AI CUP report

Members in team15

109136503 余頡　　109062470 馮少迪

109062561 陳楓翔　　109136504 林哲廣

109065466 蔡洵晟

Ⅰ. Introduction

This competition is to de-identify medical data. The goal is to catch words that contain private information and label them by type. The content of this report will include the pre-processing, hypothetical attempts, and actual results that we have performed.

The final result is through cleaning, correcting, converting, and cutting the original training data,and then use a pre-training language model such as roberta to generate word embedding on the original text data, and then input it into bilstm + crf (or only crf) for training with different parameters to get multiple different models and then use different models to predict the new test data, merge the different prediction results, and finally reach the f1 score of 76.37 on the retest set. The follow-up we will introduce in detail data pre-processing, framework, roberta and other pre-training language models, bilstm and crf,parameter settings,correction of prediction results, merging and experimental testing of various models, etc.

Ⅱ. Data pre-processing

A. NER tagging – POS pattern

The original data format is as shown in the figure below: It is composed of dialogue articles and private data.

The private information of Training data includes start, end position, entiy

醫師：你有做超音波嘛，那我們來看報告，有些部分有紅字耶
邊有看到肌肉，下面這邊就是肝臟，下面旁邊是膽囊，膽囊的
醫師：對阿，因為還要工作，你做甚麼工作？民眾：我打撲克

article_id	start_position	end_position	entit
0	55	57	68 med_exam
0	66	68	68 med_exam
0	1264	1271	10.78公分 med_exam
0	1358	1361	三多路 location
0	1374	1378	長庚醫院 location
0	1863	1865	十天 time
0	2072	2076	打撲克牌 profession

Figure 1. Original data set

text, entiy type and other data.We need to convert the data format into a format that the model can read. A text data and a corresponding label are as follows:

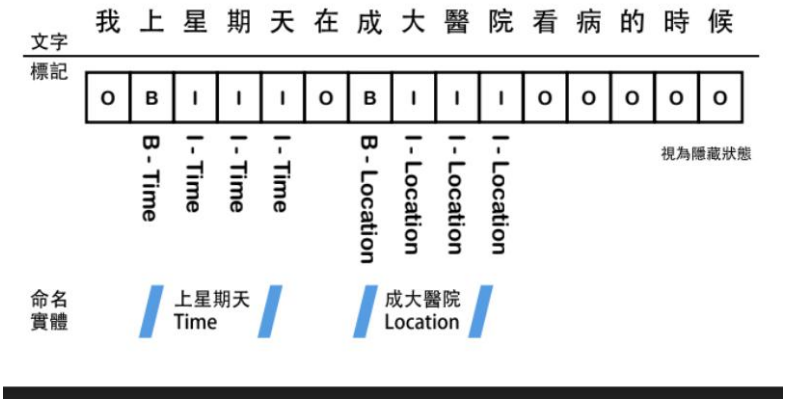


Figure 2. Text with NER label

First, we use the function of crf fomart data in the code of the baseline to convert the entire text data into the above NER format.

B. Elimination of error data

The original training data has some start positions, end positions are wrong (e.g. end <start), or the start position is incorrect. For incorrect end position, calculate the length of the entiy text and add the original start position to get the end position. For data with incorrect start pos, delete it directly to avoid training the wrong data. Improve the precision of the model.

C. Text length and form

In order to comply with the input of the pre-trained language model, first cut the entire article into each sentence according to the period, and stack each sentence into a strip of trading data with the length we want. We assume that the model accepts it during training. A longer text allows the model to

capture more contextual relationships and better recognize entity. So make the length of each training data close to the maximum length of the pre-trained language model (512). From the results, using the maximum length greatly improves the recall and precision of the model.

All full-size characters are converted into half-size characters to make the data in a unified format because it is a Chinese text.

D. Combination of Simplified Chinese training data

At first, only the data provided by the competition team was used for processing and training, but the result was not very satisfactory, so I found a lot of additional text information to improve the multiplicity of the data set, thereby reducing the bias of unknown data that cannot be identified and trained.

Many text resources we found come from Simplified Chinese, so the pre-training language model uses both simplified Chinese and traditional Chinese for training at the same time. Therefore, the original traditional Chinese text is converted into simplified Chinese, and then added to the original training data, which increases the original data by several times. Finally the recall of the model rises substantially because there are more kinds of trading data.

Finally, all data is divided into train data and valid data by 9:1

Ⅲ. Model

The following will briefly introduce a few models that we have studied and used, and detail the final model framework, parameter adjustment and integration methods used.

Pre-trained language model:

A.BERT model[1]

The full name of BERT is Bidirectional Encoder Representation from Transformers, that is, the Encoder of the two-way Transformer because the decoder cannot obtain the information to be predicted. The main innovations of the model are in the pre-train method, which uses Masked LM and Next Sentence Prediction to capture word and sentence level representations respectively.

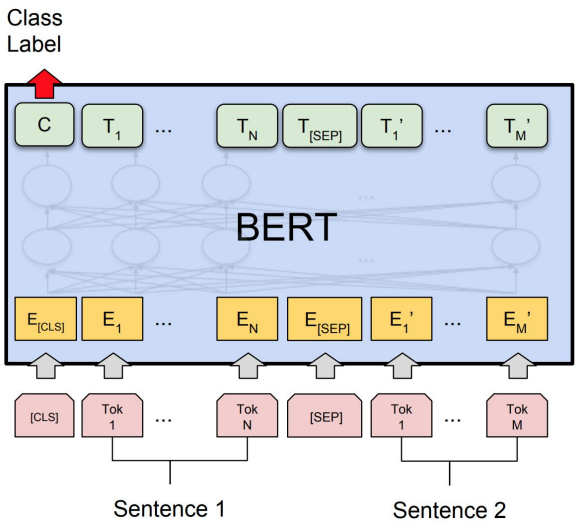


Figure 3. Structure of BERT

Bert Data reader:

a. WordPiece embedding(token embedding):

WordPiece refers to dividing a word into a limited set of common sub-word units, which can achieve a compromise between the validity of the word and the flexibility of the character.

b. Position Embedding:

Position embedding refers to encoding the position information of a word into a feature vector. Position embedding is a crucial part of introducing the position relationship of words into the model. The specific content embedded in the location refers to my previous analysis.

c. Segment Embedding:

It is used to distinguish two sentences, such as whether B is the following of A (dialogue scene, question-and-answer scene, etc.). For sentence pairs, the feature value of the first sentence is 0, an d the feature value of the second sentence is 1.

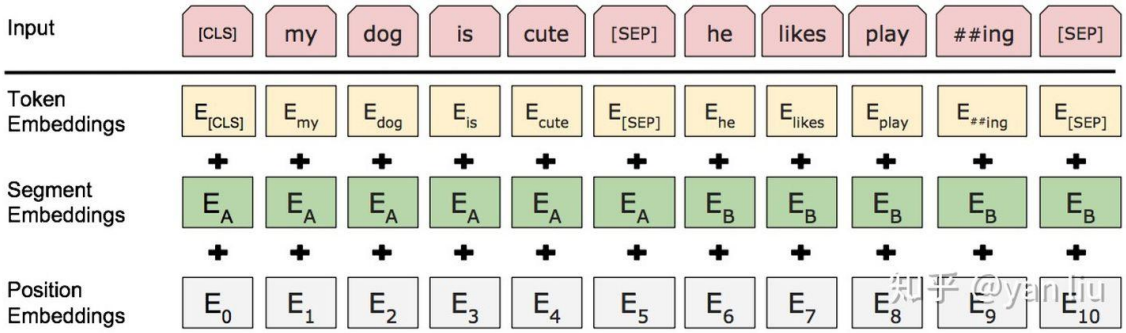


Figure 4. Bert reader embedding

B.Ernie model

ERNIE greatly enhances the general semantic representation ability by uniformly modeling the lexical structure, grammatical structure, and semantic information in the training data, and it has achieved significant results that surpass BERT in multiple tasks.0

C.Roberta model

RoBERTa introduced the paper using the original BERT architecture, but made some changes during the training process, and has more rigorous research on the impact of hyper-parameters, Improve training time (longer training time), expand the training batch set (larger batches), and use a longer sequence for training (long-term training),Dynamically generate the mask used by ML .Use standard training materials (CC-NEWS) and other changes, and of course, use more training materials.

D.Bilstm + crf[2]

Each word in the sentence is a word vector containing word embedding and word embedding. Word embedding is usually pre-trained, and word embedding is initialized randomly. All embeddings will be adjusted with the iterative process of training.

Secondly, the input of BiLSTM-CRF is the word embedding vector, and the output is the predicted label corresponding to each word.

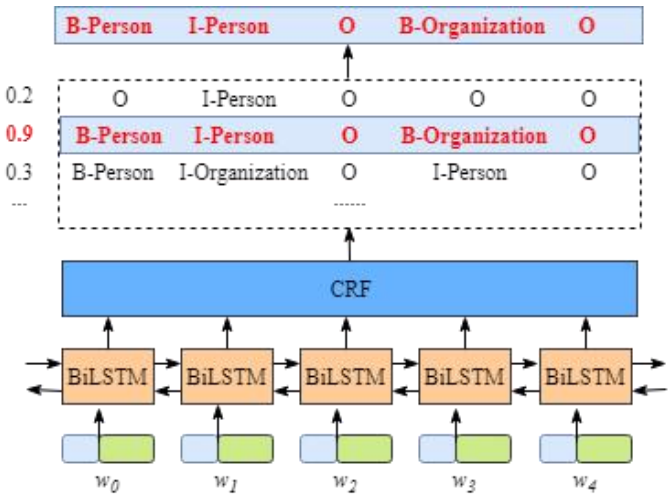


Figure 5. Structure of BiLSTM- CRF

- BiLSTM: The abbreviation of Bi-directional Long Short-Term Memory, is a combination of forward LSTM and backward LSTM. Using the LSTM model

can better capture the long-distance dependencies because LSTM can learn what information to remember and what information to forget through the training process. The input of the BiLSTM layer represents the score of each category corresponding to the word. For example, W0, the output of the BiLSTM node is 1.5 (B-Person), 0.9 (I-Person), 0.1 (B-Organization), 0.08 (I-Organization) and 0.05 (O). These scores will be the input of the CRF layer.

- CRF is a model for finding the conditional probability distribution of another set of output random variables under the condition of a given set of input random variables; its characteristic is that it assumes that the output random variables constitute a Markov random field, and the conditional random field can be used for different prediction problems , The natural language processing process is mainly a linear (linear chain) conditional random field. At this time, the problem becomes a discriminant model for predicting the output sequence from the input sequence, in the form of a log-linear model, and the learning method is maximum likelihood estimation or regularization Maximum likelihood estimation. The CRF layer can add some constraints to ensure that the final prediction result is valid. These constraints can be automatically learned by the CRF layer during training data. The possible constraints are:

a.The beginning of the sentence should be "B-" or "O", not "I-".

b."B-label1 I-label2 I-label3..." In this mode, categories 1, 2, and 3 should be the same entity category. For example, "B-Person I-Person" is correct, but "B-Person I-Organization" is wrong.

c."O I-label" is wrong, the beginning of the named entity should be "B-" instead of "I-".

With these useful constraints, erroneous prediction sequences will be greatly reduced.

- BiLSTM-CRF has reached or surpassed the CRF model based on rich features, and has become the most mainstream model of NER methods based on deep learning.

Ⅳ. System Model

A. Framework – Paddlehub

PaddleHub aims to provide developers with rich, high-quality, ready-to-use pre-trained models.

No deep learning background required, Covers 4 main categories of image,

text, audio and video, and supports one-click prediction, easy service deployment and migration learning. All models are open source and can be downloaded and used offline for free.

Figure 6. Interface of PaddleHub

B. Model structure:

We use a pre-trained language model such as roberta to generate word embedding for the original text data, and then input it into bilstm + crf (or only crf) for training with different parameters.

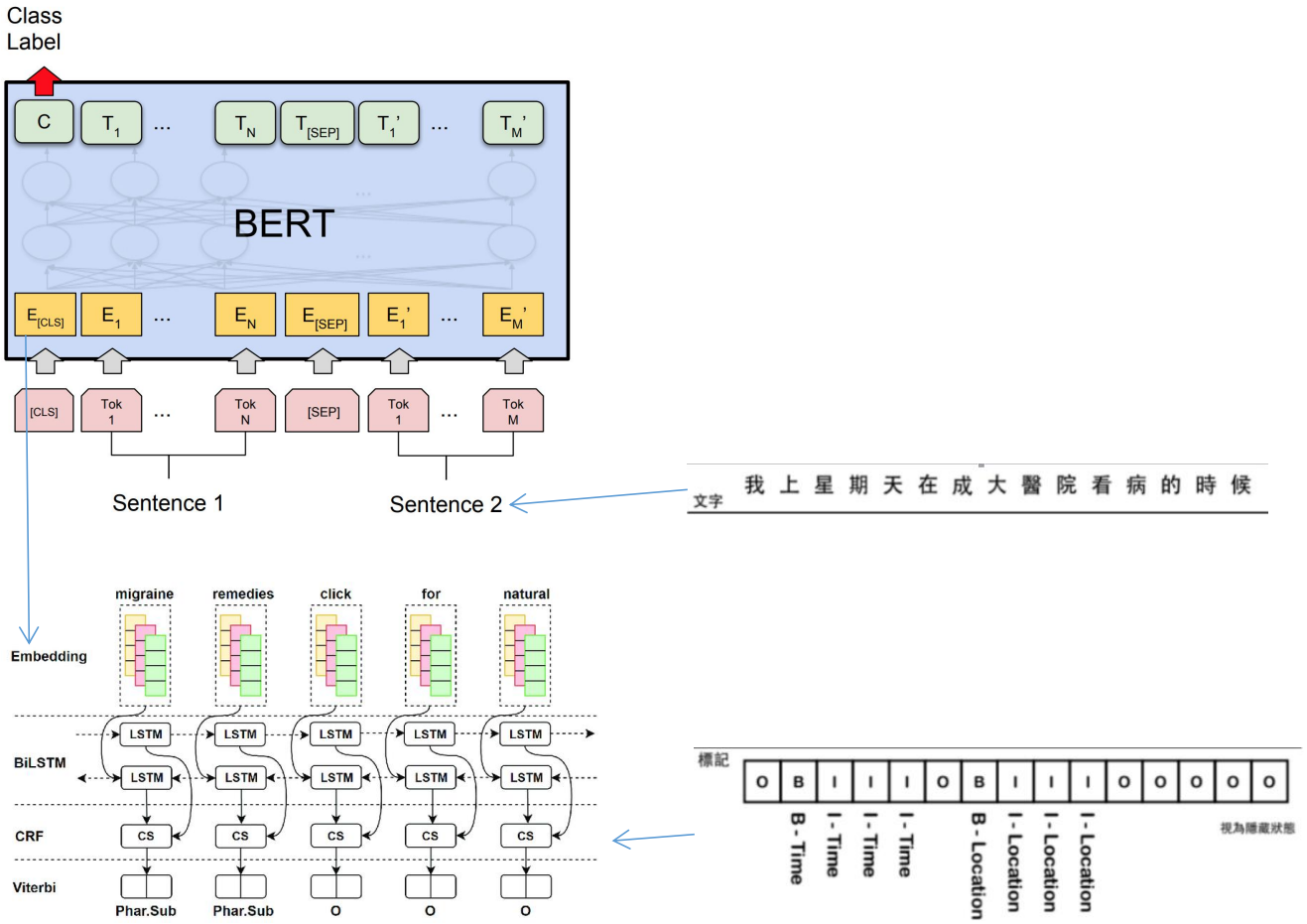


Figure 7. Combination of Roberta and Bilstm

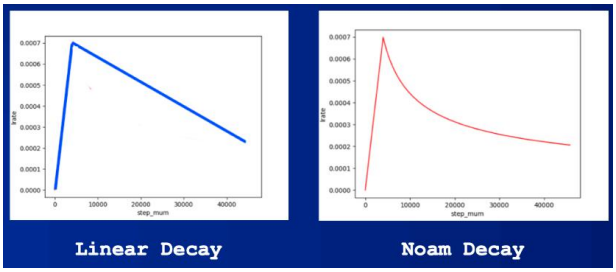
C. Parameter setting:

Roberta learning_rate: 5e-5 maximum learning rate

Bilstm lerarning_rate :1e-4
Crf learning_rate: 5e-3

Lr Hint: The roberta ontology has been pre-trained, that has weights, so it is easy to fine-tune to the best point with a small learning rate, and the next structure is to train from scratch, and training with a small learning rate not only learn Slow, and it is difficult to synchronize with the BERT ontology training. For this reason, we increase the learning rate of the downstream structure (blstm+crf) to try to synchronize the two at the end of training: when roberta is fully trained, the downstream structure is also fully trained, and the final model performance is of course the best [3].

a.lr_scheduler: linear_decay Has linear_decayThe two decay strategies of noam_decay and noam_decay can be selected as follows



b.warmup_proportion: 0.1 training warm-up ratio, if set to 0.1, the learning rate will be gradually increased to in the first 10% training steplearning_rate

c.weight_decay: 0.01 weight decay, similar to model regularization strategy, avoiding model overfitting

d.optimizer_name: optimizer name (adam)

D. Different training ways:

Method one:
Directly cut the original data 9 to 1 for training.

Method two:
Add some sentences that contain some infrequent labels to the training data to improve the recall of the most rare labels of the model.
Few labels:

```
#   'B-profession':26,
#   'B-ID':17,
#   'B-contact':16,
#   'B-clinical_event':2,
#   'B-family':27,
#   'B-education':5,
```

```
#   'B-organization':1,
#   'B-others':2
```

Method three:
Classify different labels, copy the data, and then train in batches, and finally merge the results predicted by each model. For example: med_exam, id, contract are one type, time is one type and so on.

Method four:
Find some other Chinese daily data with the same label (such as B-TIME) and add it to the original data to improve the percision and recall of the data.

Ⅴ. Additional attempt and experiment

A. Text length and Prediction:
When the length of the input test data is shorter (close to the length of a sentence), the final f1 score is higher, and when the length of the test data is too long, the f1 score is lower.

B. Entity revising :
The official has given some entity format and existence words. Use these existing words to directly match the original text, and then replace their labels.

time	profession	education	name	med_exam	location
禮拜X	Google	哈佛	小美	XX.XX	劍橋
X月X日	肝膽科	藥學	林醫師	XXXX	醫院
X號	ubereats	台大經濟系	黃醫師	X點X	冰島
X天	保險	高雄大學	王醫師	X百X	美國
X月	工程師	法律系	麻里愛	X十X	巴西
X周	送報紙	成大資工	陳小明	XX度X	紐約
X日	送牛奶	政大	王大華		日本
X年	飲料店	交大	邵介		台北
X點多	水餃店		明翰		士林
X個月	工地		林小姐		新市
X個小時	急診		楷柔		平溪
X個禮拜	ECM				屏東
XX節	金融業				左鎮
前天	銀行員				大陸
昨天	消防員				台北
今天	老師				榮總
明天					沿海地區

C. Combination of models
When the model epoch is low, the recall of the model is high and the precision is low.

When the training epoch of the model rises, it causes the model to further overfit. At this time, the precision of the model is higher but the recall is lower.

Models with higher precision predict less entiy, but each will be more accurate. Models with higher recall have a lot of entiy predictions, but the average accuracy of each is not high.
Therefore, the model with high precision can be used as a benchmark, and those entiy with more recall predictions (entiy with different start poision from the former) can be added to the result of the model with high precision. It is equivalent to combining the advantages of the two models, and the result will be further improved.

D. Experiment

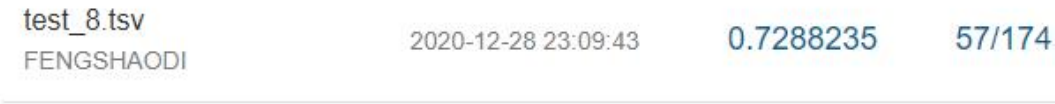
a. Use robeta model:Use only crf and training data max length is 128

b. Use bilstm + crf If the training data max length is 128, result:

c. Set the training data max length is 512, result:

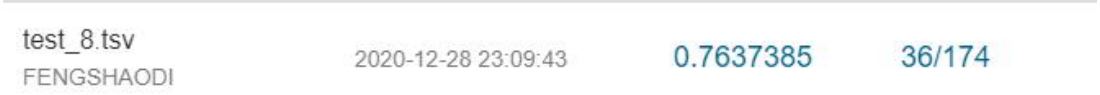
d. Add simplified Chinese text information:

e. Combine the results of the two models:



f. Replace the prediction result directly with the existing vocabulary, but the result drops:

g. Private leaderboard f1 score:



h. Use the bert model (simpler than roberta, f1-score slightly decreased):

i. Use the ernie model because ernie is only trained on simplified Chinese, resulting in bad results:

Train the roberta + bilstm + crf model from 10 epoch to 20 epoch respectively, and the f1score of the model starts to decline after the 17th epoch.

Ⅵ.Reference

[1] Detailed explanation of the principle of Google BERT model
https://zhuanlan.zhihu.com/p/46652512
[2] Introduction to the CRF layer in the most accessible BiLSTM-CRF model
https://zhuanlan.zhihu.com/p/44042528
[3] PaddleHub
https://github.com/PaddlePaddle/PaddleHub
[4] BERT's downstream structure tuning
https://zhuanlan.zhihu.com/p/107378382