Prompt Learning Classification Task Based on Improved Answer Space

Jiaxi Yang, Baoshan Sun

Abstract— **In order to solve the problem of single label word selection in prompt learning, this paper proposes a method of Probabilistic Answer Set. Add the prefixed space vector in front of the input text sentence as a template for prompt learning. Constructing an answer space set for each label word, and the probability that the text belongs to this category is obtained in the form of probability of the answer set. The experimental results show that on the THUCNews news classification datasets, the algorithm can improve the accuracy and F1-score by 1 percentage points compared with the common classification model. It can improve by 0.5 percentage points compared to the fixed prompt learning mode. It shows that the Probabilistic Answer Set algorithm proposed in this paper is better than other models significantly in the text classification task.**

Index Terms— **Chinese text classification, Prompt Learning Model, Probabilistic Answer Set**

I. INTRODUCTION

With the continuous popularization of the Internet and the rapid development of informatization, the ways and means of obtaining useful information have become more and more efficient and diversified. Various social platforms and media software have become important ways for people to obtain useful information. According to the 49th Statistical Report on the Development of China's Internet issued by the China Internet Network Information Center in 2022， as of December 2021, the number of netizens in my country has officially exceeded the 1 billion mark, the Internet penetration rate has increased by
percentage points compared with the previous year.Beside that, people spend an average of 28.5 hours online per week, an increase of 2.3 hours compared with the previous year. When people are more and more inclined to communicate on social media platforms, a large amount of text data is generated, most of which are short text data. These data contain rich information and have become an important way for people to obtain usefulinformation. These data are usually very concise and highly generalized, and has high value to research.

Manuscript received October 05, 2022.
Jiaxi Yang, School of Computer Science and Technology, Tiangong University, Tianjin, China.
Baoshan Sun, School of Computer Science and Technology, Tiangong University, Tianjin, China.

T.Mikolov et al. proposed the Word2vec model based on neural network in 2013, using Skig-Gram and CBOW algorithm to maping word vector into a
low-dimensional vector, which is widely used in various tasks of Natural Language Processing. The emergence of TextCNN, LSTM, GRU, ELMO and other models further optimizes the neural network and achieves a better effect in various tasks. Libo Han et al. integrated knowledge perception and dual attention mechanism on the basis of TextCNN, and the classification effect has been significantly improved. And then in 2017, Google officially proposed the Attention mechanism and the transformer network structure, which solved the problem of the semantic relationship of long-distance words and better expressed the representation ability between words. In 2018 , Devlin J et al. proposed the Bert model[1] by stacking the encoder structure of multi-layer transformers, which achieved a better effect on each task in Natural Language Processing. The WWM mechanism based on the Bert model[2] and the proposal of models such as Albert[3], ERNIE[4], Macbert[5] also bring the Bert model to a better advantage in the Chinese text classification field.
In the case of small data samples, in order to make the downstream tasks of the model closer to the pre-training tasks to achieve better results, Timo Schick et al. used prompt learning to solve the problem of text classification in 2020[6].In the same year, Timo Schick et al. solved the multi-token problem of label mask prediction[7] and once again proposed to expand on the basis of the small-sampletext classification model PET, replacing the multi-class maximum likelihood estimation in the original target with multiple one-vs-rest classifications[8]. After that, Hu S expands the label words and introduces external knowledge as the choice of label words[9]. In 2021, Zhang N et al. changed both label and prompt to continuous, and added the template [mask] word target, and the effect was significantly improved[10]. Kumar S et al. found the optimal prompt through multiple iterations[11]. After that, Tsimpoukelli M introduced prompt learning into multimodal learning, realizing multimodal small sample learning[12].
In this paper, we will base the prompt learning method to propose a method of Probabilistic Answer Setto classify short texts and then classify them on THUCNews news classification datasets. The training comparison have shown the effectiveness of this method.

II. RELATED WORK

Short text classification is a supervised learning process, which has experienced a total of 4 paradigms in the development stage. The first normal form mainly relies on extracting feature engineering. The second normal form focuses on how to design a neural network to learn relevant features of texts. And the third normal form focuses more on the design of the objective, designing a reasonable objective

function in the pre-training and fine-tune stages. This paper proposes an label words set based on the fourth normal form prompt learning method.

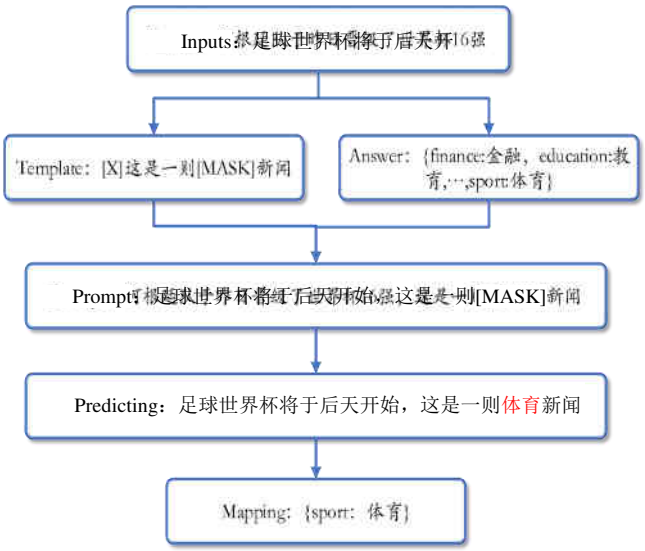
Prompt learning

For a long time, NLP tasks have adopted the solution of Pretrain + Fine-tuning (Model Tuning), but this solution requires a new fine-tune model for each task, and cannot shared. But for a large pre-trained language model, this seems to be customized for each task, which is very inefficient. The pre-training task of the traditional Bert model is the Mask Language Model (MLM). However for the downstream text classification task, it is far from the pre-training task and can't take advantage of the pre-training language model.

Therefore, prompting learning is proposed.

Prompt learning is to add a template before and after the input sentence. The main function of the template is to prompt the task type of the input sentence. At the same time,the keywords of the input sentence are masked, so that it can let the model predict the original word there. It's a good way to convert text classification tasks into MLM word prediction tasks, and better use the advantages of pre-training languagemodels. The specific process of prompt learning is shown in Figure 1.

Figure 1. Prompt Learning flow chart



Compared with the original text classification task, prompt learning changes the form of model input, adds prompt templates before and after the original input, andconverts the original text classification task into a multi-word prediction task.

Answer engineering aims to find a suitable answer space Z and a mapping from the answer to the final input y.The reason why Prompt learning can achieve few shot or even zero shot depends on the powerful generalization ability of the pre-trained language model. However, the language model is for the full vocabulary when predicting, and the full vocabulary is not necessarily required for downstream tasks. To this end, Answer engineering needsto find an answer space that matches the downstream task, that is, to construct a candidate set, and define the mapping between the answer space and the final output, and the correspondence between the candidate words and the finaloutput.

The answer space in Answer engineering can consist of tokens, fragments, or complete sentences. Tokens and fragments are more common in classification-related tasks, and complete sentences are more common in generation-related tasks. The answer space can also be manually constructed or searched through a model. The candidate sets constructed by most methods are discretized, and only a few methods are continuous.

III. CONSTRUCTION OF ANSWER SPACE

When classifying, for the same category, there may be different label words with similar meanings that can represent the same category. At the same time, for the selection of different label words, the results obtained are also very different, as shown in Table 1

Table 1. Comparison of the effects of different label words

Template	Label	Acc	F1
<X> 这是一条 [MASK] 新闻	[财经, 教学,…, 运动]	0.929	0.816
<X> 这是一条 [MASK] 新闻	[金融, 教育,…, 体育]	0.937	0.821
<X> 这是一条 [MASK] 新闻	[经济, 教育,…, 竞技]	0.933	0.819

Based on the above reasons, this paper proposes a method of Probabilistic Answer Set. By predicting the probability of each word in the set of words with similar meanings of the label words, the probability that the text belongs to this category is calculated. This method fully considers the possibility of similar words to achieve betterresults.

The vocabulary uses the jieba participle vocabulary.

Because most of the label words are nouns, all the nouns are screened out in the vocabulary as candidate words for the label words. For the label words of each category, $Y = [label_0, label_1, label_2, \cdots, label_{N-1}]$ where N is the number of categories, since the pre-training language model adopts the single word segmentation method with the granularity of words, each label word is segmented first:

$$Tokenizer(Y[n]) = \{label \quad :[x^n, x^n, \quad, x^n \quad]\} \quad (1)$$

Among them, d_n represents the length of the label word *label,* after tokenization, and then vectorizes each label word:

$$Y_E = Embedding(Tokenizer(Y)) \quad (2)$$

$$Y_E[n] = \{label_n : [e_0^n, e_1^n, \cdots, e_{d_n-1}^n]\} \quad (3)$$

By averaging the vector matrix obtained by each label word, the corresponding label word vector can be obtained:

$$Y_{vector}[n] = \frac{\sum_{i=0}^{d_n-1} e_i^n}{d_n} = \frac{e_0^n + e_1^n + \cdots + e_{d_n-1}^n}{d_n} \quad (4)$$

For each word in the filtered vocabulary, calculate the cosine similarity with each label word:

$$\cos ine(X_{word}, Y_{vector}[n]) = \frac{X_{word} \cdot Y_{vector}[n]}{\|X_{word}\| \|Y_{vector}[n]\|} \quad (5)$$

Among them, X_{word} represents the word vector in the candidate lexicon, which is also the same as the label word processing method. After word segmentation and vectorization, the vector matrix is averaged to obtain the corresponding word vector, and d_{hidden_size} represents the hidden layer word vector dimension of the pre-training language model.

Since each word in the answer set has a different weight coefficient for the label set, that is, the importance index, this paper introduces a weight matrix $\eta_{label[n]} = [\eta^0_{label[n]}, \eta^1_{label[n]}, \cdots, \eta^{top_k-1}_{label[n]}]$. The initialization of the weight matrix is based on the cosine similarity between each word in the answer set and the label word to perform softmax calculation. The value obtained after is initialized.

The value of the probability that the model predicts that the text belongs to the category, that is, the weighted sum of the probability of predicting each word in the answer set of the category:

$$P(y=Y[n]|X_{input}) = \frac{\sum_{j=0}^{top_k-1} \eta^j_{label[n]} P([MASK]=w_j^{label[n]}|X_{input})}{top_k} \quad (6)$$

By this formula, the probability that the input text belongs to each category can be obtained, and the highest probability is the category to which the input text finally belongs.

IV. EXPERIMENT

A. Experimental dataset

In the experimental part of this paper, the dataset uses the Sina news classification dataset provided by Tsinghua University, and selects 10 categories including finance, education, and politics, etc. Each category selects 10,000 as training set, 1,000 as validation set, and 1,000 as test set. As shown in Table 2.

Table 2. Experimental Data Settings

Datasets	Categories	Train	Validation	Test
THUCNews	10	100000	10000	10000

B. Experimental results and Analysis

In order to verify the superiority of probabilistic answer space in processing text classification tasks, this paper uses three benchmark models as comparative experiments in the experiment. In the same environment to test the effect of the model proposed in this paper, the selected comparison model is as follows:

1）Roberta-wwm: This method directly uses Roberta-wwm for text classification without any processing of prompt.

2）Roberta+ fixed sentence prompt: This method adds a fixed prompt template <这是一条[MASK]新闻> after the input text sentence X_{input}, then the sentence input by the model becomes $<X_{input}><$这是一条[MASK]新闻>, After that, using the Roberta model predicts the word at [MASK] to divide the text, and the selection of the label word uses the label word of the original text division.

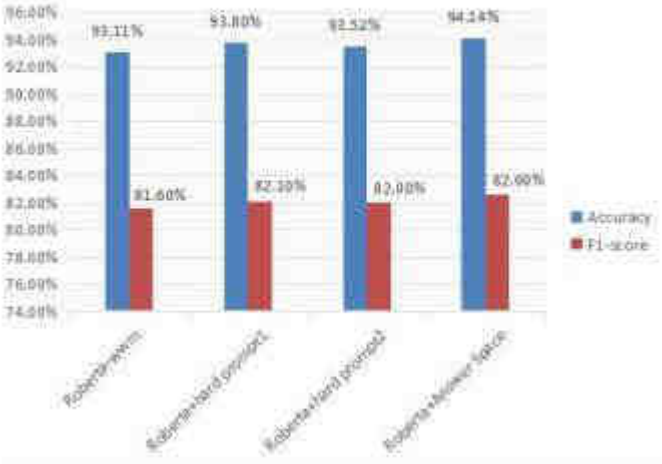
3）Roberta+ fixed sentence prompt: This method adds a fixed prompt template <播报一条[MASK]新闻> in front of the input text sentence X_{input}, then the sentence input by the model becomes <播报一条[MASK]新闻>$<X_{input}>$, and the selection of the label word uses the label word of the original text division.

Three benchmark models were selected to compare and verify the effects of this model on the THUCNews and Sohu news classification datasets. The experimental results are shown in Table 3.

Table 3. Comparison of model evaluation results

	THUCNews	
Model	Accuracy	F1-score
Roberta-wwm	93.11%	81.6%
Roberta+hard prompt1	93.70%	82.1%
Roberta+hard prompt2	93.52%	82.0%
Roberta+Answer Space	94.14%	82.6%

Figure 2. Comparison of the accuracy and F1-score of each model on the THUCNews datasets



From Table 3, Figure 2, it can be concluded that on the THUCNews news classification datasets, the effect of the model using prompt learning is significantly better than the effect of using the model to classify directly. When using a fixed sentence as a template, the effect of the suffix fixed

template is 0.28 percentage points higher than the effect of the prefix fixed template on the THUCNews datasets.

Under the condition of using the same template, the construction method of the label words also has a greatimpact on the final result. The Probabilistic Answer Set proposed in this paper is 0.3 percentage points better than the ordinary single label word when the suffix fixed sentence template is used. It shows the superiority of Probabilistic Answer Set in the construction of label word.

CONCLUSION

Based on a series of problems such as single label word selection in existing text classification tasks and prompt learning, this paper proposes a method of constructing a Probabilistic Answer Set. The Probabilistic Answer Set is added as the selection method of label words, and the probability that the text belongs to this category is obtained by calculating the probability of each word in the set. Experiments show that the accuracy and F1-score of the model can reach 94.14% and 82.6% on the THUCNews datasets respectively. Compared with other models, the classification effect of this model is significantly improved, and it has better results for news text classification.

REFERENCES

1. Kenton J D M W C, Toutanova L K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of NAACL-HLT. 2019: 4171-4186.
2. Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for chinese bert[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.
3. Lan Z, Chen M, Goodman S, et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations[C]//International Conference on Learning Representations. 2019.
4. Sun Y, Wang S, Li Y, et al. Ernie 2.0: A continual pre-training framework for language understanding[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 8968-8975.
5. Cui Y, Che W, Liu T, et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 657-668.
6. Schick T, Schütze H. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 255-269.
7. Schick T, Schütze H. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 2339-2352.
8. Schick T, Schmid H, Schütze H. Automatically Identifying Words That Can Serve as Labels for Few-Shot Text Classification[C]//Proceedings of the 28th International Conference on Computational Linguistics. 2020: 5569-5578.
9. Hu S, Ding N, Wang H, et al. Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 2225-2240.
10. Zhang N, Li L, Chen X, et al. Differentiable Prompt Makes Pre-trained Language Models Better Few-shot Learners[C]//International Conference on Learning Representations. 2021.
11. Kumar S, Talukdar P. Reordering Examples Helps during
12. Priming-based Few-Shot Learning[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 4507-4518.
13. Tsimpoukelli M, Menick J L, Cabi S, et al. Multimodal few-shot learning with frozen language models[J]. Advances in Neural Information Processing Systems, 2021, 34: 200-212.