A Medical Question Classification Approach Based on Prompt Tuning and Contrastive Learning

Qian Wang
School of Computer Science and Information Engineering
School of Artificial Intelligence
Hubei University
Wuhan, China
wqmaster@stu.hubu.edu.cn

Cheng Zeng*
School of Computer Science and Information Engineering
School of Artificial Intelligence
Hubei University
Wuhan, China
zc@hubu.edu.cn

Yujin Liu
School of Computer Science and Information Engineering
School of Artificial Intelligence
Hubei University
Wuhan, China
1298740592@qq.com

Peng He
School of Computer Science and Information Engineering
School of Cyber Science and Technology
Hubei University
Wuhan, China
penghe@hubu.edu.cn

***Abstract*—COVID-19 has profoundly impacted people's lives, and people are more concerned about medical and health issues, so it is essential to design an efficient method for classifying medical questions. Fine-tuning paradigms based on pre-trained language models have proven effective in recent years. However, PLMs based on fine-tuning paradigms are poorly robust, and there is a gap between the pre-training phase and the downstream task form, resulting in PLMs that cannot use the rich latent knowledge in downstream tasks. We propose a medical question classification method that combines prompt fine-tuning and contrastive learning and uses the large-scale knowledge graph enhancement model ERNIE 3.0 as a feature extractor to address both problems. Our approach utilizes an additional prompt template to enable PLM to unleash the potential in specific tasks and uses a contrast sample strategy to alleviate the problem of confusable samples that are difficult to distinguish. Experiments on a medical question classification dataset show that the method achieves an accuracy of 93.65 percent, with better metrics than recent work.**

Keywords—medical question classification；contrastive learning；prompt tuning；ERNIE 3.0

I. Introduction

The COVID-19 pandemic has inflicted immense damage worldwide, revealing the fragility of global healthcare delivery systems. The internet is inundated daily with concerns about the health consequences of COVID-19 sequelae[1]. Given this context, accurate medical question classification is of paramount interest. However, due to the scarcity of professional medical guides and ambiguous phrasing of medical inquiries, efficiently handling user-raised medical questions is challenging. Precise categorization of medical health questions is thus crucial.

In this paper, we ground the Medical Question Classification Task (MQC), where the goal is to accurately classify medical questions. Currently, MQC is defined as a single-label multi-class prediction problem[2], and the implementation methods are mainly classified into statistical-based machine learning methods and neural network-based deep learning methods. In recent years, with the arrival of the artificial intelligence wave, deep learning has been the most widely used method in MQC. However, there are drawbacks of insufficient extraction of problematic features and inability to utilize the corpus. Pre-trained language models solve this problem. Rasmy et al. [1] developed MedBERT, a medical knowledge-aware BERT model based on medical knowledge, which can give larger weights to medical concept words in the problem text and achieve good classification under fine-tuning. He et al. [4] proposed a BERT model to extend long and multi-type texts by integrating medical knowledge graphs, which also achieved better classification results under fine-tuning. But an important issue is overlooked: the target task for the initial training of PLMs does not match the downstream task, resulting in the inability to unleash the potential of PLMs in a specific task.

To address the problems of the medical question classification task, we propose a medical question classification method based on prompt tuning and contrastive learning. We adopt ERNIE 3.0[5] as a feature extractor to support the construction of prompt learning. ERNIE 3.0, a PLM with a large-scale knowledge base, performs surprisingly well in natural language processing tasks. Moreover, we introduce the contrastive learning approach to the current prompt-based fine-tuning task. Our contributions can be summarized as follows:

- For the medical problem classification task, we propose a model that combines prompt learning and supervised contrastive sample learning. We construct a prompt template as part of the original input to assist ERNIE 3.0 in achieving excellent classification results in our experiments.
- In the training process, we adopt the comparative sample learning strategy to alleviate the problem of insufficient medical samples.

II. Related Work

A. Knowledge Enhanced Pre-Training model ERNIE 3.0

PLMs such as BERT[6], RoBERTa[7], and GPT[8] perform impressively on various natural language understanding and generation tasks. However, most models learn only on corpora and lack experience in aspect-based tasks. Therefore, this approach performs poorly in solving downstream language understanding tasks. In this paper, we use a pre-trained language model containing billions of parameters for Chinese knowledge enhancement trained by PaddlePaddle[9]. It fuses autoregressive and self-coding networks and was trained on a 4TB Chinese corpus and a large-scale knowledge graph corpus. As shown in Figure 1, ERNIE 3.0 uses multilayer Transformer-XL[10] as the basic structure. In ERNIE 3.0 architecture, the universal representation module is used to extract initial semantic features, and these parameters are shared in the understanding and generation tasks. The task-specific representation plays the role of task-specific semantic feature extraction, and the task-specific target learns the parameters.

DOI reference number: 10.18293/SEKE2023-025

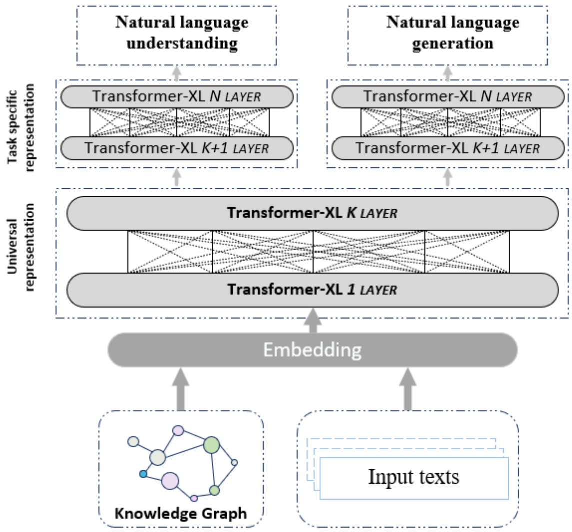


Figure 1. ERNIE 3.0 Model Structure.

B. Prompt-based Tuning

Pre-trained language models based on prompt turning have been shown to be effective for different NLP tasks [11]. Prompt universal process, as shown in Figure 2: identify PLM tasks, design prompt engineering, and answer engineering. PET [12] is a traditional prompt learning method requiring less annotated data than fine-tuning. Gao et al.[13] chooses to optimize prompt token embedding based on prompt tuning and proposes a small-sample fine-tuning method based on PLM, which also performs well in scenarios with few resources. Liu et al.[14] proposes P-tuning to insert optimized pseudo prompt tokens at the input side to find knowledge templates in continuous space automatically.

C. Contrasitve Learning

In many practical applications, researchers face challenges in collecting large-scale data and achieving precise labeling. To overcome these issues, researchers have proposed several methods for deep feature extraction from unlabeled or weakly labeled samples. The fundamental concept underlying contrastive learning is to measure the similarity of sample pairs in the feature space. By leveraging data sample labeling information, this technique pulls similar samples closer together in the feature space during training, while increasing the distance between dissimilar samples. SimCSE [15] implements textual sentence-level semantic representation based on contrastive learning and dropout, unsupervised SimCSE uses dropout to construct positive and negative samples and supervised SimCSE constructs positive and negative samples with the help of contradictory labels contained in the NLI dataset. However, in supervised situations, the scarcity of NLI data is not conducive to enhancing the semantic representation of sentences. Khosla et al.[16] introduced contrastive learning to the supervised scenario and achieved significant results in text classification task.

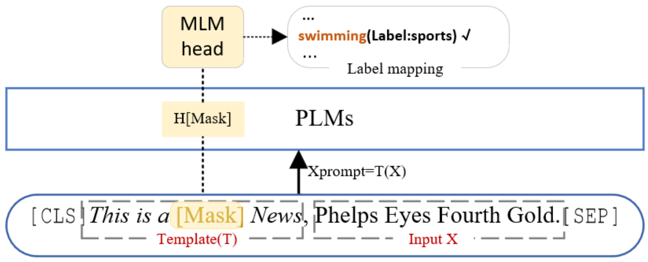


Figure 2. Using prompt learning in text classification

III. Approach

In this section, we discuss our proposed ERNIE 3.0 medical question classification method based on prompt-tuning and contrastive learning. The model's primary steps are prompt template construction and contrastive sample learning, as shown in Figure 3. The steps are detailed as follows:

A. Prompt Engineering

Prompt template construction: In this paper, the prompted approach treats the classification task as a masked language modeling (MLM) problem. The training data for our classification task is composed of $D=\{x_i, y_i\}$, where x is the set of input texts and y is the set of labels corresponding to the texts, with each text corresponding to a label. We use specific [Mask] token and template T to construct the input instance x_p and set the template T as "This is a [Mask][Mask] intention" for this task, while we define a set of characters P, p∈P to be used for filling the prompt template.

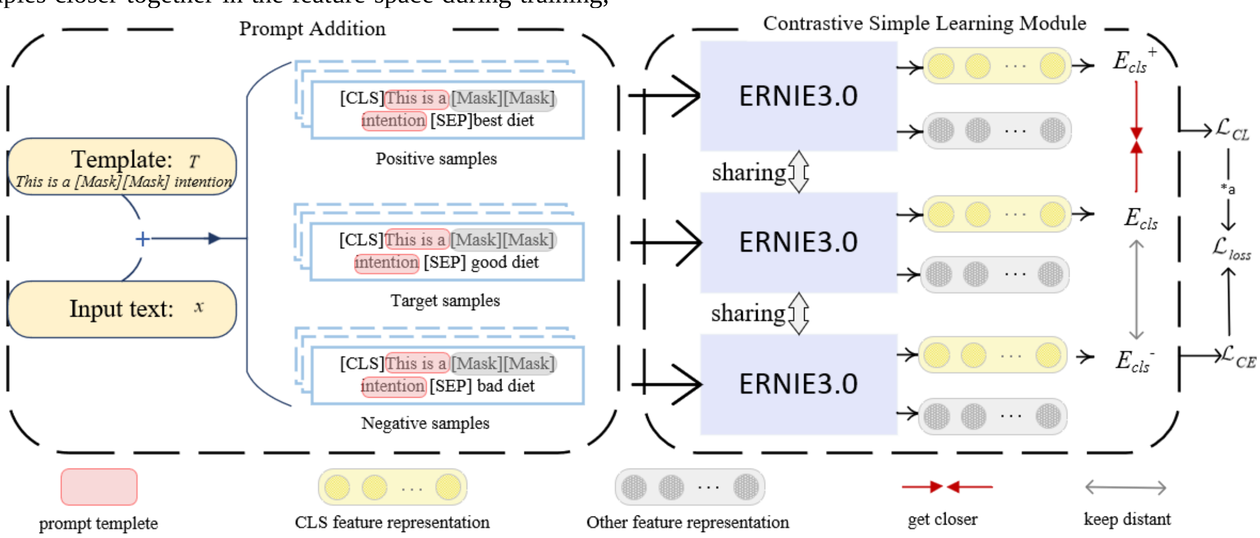


Figure 3. The framework of ERNIE 3.0-CL on Prompt turning.

Combination of input and template: Before prompt-based contrastive learning, we need to convert x to x_{prompt}, as shown in Figure 3, where we place the template in front of the sentence, and the final input to the model is：

$$x_{prompt} = \{[CLS], T, [SEP], x\} \quad (1)$$

Answer Mapping: We use $h \in \mathbb{R}^{\ell \times d}$ as the [Mask] hidden layer state of the PLM, where ℓ represents the length of x_{prompt}, d represents the hidden layer size. To map hidden vectors to relational labels, we define a verbalizer V_φ as a mapping method. The probability score of category i is：

$$p(y_i | x) = p(V_\varphi | x_{prompt}) = \frac{\exp(W_i \cdot h)}{\sum_{j=1} \exp(W_j \cdot h)} \quad (2)$$

where W is the static embeddings of the label prompt token, with size $h \times v$.

We build prompt input sentences and make the model predict the label with the highest correlation to [mask]. Consequently, our classification task is transformed into cloze-style task.

B. Contrasitve Simple Learning Module

Large-scale labeled data collection is complex and expensive in many application situations. To address this problem, scholars have proposed various methods for engineering features on unlabeled data. Many scholars have proved self-supervised and supervised contrastive learning methods to improve the effectiveness of sample features and the robustness of models. However, self-supervised contrastive learning does not make effective use of supervised information, and it is not applicable to the question classification task. In this paper, we use the supervised contrastive learning strategy to expand the features; referring to the previous study [16], we define the i-th sample as the target sample, the j-th sample as the positive sample, and the remaining N-2 samples as the negative samples. The following contrastive loss is defined for supervised tasks:

$$\mathcal{L}_{cl} = \frac{1}{N} \sum_{i \in I} \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}_i} -\log \frac{\exp(z_i \cdot z_p / \gamma)}{\sum_{a \in \mathcal{A}_i} \exp(z_i \cdot z_a / \gamma)} \quad (3)$$

where $\mathcal{P}(i)$ is the set of indices of all positive simples in the mini-batch, γ is a temperature coefficient, N is the number of batch size, z is the normalized representation of the ERNIE 3.0 feature vectors h based on the prompt fine-tuning method. Although this approach effectively compares positive and negative samples, we still need a cross-entropy loss to fine-tune the classifier:

$$\mathcal{L}_{CE} = \frac{1}{N} \sum_{i \in I} -\log \frac{\exp(z_i \cdot z_p)}{\sum_{a \in \mathcal{A}_i} \exp(z_i \cdot z_a)} \quad (4)$$

Finally, the total model loss is:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha * \mathcal{L}_{CL} \quad (5)$$

where α is the hyperparameter used to control the contrastive learning and cross-entropy loss.

IV. EXPERIMENTS

A. Experimental dataset and settings

In this paper, we use a crawler tool to crawl texts from mainstream medical websites to construct a medical question dataset for experiments, using manual annotation. Based on the crawled data, we integrated and labeled 5 categories: disease diet, seasonal diet, sports and fitness, weight loss and beauty, and dietary contraindications.

TABLE I. DATA DISPLAY

Label	Text
seasonal diet	The principles of winter health care.
disease diet	What to eat after injury fracture.
dietary contraindications	People who should not eat watermelon.

The experiment in this paper is a multiclass classification task, and we use the accuracy rate and F1 value as the main evaluation metric.

B. Implementation Details

The model is trained using AdamW[17] gradient descent algorithm, with batch size of 64, maximum rounds of 50, and initial learning rate of 2E-5. We set dropout to 0.1 in all layers of the model. The hyperparameters in the contrastive learning module are the contrastive learning loss fusion factor α, the temperature coefficient γ. After several experimental comparisons, we obtained the optimal model parameters with partial hyperparameters of α set to 0.01 and γ set to 0.1.

C. Baseline models and approachs

We compare with several baselines in related work, using BERT and RoBERTa as pre-trained language models. We also compare the fine-tuning based training method of PLMs with the integration of other advanced methods SimCSE, R-drop.

V. ANALYSIS OF EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed method, we used the proposed method to compare each baseline and conducted experiments on the medical question dataset, and the results are shown in TABLE II. , from which we can obtain the following conclusions.

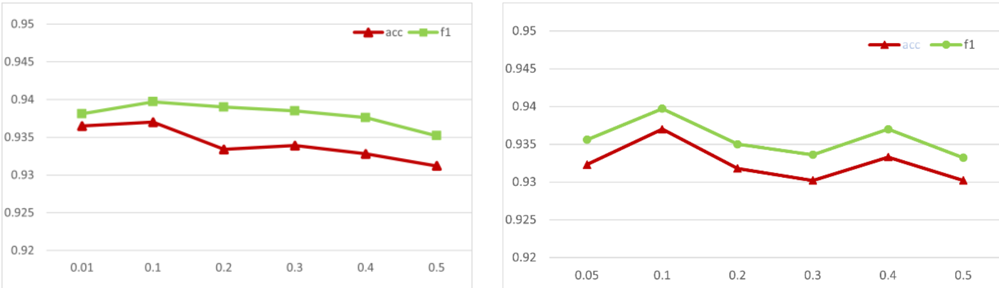
- BERT and RoBERTa have slight difference in experimental effect, ERNIE 3.0 introduces large-scale knowledge graph in pre-training model to get more text features and better classification effect.
- TABLE II. shows that our method works well for prompt-based methods. The Prompt-Tuning-based ERNIE3.0-CL method proposed in this paper improves the accuracy by 0.8 percentage points and the F1 value by 0.6 percentage points. Its performance is also better than PLMs using the fine-tuning paradigm.
- All ERNIE 3.0 fusion models are more effective than the baseline ERNIE 3.0. Our proposed ERNIE 3.0-CL compares better with other methods of ERNIE 3.0 fusion with contrastive learning, ERNIE 3.0-Rdop, and ERNIE 3.0-SimCSE. We believe this is because contrastive learning learns a more generalized representation from positive and negative case samples; contrastive learning module does improve the model's ability to discriminate between negative samples.

TABLE II. THE COMPARATIVE RESULTS OF THE EXPERIMENTS

	Model	*Acc*	*F1*
Fine-tuning	BERT	0.9213	0.924
	BERT-Rdrop	0.9228	0.9284
	BERT-SimCSE	0.9259	0.9306
	RoBERTa	0.9265	0.9303
	RoBERTa-Rdrop	0.9276	0.9306
	RoBERTa-SimCSE	0.927	0.9318
	RoBERTa-CL	0.9291	0.9323
	ERNIE3.0	0.9281	0.9334
	ERNIE3.0-Rdrop	0.9286	0.9314
	ERNIE3.0-SimCSE	0.9307	0.9327
Prompt-tuning	**ERNIE 3.0-CL**	**0.9365**	**0.9391**

VI. HYPERPARAMENTER INFLUENCE

In supervised contrastive learning, there are some hyperparameters that can affect the model's performance and training process. In order to explore the effectiveness of contrastive learning, we set different fusion ratio α and temperature coefficient γ to control contrastive learning loss. The correct setting of temperature parameter can make the model learn hard negatives better. As shown in Figure 4, after experimental comparison of the two hyperparameters we set, when the temperature coefficient is set to 0.1, and the fusion ratio is set to 0.01, our model gets the best result.



（a）Loss function fusion ratio

(b) Temperature coefficient

Figure 4 . Hyperparameter comparison.

VII. CONCLUSION

In this paper, we design a novel supervised contrastive learning method for medical question classification using ERNIE 3.0 pre-trained language model and prompt fine-tuning. The classification problem is transformed into a completion fill-in-the-blank problem through a manual prompt template to learn words mapped to labels and using a contrastive sampling module to pull in similar samples and push negative samples from different classes. We explore the efficiency of our model on the medical question classification task and outperform recent work in experimental tasks. However, our proposed model has limitations: The prompt template uses manual templates, which have more parameters to optimize and are prone to overfitting. In our subsequent work, we will take the automatic generation of prompt templates and the optimization of sampling strategies as our direction and try experimenting with different tasks.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 62102136), the Key R & D projects in Hubei Province （No.2021BAA188, 2021BAA184, 2022BAA044）, the Science and Technology Innovation Program of Hubei Province (No.2020AEA008).

REFERENCES

[1] S. Roy, S. Chakraborty, A. Mandal, and G. Balde, “Knowledge-Aware Neural Networks for Medical Forum Question Classification,” arXiv preprint arXiv:2109.13141, 2021.

[2] R. Jalan, M. Gupta, and V. Varma, “Medical forum question classification using deep learning,” In European Conference on Information Retrieval. Springer, 45–58, 2018.

[3] L. Rasmy, Y. Xiang, Z. Q. Xie, C. Tao, and D.G. Zhi, “Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction,” arXiv preprint arXiv:2005.12833,2020.

[4] Y. He, C. Wang, S. Zhang, N. Li, Z. R. Li, and Z. Y. Zeng. “KG-MTT-BERT: Knowledge Graph Enhanced BERT for Multi-Type Medical Text Classification,” arXiv preprint arXiv:2210.03970,2022.

[5] Y. Sun, S.H. Wang, S.K. Feng, S.Y. Ding, C. Pang, and J.Y. Shang, ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation.arXiv preprint arXiv:2107.02137,2021.

[6] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. “BERT: pre-traini ng of deep bidirectional transformers for language understanding,” In Proceedings of NAACL-HLT, pages 4171–4186. 2019.

[7] Y.H. Liu, M. Ott, N. Goyal, J.F. Du, M. Joshi, and D.Q. Chen, “Rober ta: A robustly optimized bert pretraining approach,” arXiv preprint ar Xiv:1907.11692, 2019.

[8] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. “Improvin g language understanding by generative pre-training,” URL https://s3-us-west-2. amazonaws. com/openaiassets/researchcovers/languageuns upervised/language understanding paper. pdf, 2018.

[9] Y.J. Ma, D.H. Yu, T. Wu, and H.F Wang. “Paddlepaddle: An open-source deep learning platform from industrial practice,” Frontiers ofData and Domputing, 1(1):105–115, 2019.

[10] Z.H. Dai, Z.L Yang, Y.M. Yang, J. Carbonell, V. Le, and R. Salakhut dinov. “Transformer-XL: Attentive Language Models beyond a Fixed -Length Context.,” arXiv preprint arXiv:1901.02860, 2019.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, and P. Dhariwa l, “Language models are few-shot learners,”. In Proceedings of NeurIP S, volume 33, pages 1877–1901,2020.

[12] T. Schick and H. Schütze. “Exploiting cloze-questions for few-shot te xt classification and natural language inference,” In Proceedings ofEA CL,pages 255–269, 2021.

[13] T.Y. Gao, A. Fisch, and D.Q. Chen. “Making pre-trained language mo dels better few-shot learners,”. In Proceedings of the 59th Annual Mee ting ofthe Association for Computational Linguistics and the 11th Inte rnational Joint Conference on Natural Language Processing, 2021.

[14] X. Liu, Y.N. Zheng, Z.X. Du, M. Ding,Y.J. Qian, Z.L. Yang, and J. T ang. “Gpt understands, too,” arXiv preprint arXiv:2103.10385,2021.

[15] T.Y.Gao,X.C.Yao, and D.Q.Chen, “SimCSE: Simple Contrastive Lear ning of Sentence Embeddings”arXiv preprint arXiv:2104.08821,2021.

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, and Yong long Tian, Supervised contrastive learning. In NeurIPS, volume 33, pa ges 18661–18673, 2020.

[17] I. Loshchilov and F. Hutter. “Decoupled weight decay regularization,” In ICLR,2018