Does BERT know Which Answer Beyond the Question?

CCKS 2021：面向中文医疗科普知识的内容理解（二）医疗科普知识答非所问识别 冠军方案

Medical question answering classification (MQAC) aims to construct an intelligent medical question answering classification system,which can use massive medical information content to quickly meet user needs and ensure the accuracy and authority of the content.

Our in novative approach has improved by an average of 1.164% compared to the baseline. The experiment fully shows that our innovative method is very effective. The method we proposed won the CCKS competition and was significantly ahead of the second opponent (close to 1%) in the final, showing extremely high practicality and effectiveness.

YIXUANWENG， FEI XIA， MAOJIN XIA， QIANG YU， JINFENG HUANG

Abstract

This poster presents our proposed framework for the Chinese MQAC organized by the 2021 China conference on knowledge graph and semantic comput ing (CCKS) competition, which requires correct classification of whether the answer can satisfy the related question. After the preliminary ex periment, we analyzed the bad cases in-depth and found hard instances and insufficient model generalization capabilities. In order to solve these problems, we have proposed a series of innovative strategies, including five categories of methods: 1. Four different adversarial training meth ods 2. Hard instances identification and multi-round training methods 3. Target instances constructed by similarity 4. Validation set retraining with a small learning rate 5. Medical word vector combined with Easy Data Augmentation(EDA) method for text data augmentation.

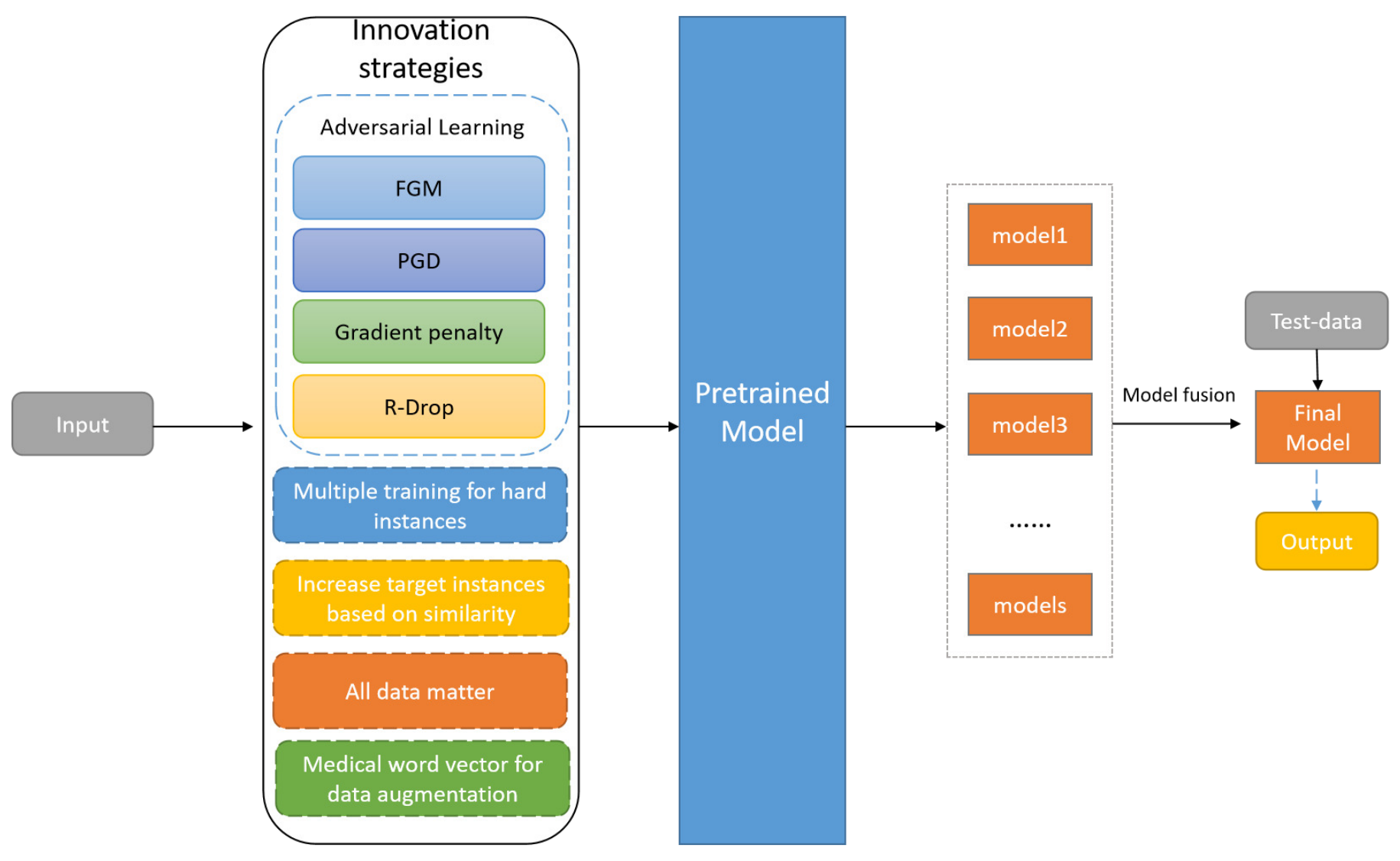


Fig. 1. The overall framework of the integration of innovation strategies.

R-Drop

R-Drop uses a simple dropout twice method to construct positive samples for comparative learning, significantly improving the experimental re sults in supervised tasks. Fig 2 shows the framework of RDrop.

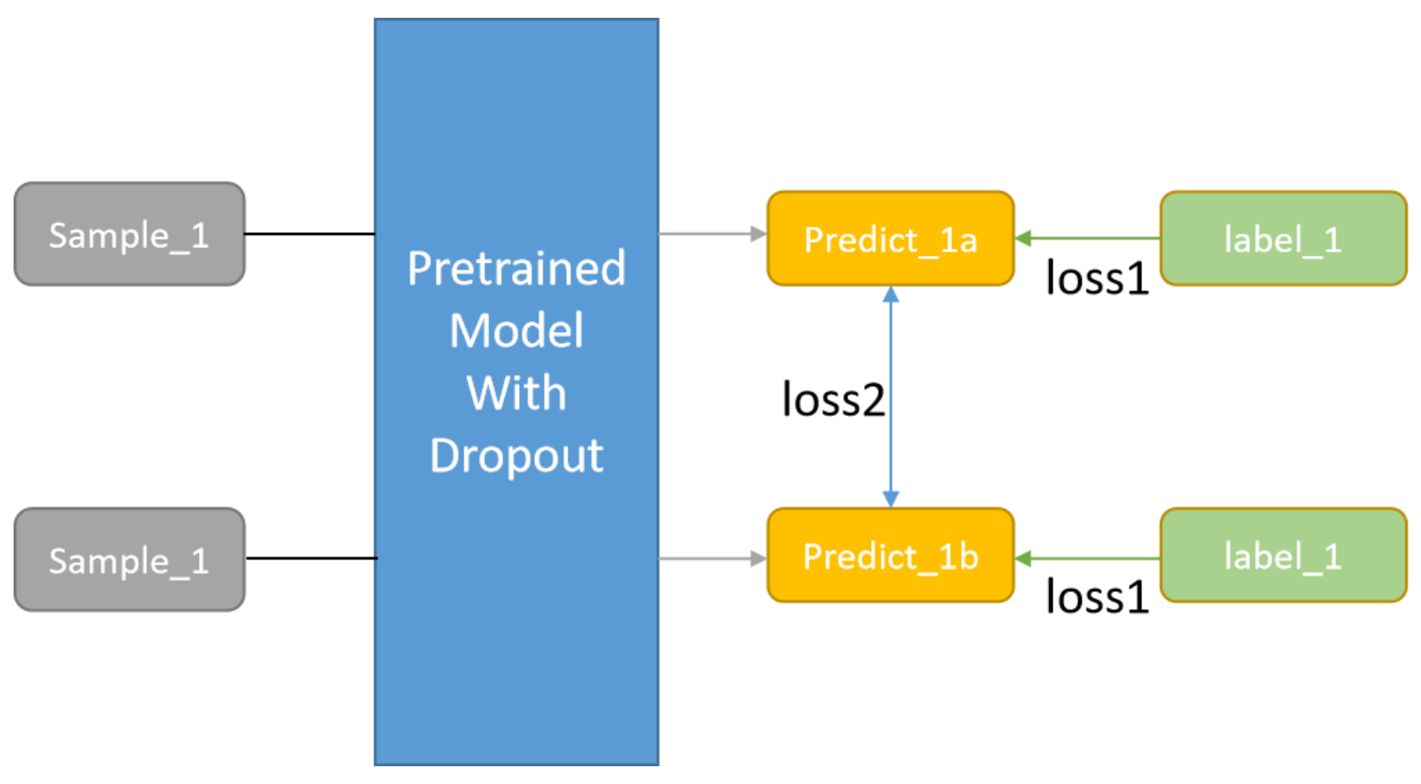


Fig. 2. Schematic diagram of method RDrop.

Multiple Training for Error-prone and Difficult Instances

In the bad case study, we found that many examples are hard and often incorrectly predicted by the model. We used the Chinese medical dialogue dataset to find this part of the examples. First, we used the baseline model to pseudo-label the Chinese medical dialogue dataset. Then we use this part of pseudo-label data to train a selection model for selecting error-prone and challenging samples. We use model to predict the original labeled training set and record the incorrect and low-confidence data as error-prone and complex samples. After that, we will add this part of the data in each training process and let the model train on this data multiple times. We found that this method can enhance the model’ s classification accuracy for complex samples.

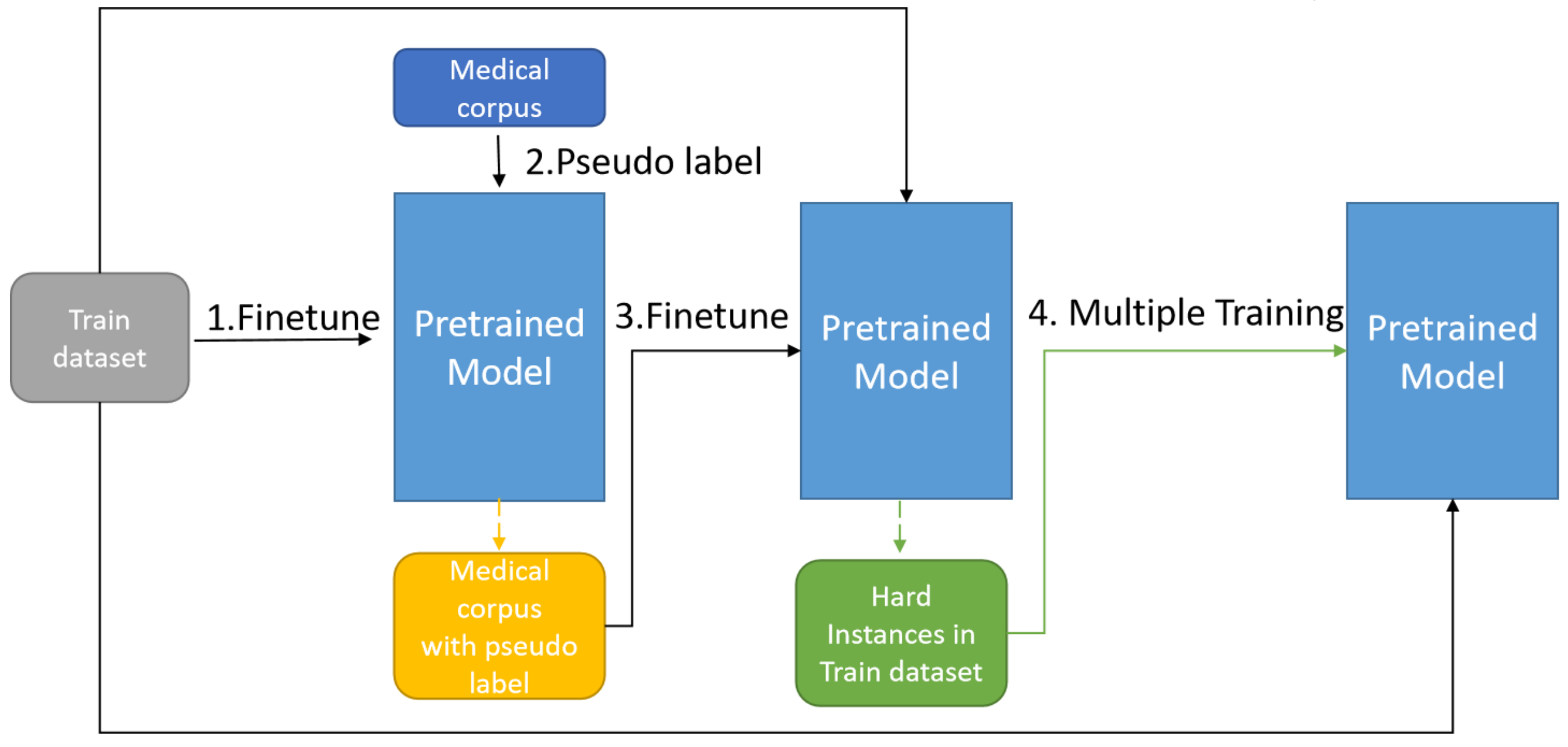
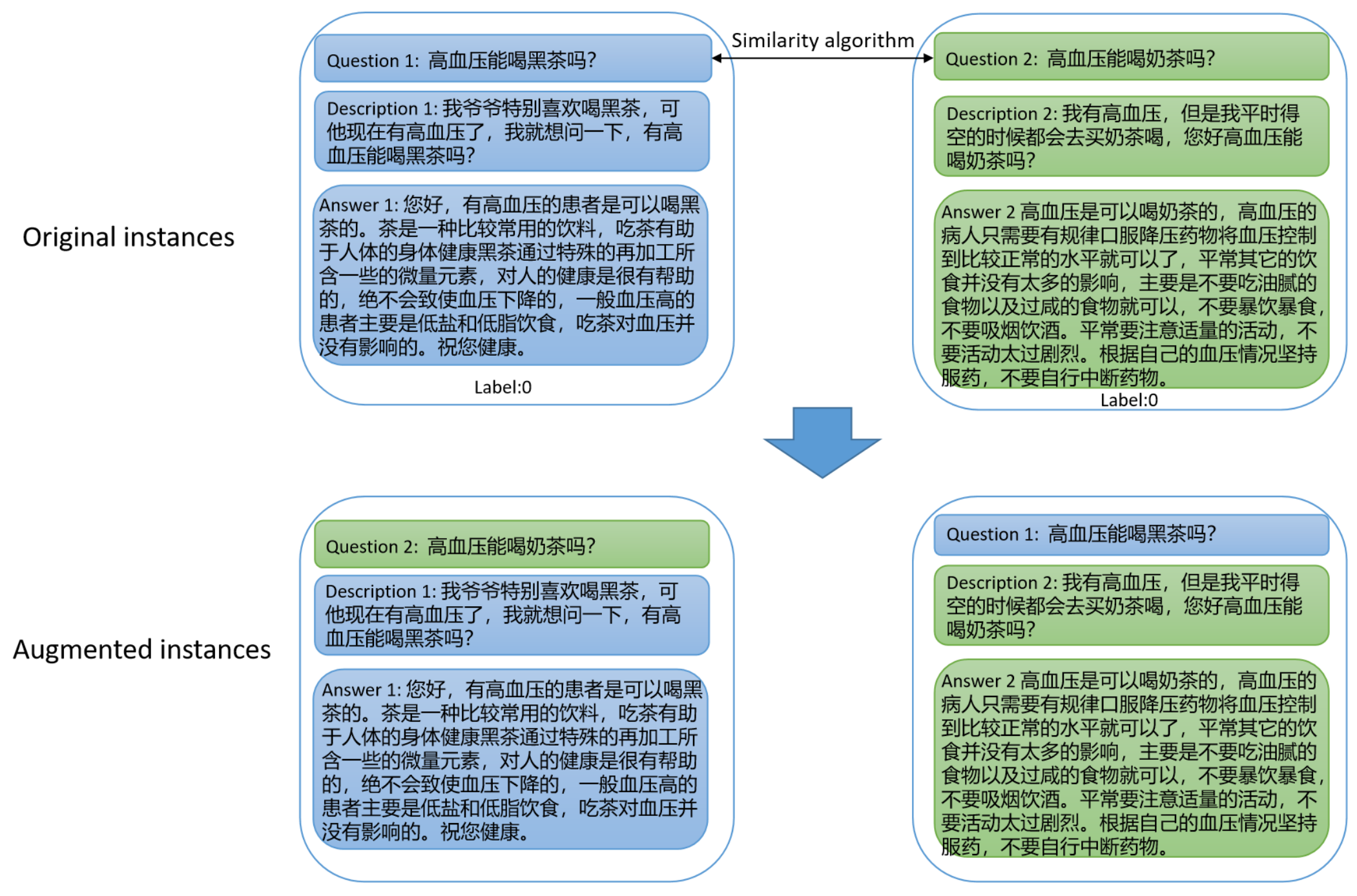


Fig. 3. Multiple Training for Error-prone and Difficult Instances.

Minimum Edit Distance Search to Increase target instances

In order to increase the number of samples with label 1, we also manually constructed the data. After analyzing the data, we found that even if the two questions are very similar, the corresponding answer is very different. In short, in the standard question and answer (label 0), replace the question with another very similar question, and the answer will be a non-ideal answer (label 1). More over, samples constructed in this way are generally more complex, which helps the model learn from hard samples. So we first find out the most similar set of answers in the training set and exchange the answers with each other, which constitutes two sets of answer data. takes much time, we finally used the classic edit distance method in measuring similarity to achieve this goal. Its performance is also excellent.



Question 1: 高血压能喝黑茶吗?

Description 1: 我爷爷特别喜欢喝黑茶，可他现在有高血压了，我就想问一下，有高血压能喝黑茶吗?

Answer 1: 您好，有高血压的患者是可以喝黑茶的。茶是一种比较常用的饮料，吃茶有助于人体的身体健康黑茶通过特殊的再加工所含一些的微量元素，对人的健康是很有帮助的，绝不会致使血压下降的，一般血压高的患者主要是低盐和低脂饮食，吃茶对血压并没有影响的。祝您健康。

Question 2: 高血压能喝奶茶吗?

Description 2: 我有高血压，但是我平时得空的时候都会去买奶茶喝，您好高血压能喝奶茶吗?

Answer 2 高血压是可以喝奶茶的，高血压的病人只需要有规律口服降压药物将血压控制到比较正常的水平就可以了，平常其它的饮食并没有太多的影响，主要是不要吃油腻的食物以及过咸的食物就可以，不要暴饮暴食，不要吸烟饮酒。平常要注意适量的活动，不要活动太过剧烈。根据自己的血压情况坚持服药，不要自行中断药物。

Question 2: 高血压能喝奶茶吗?

Description 1: 我爷爷特别喜欢喝黑茶，可他现在有高血压了，我就想问一下，有高血压能喝黑茶吗?

Answer 1: 您好，有高血压的患者是可以喝黑茶的。茶是一种比较常用的饮料，吃茶有助于人体的身体健康黑茶通过特殊的再加工所含一些的微量元素，对人的健康是很有帮助的，绝不会致使血压下降的，一般血压高的患者主要是低盐和低脂饮食，吃茶对血压并没有影响的。祝您健康。

Question 1: 高血压能喝黑茶吗?

Description 2: 我有高血压，但是我平时得空的时候都会去买奶茶喝，您好高血压能喝奶茶吗?

Answer 2 高血压是可以喝奶茶的，高血压的病人只需要有规律口服降压药物将血压控制到比较正常的水平就可以了，平常其它的饮食并没有太多的影响，主要是不要吃油腻的食物以及过咸的食物就可以，不要暴饮暴食，不要吸烟饮酒。平常要注意适量的活动，不要活动太过剧烈。根据自己的血压情况坚持服药，不要自行中断药物。

Fig. 4. Minimum Edit Distance Search to Increase target instances.

All data matter

Small learning rate fine-tuning in the validation set In order to make full use of the value of data, we also thought of validation set data. To measure the effect of model classification, we divide the original data into a training set and a validation set.

After every epoch training, we will verify its performance on the validation set, and finally, we choose the model that performs best on the validation set. The model has not seen these validation set data during the training phase, so it may be helpful if retrained. At first, let the currently trained optimal model train another round on the validation set, but there is no improvement in the expected effect. Instead, the model reduces part of the classification ability due to this part of the addition.

We analyze that it may have the following reasons: 1. With the catastrophic forgetting of deep learning, the model may forget the memory as training increases. 2. The distribution of the verification set may be different. The model may have learned some features, but its generalization ability may be weakened.

Therefore, we tried to reduce the learning rate and make the model learn less on these unseen data. Finally, the experiment showed that this method is very significant. This method is practical and straightforward and can be used in any task that uses deep learning. It has a general improvement. We think it is one of the powerful innovations of this experiment

Testset A data In the final test phase, we also used the test data from the A test phase. Considering our excellent performance in the A test phase and high accuracy, we believe that this part of the pseudo-label data with high accuracy is also valuable. So we add this part of the data to the model training, and use the trained model to predict the results. Unsurprisingly, our results have improved again, which proves that it is indeed effective. All data matter.

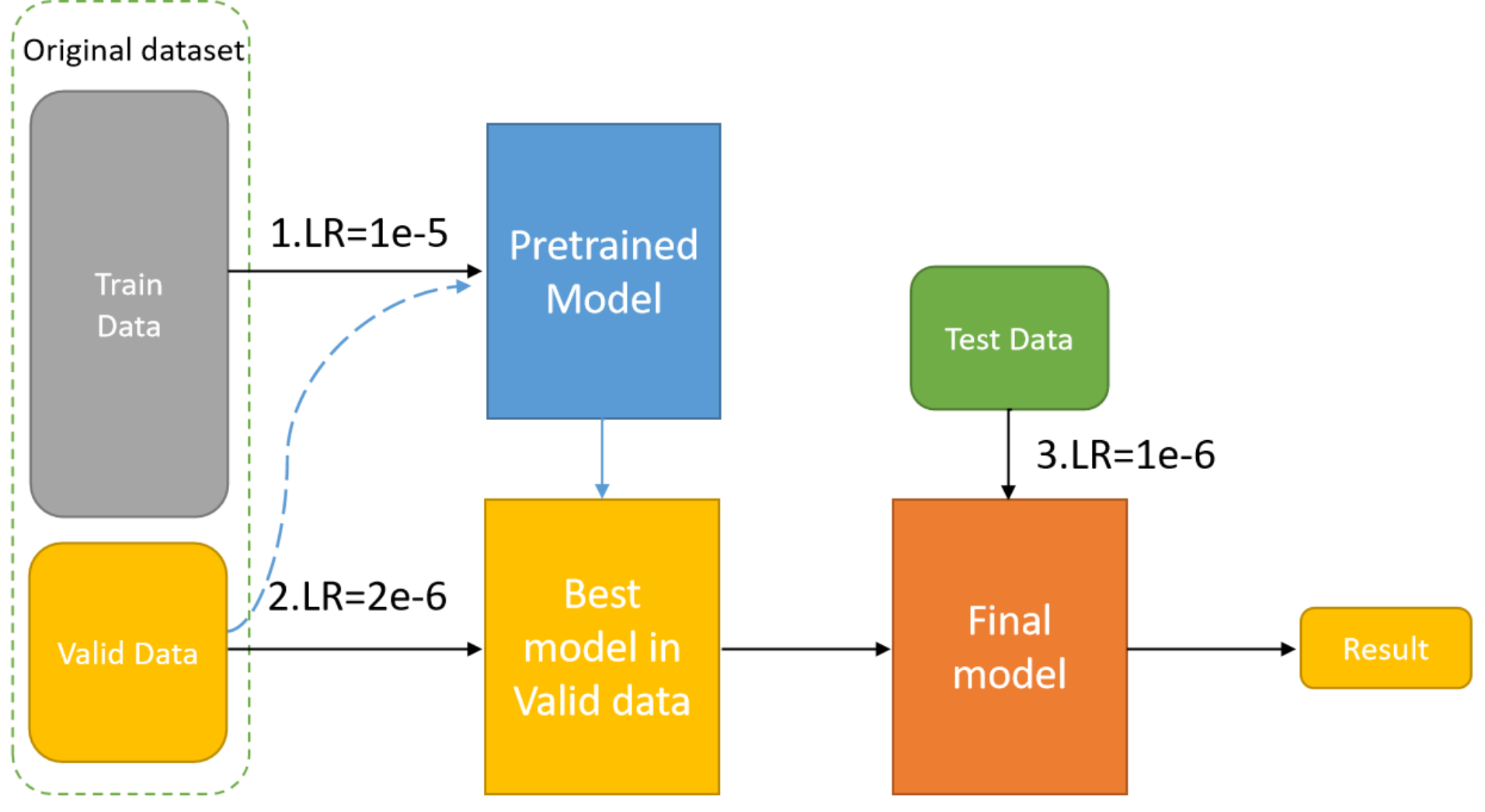


Fig. 5. Small learning rate fine-tuning in the validation set.

TestA			TestB		
Rank	Team Name	Score	Team Name	Score	Rank
1	DeepBlueAI	84.963	our Team	70.698	1
2	our Team	84.498	DeepBlueAI	69.972	2
3	FREE	84.436	Space Oddity	69.778	3
4	Space Oddity	84.379	united	68.96	4

wengsyx@gmail.com