介绍文档

参赛队伍：whaido
最终排名：第四名
竞赛地址：https://biendata.com/competition/datagrand/

1．运行环境及参数

运行环境及参数详见代码模型包中 requirements.txt

复现结果可直接参考 2.3 节

2. 模型简介

参赛模型主要采用 bert 预训练+fineuning 模式

BERT-Base, Chinese: Chinese Simplified and Traditional, 12-layer, 768-hidden, 12-heads, 110M parameters

模型相关参数查看 bert_base/bert_config.json

基于谷歌开源的 bert 模型。利用提供的 corpus.txt 进行预训练，再利用 train.txt 进行 fine-tuning，最终对 test.txt 进行预测。

2.1 预训练

数据：
corpus.txt 转化为 corpus_bert.txt(每行中间加空行，划分段落，bert 预训练中会预测上下句，段落内的句子关系较重要，段落是按空行划分的)----该步骤对句子关系有重要意义

Bert 基本信息：
bert_base----bert_config.json, vocab.txt(词汇表由 corpus 生成)，不采用 bert 原始预训练获得的 init_checkpoint，因 corpus 数据是脱敏的

bert_config.json

修改权重的 vocab_size,与 vocab.txt 相同

准备预训练语料

主要利用文件 create_pretraining_data.py

```
python create_pretraining_data.py   --
input_file=datagrand/corpus_bert.txt   --
output_file=tmp/tf_examples.tfrecord   --
vocab_file=bert_base/vocab.txt   --do_lower_case=True   --
max_seq_length=200   --max_predictions_per_seq=30   --
masked_lm_prob=0.15   --random_seed=12345   --dupe_factor=5
```

为了节省预训练时间，max_seq_length 在参考了 corpus.txt 句子长度分布区间后，设置为 200

进行预训练

预训练关键参数

```
train_batch_size=32   --max_seq_length=200   --
max_predictions_per_seq=30   --num_train_steps=1500000   --
num_warmup_steps=25000   --learning_rate=1e-4
```

通过不同参数设置，上述的关键参数数值表现最好，训练结果
global_step = 1500000
loss = 0.5360788
masked_lm_accuracy = 0.8663153
masked_lm_loss = 0.53681165
next_sentence_accuracy = 1.0
next_sentence_loss = 0.0

2.2 Fine-tuning

Fine-tuning 使用的训练数据采用 BIEO 标记，即 Begin, Intermediate, End, Other

Fine-tuning 主要做的任务是 ner，参考 https://github.com/ProHiryu/bert-chinese-ner

BERT_NER.py 主要修改自 run_classifier.py

```
python BERT_NER.py --do_train=True --do_eval=True --
do_predict=False --data_dir=datagrand/ --
bert_config_file=bert_base/bert_config.json --
init_checkpoint=tmp/pretraining_output/ --
vocab_file=bert_base/vocab.txt --output_dir=./ner_result/
```

fine_tuning 训练关键参数

```
num_train_epochs", 20
max_seq_length", 200
learning_rate", 5e-5
```

2.3 预测

```
python BERT_NER.py --do_train=False --do_eval=False --
do_predict=True --data_dir=datagrand/ --
bert_config_file=bert_base/bert_config.json --
init_checkpoint=tmp/pretraining_output/ --
vocab_file=bert_base/vocab.txt --output_dir=./ner_result/ --
max_seq_length=512
```

目前采用方式是将预测文件路径名称写死，如需修改预测文件
可在 BERT_NER.py 修改函数（line208）

```
get_test_examples
```

预测结果存储在 ner_result/ eval_submit_result_BIEO.txt 中

3. 队伍介绍

队伍名称：whaido

4. 参考

https://github.com/google-research/bert
https://github.com/ProHiryu/bert-chinese-ner