

哈尔滨工业大学 2018 年秋季学期

计算机学院本科生  
“中文信息处理”课  
实验报告（二）

报 告 题 目： 中文名实体识别

姓 名： 肖松

学 号： 1160300527

学 生 专 业： 计算机科学与技术

任 课 教 师： 刘秉权

2018 年 12 月 15 日

# 报 告 正 文

## 1.实验内容

- (1) 使用任意方法实现任一类中文名实体识别;
- (2) 给定足够规模的测试文本,在其上标注至少 100 个实体识别结果;
- (3) 计算出实体识别的准确率和召回率,并给出计算依据;
- (4) 针对识别结果中存在的问题给出具体分析;
- (5) 提交实验报告,给出详细实验过程、结果和结论;提交源代码、可执行程序 and 程序中使用的其他资源。

## 2. 实验要求和目的

1. 自己构造必要的知识库;
2. 自己准备足够规模的语料;
3. 编程环境、汉字编码不限。

## 3.实验环境

win10

Python 3.7.0

CRF++-0.58

## 4.程序主要算法

条件随机场 (CRF) :

条件随机场定义:令  $G = (V, E)$  表示一个无向图, 任意一个节点  $v$  对应一个随机变量  $Y_v$ , 因此  $Y = \{Y_v | v \in V\}$ ,  $Y$  中元素与无向图  $G$  中的顶点一一对应。当在条件  $X$  下, 随机变量  $Y_v$  的条件概率分布服从图的马尔可夫属性:  $p(Y_v | X, Y_w, w \neq v) = p(Y_v | X, Y_w, w \sim v)$ , 其中  $w \sim v$  表示  $(w, v)$  是无向图  $G$  的边。这时我们称  $(X, Y)$  是一个条件随机场。

我们可以推出最终条件随机场的条件概率为:

$$P(Y|X) = \frac{1}{Z(X)} \exp(\sum_{k=1}^K \lambda_k F_k(Y, X)) \text{ 其中,}$$

$$Z(X) = \sum_y \exp(\sum_{k=1}^K \lambda_k F_k(Y, X))$$

训练 CRF 主要就是要训练特征函数的权重（即  $\lambda_k$ ），对于训练集  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ ，采用极大似然估计法计算权重参数，条件概率的对数似然函数为：

$$L(\lambda) = \sum_{x,y} \tilde{p}(x, y) \sum_{i=1}^n (\sum_{k=1}^K \lambda_k f_k(y_{i-1}, y_i, x, i)) - \sum_x \tilde{p}(x) \log Z(x)$$

其中  $\tilde{p}(x, y)$  为训练样本集中  $xy$  的经验概率，它等于  $xy$  同时出现的次数除

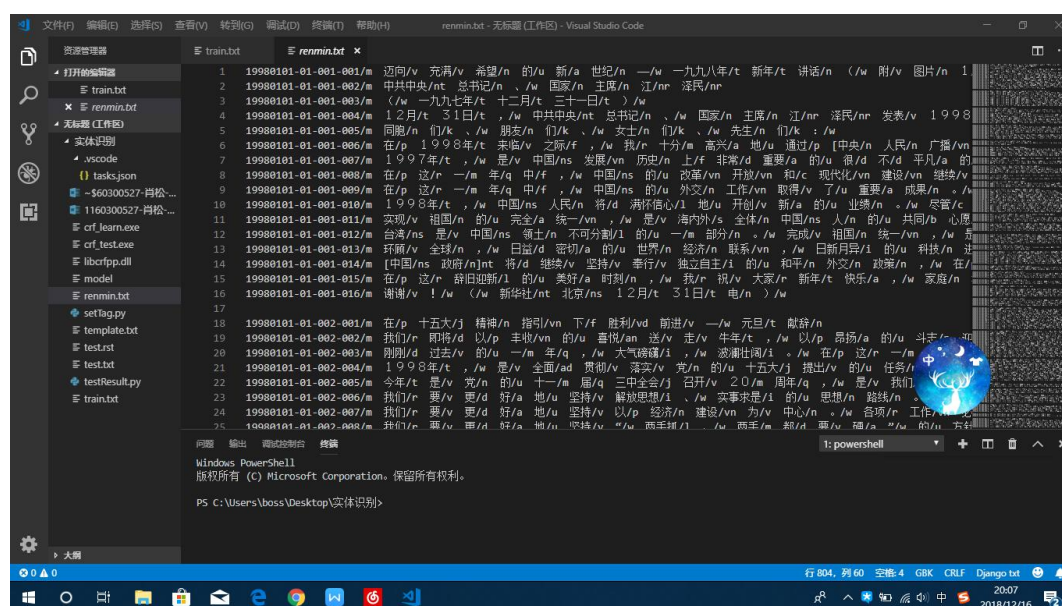
以样本空间容量； $\tilde{p}(x)$  为训练样本集中  $x$  的经验概率，它等于  $x$  出现的次数

除以样本空间容量。

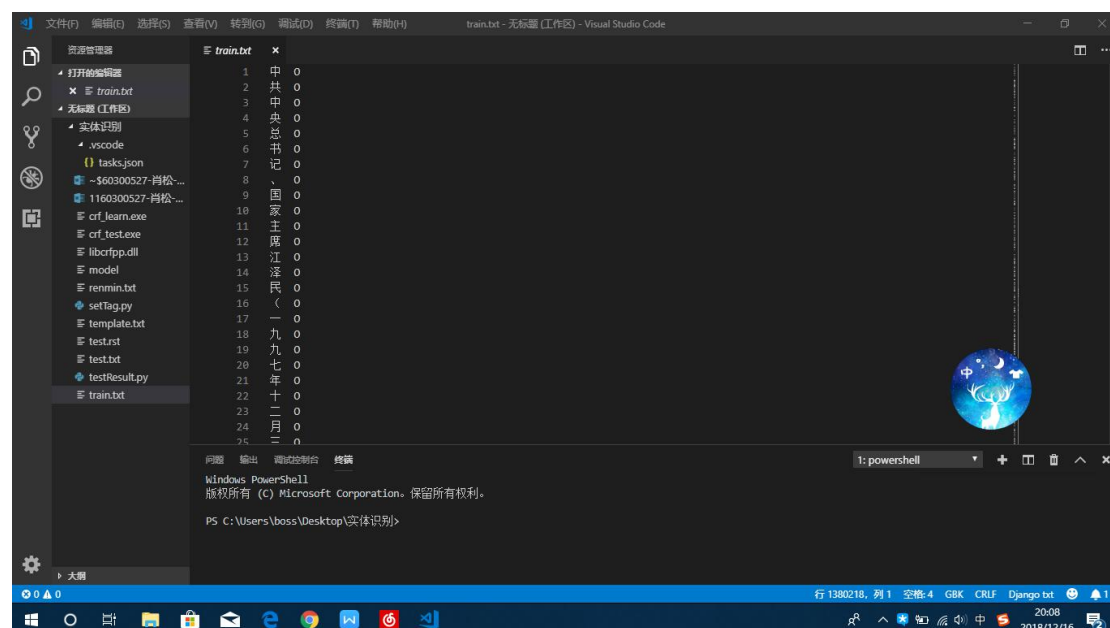
只需要对似然函数求导为 0，进行优化求解即可

## 5. 实验过程

1) 对语料进行预处理。在本次实验中，采用的 1998 年人民日报分词数据集，格式如下：



需要将其处理为 crf++ 需要的训练数据格式：



其中，O 标识不是查找的中文名，S 标识该中文名只有一个字符，B 表示中文名的第一个字符，M 标识中文名的非开始/结束字符，E 表示中文名的最后一个字符

在处理过程中，需要将语料转换为对应的标注，并将形如[华北/ns 电管局/n]nt 进行合并，合并为华北电管局/nt。同时，从中去除部分数据作为测试集数据。在本次实验中，我抽取了 25%的数据用作测试。

具体实现见代码 setTag.py

2) 编写特征模板文件，该文件描述了用来训练以及进行测试的特征。

```
# Unigram
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-2,0]
U06:%x[0,0]/%x[-1,0]
U07:%x[0,0]/%x[1,0]
U08:%x[-1,0]/%x[-2,0]
U09:%x[1,0]/%x[2,0]
```

```
U10:%x[-1,0]/%x[1,0]
```

```
# Bigram
```

```
B
```

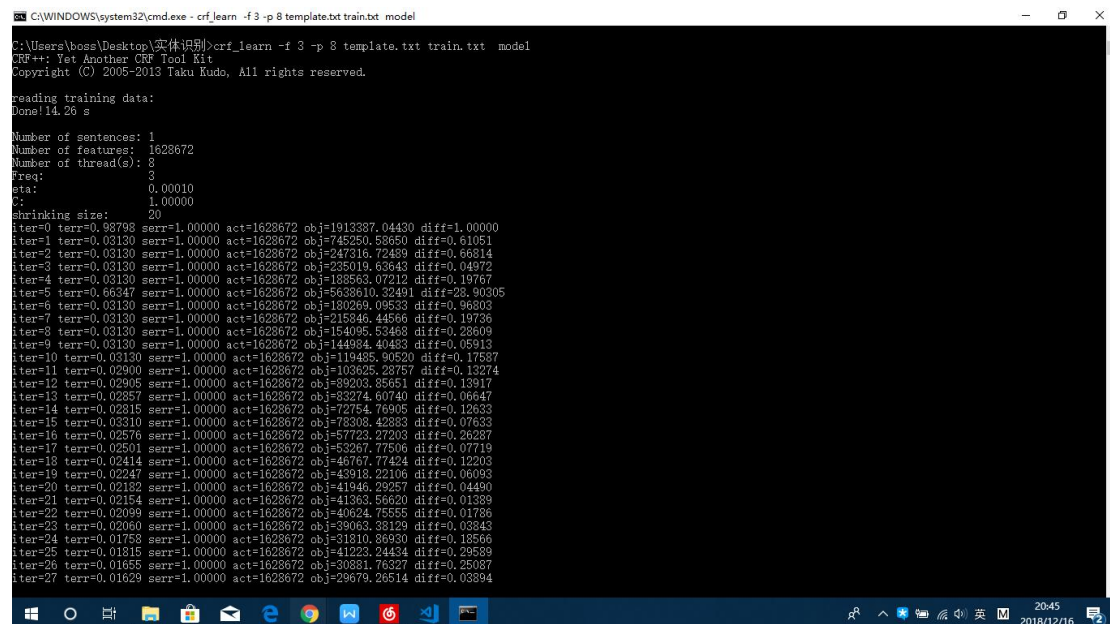
3) 执行 `crf_learn -f 3 -p 8 template.txt train.txt model` 命令，完成模型的学习，模型文件为 `model`。

其中，各参数含义为：

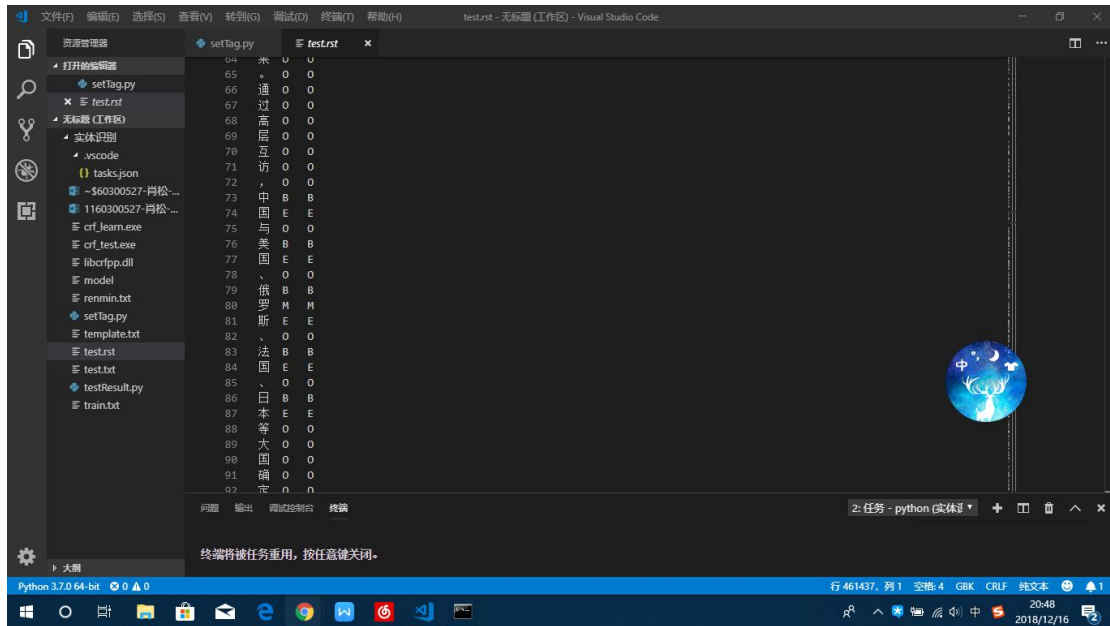
-f, -freq=INT 使用属性的出现次数不少于 INT(默认为 1)

-p, -thread=INT 线程数(默认 1)，利用多个 CPU 减少训练时间

训练过程如下：



4) 通过 `crf_test -m model test.txt > test.rst` 命令完成对测试集数据的标注，标注后的 `test.rst` 文件如下：



第一列为进行标注的中文字符，第二列为按照模型进行标注的结果，第三列为人工标注的结果

6. 根据对测试集的标注结果计算模型的各项参数：Precision, Recall, F1-Score.

```
1. def test(path):
2.     file = open(path)
3.     tags = 0#被标记字的总数
4.     right = 0#标记正确的数目
5.     precisionTags = 0#所有标记出来的数目
6.     recallTags = 0#应该被标记的字的数目
7.     for line in file:
8.         line = line.strip()
9.         if(len(line)==0):
10.            continue
11.         tags += 1
12.         _word, tag_real, tag_point = line.split()
13.         if tag_point == tag_real and tag_point != '0':
14.             right += 1
15.         if tag_point != '0':#被标注的数据
```

```

16.         precisionTags += 1
17.         if tag_real != '0':#应该被标注的数据
18.             recallTags += 1
19.     precision = float(right)/precisionTags
20.     recall = float(right)/recallTags
21.     f1Score = 2*precision*recall/(precision+recall)
22.     print("precision:%f, recall:%f, F1Score:%f\n"%(precision,recall,f1Score))
23.
24.
25. test('test.rst')

```

## 6.实验结果

测试内容：

Precision:所有标注正确的词的数量/所标注出来的词的总数

Recall:所有标注正确的词的数量/应该标注的词的数量

F1-Score:  $F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$

执行结果如下：

```

> Executing task in folder 实体识别: C:/Users/boss/AppData/Local/Programs/Python/Python37/python c:\Users\boss\Desktop\实体识别\testResult.py <
precision:0.907090, recall:0.841584, F1Score:0.873110

```

分析 test.rst 文件我们可以看出：

1) 人工标注为地名而模型未能成功标注的结果：

， O O

中 O B

国 O E

驻 O O

阿 O B

联 O M

酋 O M

使 O M

馆 O E

内 O O

对于中国，阿联酋这两个地名，模型未能成功识别

2) 人工未标注为地名而模型识别为地名的情况:

(1)

增 O O

加 O O

从 O O

巴 B O

自 M O

治 M O

区 E O

进 O O

口 O O

(2)

为 O O

什 O O

么 O O

维 B O

也 M O

纳 E O

音 O O

乐 O O

会 O O

在人工标注中，巴自治区以及维也纳未被标注为地名，而在模型中，将其识别为地名

分析上述结果，我们可以看出：CRF 算法通过对地名前后的词语进行提取，将其纳入模型的考虑范围中。在判断一个词是否为地名时，主要通过对前后的特征进行匹配，从而进行判断。因此，当前后不具有明显特征时，不能够对特征进行有效提取。同时，对于



具有多重含义的词进行识别时，可能会识别错误

## 7.实验结论和体会

CRF 可以有效对中文名词进行识别，其主要通过对特征进行提取，生成模型，然后用模型来进行匹配。由于本次实验中，仅通过对词进行特征提取，效果没有达到最佳，为了提升识别的效果，可以将语料的词性作为特征进行提取或者扩充训练的语料。同时，可以对语料进行事先整理，提升识别效果