random initialization of prompt tokens and the removal of prompt characters (ablation experiments 1 and 2) result in a reduction of 2% and 11.5% in Macro-F1, respectively. These experiments serve as the evidence of the fact that we cannot optimize randomly initialized parameters effectively in few-shot setting and the effectiveness of our proposed prompt characters design method, which utilizes the four angles of essay coherence evaluation criteria as prompt characters. In comparison to the proposed method, the ablation experiments involving the use of prompt characters but only one [MASK] token for prediction, and the use of only one [MASK] token without prompt characters (ablation experiments 3 and 4) result in a reduction of 7.3% and 1.21% in Macro-F1, respectively. These findings provide evidence for the effectiveness of our proposed multi-angle prediction template, which utilizes multiple [MASK] tokens to predict the results from multiple assessment angles, as well as the utilization of the voting mechanism. The ablation experiment involving the removal of the slicing operation (ablation experiment 5) results in a decrease of 2.06% in Macro-F1, compared to the approach using the proposed slicing operation as described in the paper. This observation provides further evidence for the effectiveness of our use of the slicing operation to alleviate the problem of lacking a substantial amount of supervised training data.

6 Conclusion

In this paper, we present a novel deep learning approach to address the task of coherence classification of Chinese essays. We propose a unique multi-angle prediction prompt template construction method. Our approach employs four distinct criteria of evaluating the coherence of Chinese essays to be prompt characters set. Additionally, we introduce a [MASK] token following each group of prompt characters, allowing for the prediction bases on the corresponding assessment angle. To ensure effective training, we establish rigorous mechanisms that require accurate predictions for each [MASK] token. During inference, we leverage a voting mechanism to obtain the final coherence classification prediction for Chinese essays. We also apply a slicing operation to all texts to mitigate

the challenge of limited training data. The proposed method achieves excellent results in NLPCC2023 SharedTask7 Track1, which proves the effectiveness of the proposed method. In theory, even when we scale with larger datasets, we can still employ the template used in the paper. By leveraging a substantial amount of supervised data to optimize prompt and other parameters better, we can achieve more accurate multi-angle predictions. When we need predict the coherence of other forms of text and if the text can also be evaluated from multiple angles, we can construct prompt templates following the method proposed in the paper and implement the final prediction using a voting mechanism. In theory, this approach can yield favorable results in the scenario.

Acknowledgments. This work was supported by Improvement of Innovation Ability of Small and Medium Sci-tech Enterprises Program (No. 2023TSGC0182) and Tai Shan Industry Leading Talent Project.

References

1. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
2. Chen, X., et al.: Adaprompt: adaptive prompt-based finetuning for relation extraction. arXiv preprint arXiv:2104.07650 (2021)
3. Davison, J., Feldman, J., Rush, A.M.: Commonsense knowledge mining from pretrained models. 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), pp. 1173–1178 (2019)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Han, X., Zhao, W., Ding, N., Liu, Z., Sun, M.: Ptr: prompt tuning with rules for text classification. AI Open **3**, 182–192 (2022)
6. Hu, S., et al.: Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. arXiv preprint arXiv:2108.02035 (2021)
7. Jiang, Z., Xu, F.F., Araki, J., Neubig, G.: How can we know what language models know? Trans. Assoc. Comput. Linguistics **8**, 423–438 (2020)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 1–35 (2023)
10. Liu, X., et al.: GPT understands, too. arXiv preprint arXiv:2103.10385 (2021)
11. Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.H., Riedel, S.: Language models as knowledge bases? arXiv preprint arXiv:1909.01066 (2019)
12. Qin, G., Eisner, J.: Learning how to ask: querying LMS with mixtures of soft prompts. arXiv preprint arXiv:2104.06599 (2021)
13. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 5485–5551 (2020)
14. Schick, T., Schütze, H.: Exploiting cloze questions for few shot text classification and natural language inference. arXiv preprint arXiv:2001.07676 (2020)
15. Vaswani, A., et al.: Attention is all you need. Advances in neural information processing systems 30 (2017)