# 中文公司名称识别 实验 指导手册

## CRF模型训练

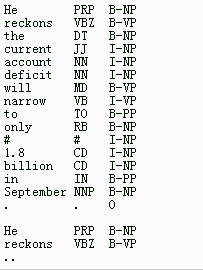
常用软件下载：

CRF++：<https://github.com/zhangsheng/crfpp>

Wapiti：<https://wapiti.limsi.fr/>

* 1. 训练数据格式介绍

下面是一个训练文件的例子：

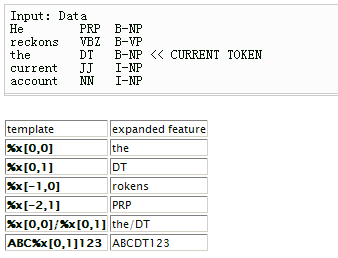


训练文件由若干个句子组成（可以理解为若干个训练样例），不同句子之间通过换行符分隔，上图中显示出的有两个句子。每个句子可以有若干组标签，最后一组标签是标注，上图中有三列，即第一列和第二列都是已知的数据，第三列是要预测的标注，以上面例子为例是，根据第一列的词语和和第二列的词性，预测第三列的标注。

测试文件与训练文件格式自然是一样的，可以没有最后一列。

* 1. 模型 特征模板介绍

模板文件中的每一行是一个模板。每个模板都是由%x[row,col]来指定输入数据中的一个token。row指定到当前token的行偏移，col指定列位置。

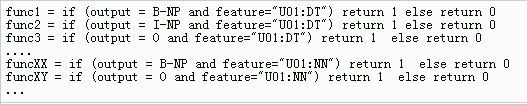


由上图可见，当前token是the这个单词。%x[-2,1]就就是the的前两行，1号列的元素（注意，列是从0号列开始的），即为PRP。

有两种类型的模板，模板类型通过第一个字符指定。即U和B。

**Unigram template: first character, 'U'**

当给出一个"U01:%x[0,1]"的模板时，CRF++会产生如下的一些特征函数集合(func1 ... funcN) 。



这几个函数我说明一下，%x[0,1]这个特征到前面的例子就是说，根据词语(第1列)的词性(第2列)来预测其标注(第3列)，这些函数就是反应了训练样例的情况，func1反映了“训练样例中，词性是DT且标注是B-NP的情况”，func2反映了“训练样例中，词性是DT且标注是I-NP的情况”。

模板函数的数量是L\*N，其中L是标注集中类别数量，N是从模板中扩展处理的字符串种类。

**Bigram template: first character, 'B'**

这个模板用来描述二元特征。这个模板会自动产生当前output token和前一个output token的合并。注意，这种类型的模板会产生L \* L \* N种不同的特征。

* 1. crf模型命令介绍
     1. CRF++ 使用
        1. 训练程序

命令行：

**% crf\_learn template\_file train\_file model\_file** 这个训练过程的时间、迭代次数等信息会输出到控制台上（感觉上是crf\_learn程序的输出信息到标准输出流上了），如果想保存这些信息，我们可以将这些标准输出流到文件上，命令格式如下：

**% crf\_learn template\_file train\_file model\_file >> train\_info\_file**

有四个主要的参数可以调整：

**-a CRF-L2 or CRF-L1**

规范化算法选择。默认是CRF-L2。一般来说L2算法效果要比L1算法稍微好一点，虽然L1算法中非零特征的数值要比L2中大幅度的小。

**-c float**

这个参数设置CRF的hyper-parameter。c的数值越大，CRF拟合训练数据的程度越高。这个参数可以调整过度拟合和不拟合之间的平衡度。这个参数可以通过交叉验证等方法寻找较优的参数。

**-f NUM**

    这个参数设置特征的cut-off threshold。CRF++使用训练数据中至少NUM次出现的特征。默认值为1。当使用CRF++到大规模数据时，只出现一次的特征可能会有几百万，这个选项就会在这样的情况下起到作用。

**-p NUM**

如果电脑有多个CPU，那么那么可以通过多线程提升训练速度。NUM是线程数量。

带两个参数的命令行例子：

**% crf\_learn -f  3 -c 1.5 template\_file train\_file model\_file**

**例如：**

crf\_learn -p2 -f 3 -c 4.0 template train.data model

crf\_test -m model test.data

* + - 1. 测试程序

命令行：

**% crf\_test -m model\_file test\_files**有两个参数-v和-n都是显示一些信息的，-v可以显示预测标签的概率值，-n可以显示不同可能序列的概率值，对于准确率，召回率，运行效率，没有影响，这里不说明了。

与crf\_learn类似，输出的结果放到了标准输出流上，而这个输出结果是最重要的预测结果信息（测试文件的内容+预测标注），同样可以使用重定向，将结果保存下来，命令行如下。

**% crf\_test -m model\_file test\_files >> result\_file**

* + 1. wapiti 使用
       1. 训练程序

命令行：

**% wapiti train [options] [input data] [model file]**

**主要参数：**

**-a 选择算法 l-bfgs、sgd-l1、bcd、rprop、rprop+、rprop-等六种算法可供选择**

**-p 选择模板文件**

**-m 可选择模型文件用于重新训练**

**-t 使用的线程数**

**-i 指定最大迭代次数**

**-1 指定L1回归惩罚系数，该系数为0时只使用L2回归**

**-2 指定L2回归惩罚系数，该系数为0时只使用L1回归**

* + - 1. 测试程序

命令行：

**%wapiti label [options] [input data] [output data]**

**主要参数：**

**-m： 模型文件**

## 中文公司名称识别 实例介绍

* 1. 采用的特征介绍

实例中仅使用了文本特征和词性特征。

文本特征：即使用词语本身作为一列特征，文本中会包含大量用于指示公司机构名称的信息，例如“公司”、“部门”等等词语都可以用来表示这是一个公司名或者机构名。

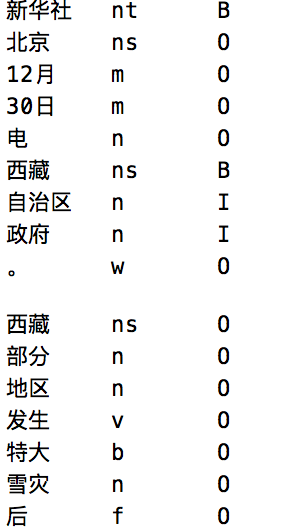
词性特征：词性包含名词、动词、形容词等等，而名词又可细分为人名、地名、其他专有名词等等，这些词性对于我们的公司名称识别也可以起到非常显著的帮助。

其他特征：可以根据具体任务的不同采用各种不同的特征，例如：语法特征、情感特征、词长等等，甚至可以不进行分词，将每个字作为一个token来进行处理。

* 1. 特征如何转化为所要求的训练数据格式

根据token粒度的不同、特征选取的不同可以生成不同的训练数据，但是训练数据的格式却是始终相同的。每一行表示一个token，每一列表示一维特征，不同列之间用制表符（\t）隔开，最后一列表示标注的结果。不同句子之间用空行隔开。

下面是一个训练语料的样例：



其中“新华社北京12月30日电西藏自治区政府。”为一个句子，而每一行为该句子的一个词，共分为三列，第一列为当前文本，第二列为词性，第三列为标注结果。

* 1. 模板选择

不同的模板训练出来的模型结果可能完全不同，我们先选用仅使用词语的模板，看一下实验的效果。

针对该任务采用的特征如下：

|  |  |
| --- | --- |
| U01:%x[-3,0] | 当前词前面第三个词 |
| U02:%x[-2,0] | 当前词前面第二个词 |
| U03:%x[-1,0] | 当前词前面第一个词 |
| U04:%x[0,0] | 当前词 |
| U05:%x[1,0] | 当前词后面第一个词 |
| U06:%x[2,0] | 当前词后面第二个词 |
| U07:%x[3,0] | 当前词后面第三个词 |

使用该模板得到的结果如下所示，其中第三列为标注结果，第四列为测试结果:

上海 ns B B

交通 n I I

大学 n I I

明年 t O O

将 d O O

迎来 v O O

120周年 m O O

庆 vg O O

。 w O O

许多 m O O

企业 n O O

校友 n O O

纷纷 d O O

慷慨解囊 i O O

， w O O

例如 v O O

著名 a O O

的 uj O O

联想 nz B B

集团 n I I

董事长 n O O

杨元庆 nr O O

， w O O

腾 v B O

讯 ng I O

公司 n I O

， w O O

红海 ns B O

公司 n I O

等 u O O

校友 n O O

我们在词语模板的基础上加入词性的模板与词语和词性的合并，新的模板如下：

|  |  |
| --- | --- |
| U01:%x[-3,0] | 当前词前面第三个词 |
| U02:%x[-2,0] | 当前词前面第二个词 |
| U03:%x[-1,0] | 当前词前面第一个词 |
| U04:%x[0,0] | 当前词 |
| U05:%x[1,0] | 当前词后面第一个词 |
| U06:%x[2,0] | 当前词后面第二个词 |
| U07:%x[3,0] | 当前词后面第三个词 |
| U08:%x[-3,1] | 当前词前面第三个词词性 |
| U09:%x[-2,1] | 当前词前面第二个词词性 |
| U10:%x[-1,1] | 当前词前面第一个词词性 |
| U11:%x[0,1] | 当前词词性 |
| U12:%x[1,1] | 当前词后面第一个词词性 |
| U13:%x[2,1] | 当前词后面第二个词词性 |
| U14:%x[3,1] | 当前词后面第三个词词性 |
| U15:%x[-2,1]/%x[-1,1]/%x[0,1] | 当前词前面第二个词词性+当前词前面第一个词词性+当前词词性 |
| U16:%x[-1,1]/%x[0,1]/%x[1,1] | 当前词前面第一个词词性+当前词词性+当前词后面第一个词词性 |
| U17:%x[0,1]/%x[1,1]/%x[2,1] | 当前词词性+当前词后面第一个词词性+当前词后面第二个词词性 |
| U18:%x[-1,1]/%x[0,1] | 当前词前面第一个词词性+当前词词性 |
| U19:%x[0,1]/%x[1,1] | 当前词词性+当前词后面第一个词词性 |
| U20:%x[0,0]/%x[0,1] | 当前词+当前词词性 |
| B | bigram特征 |

使用该模板得到的结果如下所示：

上海 ns B B

交通 n I I

大学 n I I

明年 t O O

将 d O O

迎来 v O O

120周年 m O O

庆 vg O O

。 w O O

许多 m O O

企业 n O O

校友 n O O

纷纷 d O O

慷慨解囊 i O O

， w O O

例如 v O O

著名 a O O

的 uj O O

联想 nz B B

集团 n I I

董事长 n O O

杨元庆 nr O O

， w O O

腾 v B O

讯 ng I O

公司 n I O

， w O O

红海 ns B B

公司 n I I

等 u O I

校友 n O I

可以看到新的模板对于红海公司进行了抽取，但是抽取结果仍然不是特别准确，这就需要我们标注更多的训练语料同时进一步的调整模型。

* 1. 存在问题分析与思考

常见问题：

* 分词工具的选用？

不同的分词工具之间可以会有细微的差别，但是对于最终结果影响不大，但是要注意训练语料和测试语料选用的分词工具要保持一致。

* 特征选取？

面对不同的任务要选取不同的特征，可以通过实验的方法来对比不同特征的不同表现，从而选取最优的特征。

* 特征模板如何建立

相同的特征文件不同的特征模块可能导致的结果也会千差万别。