cs224n

lecture 7:tensorflow 入门

1. Tensorflow是一个图计算模型，中心思想是将数值运算以图的形式描述，图的节点是运算的封装，输入输出可为任意数量。图的边是张量(n维数组)。
2. placeholder：类似于形参，在run时传递参数进去

constant：常量，创建时初始化

variable：变量，需要初始化

1. session机制：启动图计算通过创建session对象实现
2. 训练方法：定义损失、定义optimizer

了解有限，后面做Bilstm+crf实验再仔细看看文档和教程

CRF++ 实验过程

1. 一开始学习模型出现的问题

"reading training data:tagger.cpp(399) [feature\_index\_->buildFeatures(this)] 0.00s" 建立标签时出现问题，即最后一列不能做为template特征，于是更改了默认的unigram的template格式，去掉了包含第2列的部分

1. 然后训练并测试了官方给出的样例 没啥问题
2. 下载了北大人民日报语料库，预处理后用10个左右简单的unigram做为template训练（某个example的template），然而测试输出全为single word，猜测是自己的特征函数设置出了问题。

查了特征函数要求：

①%x[row,col]代表相对当前词偏移一定行列数后的特征词

②%x[0,0]/%x[0,1]这种表示两部分结合做为特征

③最后的标签列不能做为特征

④其他template 窗口大小n大概在2-3

于是我把特征函数改成最简单的几个(-1,0,1)组合，训练后仍全为single word，网上也没有查到可能错误原因，自己猜测原因：参数没设好过拟合、特征函数没设好、分词训练样本没有z隐变量，测试时需要填充

1. 在给定的4个example上先做实验，看一下参数的影响

默认参数为：

-f 1 使用训练数据中至少f次出现的特征

训练算法 CRF-L2

-c 1 代价参数

-m 10k LBFGS最大迭代次数

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 输出标记数 | 列数(=隐变量个数+2) | 准确率 | 召回率 | F值 |
| BaseNP | 3 | 3 | 95.57% | 100% | 97.73% |
| Chunking | 14 | 3 | 92.99% | 100% | 96.37% |
| SEG | 2 | 3 | 82.16% | 100% | 90.21% |

前两个准确率好高...这才训练了没几秒..

以下用第一个、第三个来调参：

|  |  |
| --- | --- |
| SEG | 准确率 |
| 默认 | 82.16% |
| -a CRF-L1 | 79.41% |
| -c 100 | 81.86% |
| -c 10000 | 82.26% |
| -c 0.01 | 71.76% |
| -f 3 | 81.55% |
| -f 5 | 77.47% |

|  |  |
| --- | --- |
| BaseNP | 准确率 |
| 默认 | 95.57% |
| -a CRF-L1 | 95.89% |
| -c 100 | 95.46% |
| -c 10000 | 95.41% |
| -c 0.01 | 94.40% |
| -f 3 | 95.09% |
| -f 5 | 95.52% |

实验结论：

1. CRF-L2比CRF-L1更好，默认为L2
2. –c 代价常数增大并没有出现过拟合，甚至可能提高准确率
3. –f 至少选择出现f次的值增大，选取属性变少，效果可能会变差，默认为1效果最好，但是训练轮数大概比3或5增加20%-50%
4. 于是回原来北大语料库模型调参，做了如下尝试：

（1）c使用0.01或1000，输出仍全为single word，似乎不是拟合程度的问题

（2）f使用训练数据中至少f次出现的特征，网上其他处理这些大语料库的资料f值均为3，若调为1，特征太多，训练轮数可能会增加一半（由以上实验经验），仍未尝试

（3）修改特征函数模板为[-1,0,1]组合，只留下8个最简单形式，输出仍为single word，目前还未查阅到特征函数的具体规定，似乎是依经验设定，符合基本要求即可

1. 接下来可能去看一下源码，联系一下博主看看问题可能出在哪儿。

个人目前想法：官方给出的的例子至少有一个中间特征，类似于 词、词类型、词性分类。而北大人民日报语料库只有字+词性分类，测试时自行需要填充一列，可能会影响结果。然而博客文章填充了也出了结果...无奈...

今晚仔细检查了程序发现问题了，句与句之间要有一行空行分隔开，原来全为single word的原因估计是训练出来的模型驾驭不了1MB长的句子哈哈哈

测试结果是：

=== SUMMARY:

=== TOTAL INSERTIONS: 1400

=== TOTAL DELETIONS: 1280

=== TOTAL SUBSTITUTIONS: 2417

=== TOTAL NCHANGE: 5097

=== TOTAL TRUE WORD COUNT: 106873

=== TOTAL TEST WORD COUNT: 106993

=== TOTAL TRUE WORDS RECALL: 0.965

=== TOTAL TEST WORDS PRECISION: 0.964

=== F MEASURE: 0.965

=== OOV Rate: 0.026

=== OOV Recall Rate: 0.650

=== IV Recall Rate: 0.974

### tag2word.utf8 1400 1280 2417 5097 106873 106993 0.965 0.964 0.965 0.026 0.650 0.974

参数：-c 4 –f 3

训练时间：6900s

准确率、召回率、F值都为0.965，很优秀了

其他：制作讨论班PPT：内容为混合模型与EM算法