中文数据集

2014 人民日报 下载地址 http://www.unopencity.com/project/data-detail/19/

bert 模型

chinese_L-12_H-768_A-12
https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip

运行

先对数据集进行分割并处理：运行 \data\prepro_peopledaily 的注释部分的代码， 将 2014_corpus.txt 分成了 2014_train.txt 2014_dev.txt 2014_test.txt 三部分，存在 \data\raw\LREC 中。

运行 \dataprocess_peopledaily.py 对 2014_train.txt 2014_dev.txt 处理 形成 pd_train.json 和 pd_dev.json 文件 存在\data\raw\dataset\lrec 中。

运行 \bert\cn_punctor.py 代码结构几乎和英文的相同不再赘述

模型结果保存在 \bert\output_pd 中

结果

只训练了 1 个 epoch 在 2014_test.txt 上进行测试

```
-----------------------------------------------
PUNCTUATION      PRECISION RECALL    F-SCORE
,COMMA           83.44     83.39     83.41
.PERIOD          83.36     74.78     78.84
?QUESTIONMARK    78.13     77.46     77.79
-----------------------------------------------
Overall          83.37     81.06     82.20
ERR: 2.27%
SER: 27.5%
```