体系结构及步骤:

一、one数据预处理

二、two 切分数据,映射数据,生成dict.pkl

三、three 数据增强

四、four 模型构建

五、训练

一、one数据清洗

1.查看数据

2.确定类型数量 char,bound,flag等

3.对文章进行B I 标签标注

4.数据预处理,根据标签切分.txt文章

二、特征标注及提取

1.对数据进行分析,判断人工标记的训练集有多少个类(已标记的有15个类)

.ann文件包含人工标记的类别,边界,等信息

.txt文件包含后面需要用的测试集,验证集的整个文本

2.得到标签和下标的映射(给15个类添加BI标签,)

命名实体识别不jieba分词,是一个个字自己分词,bioes,

3.构建切分函数,调用one中切分工具

先把文章切分为句子,.txt---texts

根据需求切分文章,但注意由于有边界特征,所有不能先去掉换行等符号.不然无法对应

4.将切分好的句子储存进一个字典里---data[‘word’]

5. (def process\_text:) # 提取特征

(1)读.txt数据,按照规定的方法切句子,得到切好的句子texts,保存到data[‘word’]里

(2)读texts文件,每个字标注为O,得到tag\_list,

(3)读.ann文件,获取所有的边界信息

(4)读(第一次)tag\_list文件,根据(3),将texts中的15个类别标注为B- I-

(5)根据texts的长度---切分OBI

(6)读第二次tag\_list文件,每个字标记为M

(7)使用jieba库进行词性分词,根据词长标注为SBE

(8)根据 texts的长度---切分 SBE

(9)分别将切好的类别,标签,词性保存进data字典里

(10)使用cnradical库,进行部首和拼音的辨认---直接用切分好的句子texts

(11)将切分好的部首和拼音同样保存进data字典里

(12)标注和提取完成,可以开始清理数据

6.0合并和储存切分的文件

6.分测试集和验证集

7.cpu多线程

8.统计字典

9.映射

(1)根据下标拿到所有切分好的句子

(2)对每一个特征进行处理

(3)转id

三、数据增强

1.构建映射函数,调用two中映射规则进行id映射,

2.数据增强

(1)两句话两句话拼接

(2)三个句子进行拼接

3.提供id映射文件接口:

(1)dict.pkl文件接口

4.数据分批

(1)给dict.pkl进行数据分批

(2)提供批次接口

四、 four 模型构建

1.接收class Model (3)的输出值,生成最终向量送入lstm

2.构建bi-lstm双向循环神经神经网络

①神经网络输出:

return output/lengths

4.class Model

(1)定义参数:

①包括接收始祖数据dict.pkl/生成矩阵/lstm参数/

(2)构建全流程

①定义接收数据-计算lstm输出值-计算loss-优化器-保存模型

(3)定义接收始祖数据函数

①接收始祖数据批次样本,来自three

return给3-(1)的外部函数

(4)crf

五、训练

1.数据准备

2.读取数据

通过批次数据接口,载入批次数据 class BatchManager(object):

3.载入模型

通过2读入的数据,载入模型开始训练