seqlabel-crf命名实体识别

1、环境

- linux系统
- python3.6

2、文件介绍

（1）corpus_get.py文件

该文件是将领域内的数据，通过已有的crf模型（基于crf训练好的命名实体模型，训练数据为人民日报）和特定领域词典，对领域数据进行标注。

其中一些文件路径需要修改，词典路径，领域语料路径，命名实体模型路径，生成文件路径。

（2）data.py文件

该文件是文本处理文件，包括加载语料数据，加载模型。

（3）feature.py文件

该文件是一些特征分析阶段，包括分词和词性标注工具的选择，前缀树的构建，用于匹配词典，以及特征提取器。

（4）inference.py文件

该文件的作用主要是用于读取训练得到的序列标注模型，同时也是命名实体的接口。

（5）metrics.py文件

该文件用于对数据进行标签。

（6）opts.py文件

主要是模型训练过程中的一些参数设置。

（7）train.py文件

基于crf的命名实体标注训练模型。

（8）utils.py文件

该文件包括路径的检测，特征的获取，crf模型参数的架子啊，标签的标记，结果的格式化等一些文件。

2、模型的训练

使用train.py文件进行模型训练，其训练参数可以在opts文件中修改，直接python train.py文件即可训练模型。

或者使用如下方式训练：

```
python3.6 script/train.py -max_iterations 100 \
  -template_path w2 \
  -k_fold 5 \
  -n_jobs 5 \
  -drop_vocab_pro 0.5 \
  -model_name $tag \
  -vocab_path $file \
  -train_data_path out/nerdata/$tag.nerdata \
  -save_path out/crf_save
```

在训练的时候，可以输入参数。

2、模型的使用

使用inference.py文件，调用其中函数，进行命名实体识别。其方法如下：

```
model_path = '/data/menghao/yun2space/nlp_code/seqlabel/crf_save1/crf_$5_test_f1_73.77.model'
test = InferenceModel(model_path)
test1 = InferenceModel('/data/menghao/yun2space/nlp_code/seqlabel/crf_save_question/crf_$5_test_f1_80.96.model')
print(test(['小规模纳税人为什么可以底销项税不能底进项税']))
```