作业2：文本序列标注问题研究（处理文本序列）

实验前的准备工作：

1. 复习序列标注问题一章
2. 理解隐式马尔科夫和深度神经网络处理序列标注问题的原理、方式
3. 准备实验环境（本次实验采用Pytorch和numpy）

提交实验报告文档，包含2 个主要部分：

**复习部分：**

问题的主要展现形式：（1）文本序列的词性标注；（2）中文切词；（3）命名实体识别（人名，地名，机构名）

文本序列标注问题定义：对一个序列中的每一个**观测量**，给出一个对应的**标签**。

**观测量、标签**不是**独立分布**的，他们之间存在**结构依赖关系**，会相互影响；这种结构依赖关系表现为较强的局部结构依赖，但是局部结构依赖会传导为全局依赖。

全局序列标注模型：能够对序列标注问题进行一个全局的建模，并确定一个全局最优的决策。从而解决局部最优导致的问题。

常用的全局序列标注模型：

（1）隐式马尔可夫模型；

（2）最大熵隐式马尔可夫模型；

（3）条件随机场（CRF）；

HMM回答了以下的问题：

Evaluation: 我们观察的序列有多少出现的可能性；

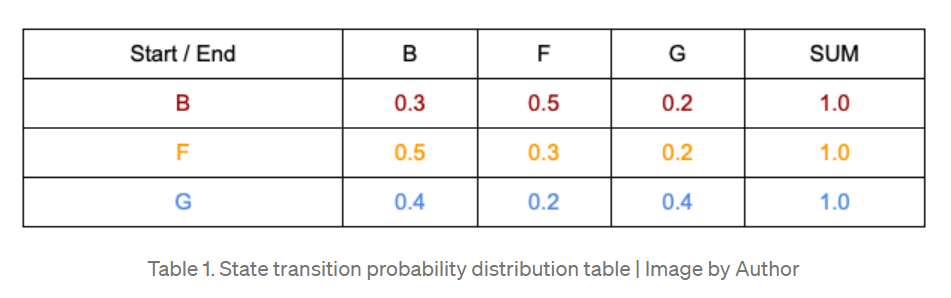
Decoding:

Learning:

HMM model:

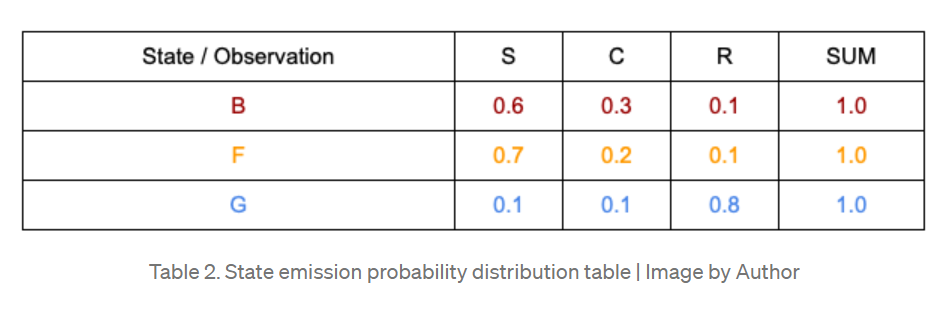
Hidden state and Observation symbols:

State transition probability distribution:



隐藏状态之间的转移

State emission probability distribution:



Observation sequence:

隐藏状态之间的迁移：

一旦一个系统可以作为HMM被描述，就可以用来解决三个基础问题。

1. 评估：给定HMM，求某个观察序列的概率；2.解码：3. 学习；

**实验部分：**

### 基于隐式马尔科夫模型的序列标注

**下面的结果是利用ner\_train\_rate数据进行训练；在ner\_dev.key上进行test；评价指标是利用seqeval包进行计算的；**

2-阶隐式马尔科夫模型的标注效果

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | P | R | F1 |
| LOC | 0.75 | 0.59 | 0.66 |
| MISC | 0.69 | 0.53 | 0.59 |
| ORG | 0.52 | 0.43 | 0.47 |
| PER | 0.65 | 0.49 | 0.55 |
| Overall | 0.67 | 0.52 | 0.58 |

3-阶隐式马尔科夫模型的标注效果

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | P | R | F1 |
| LOC | 0.88 | 0.70 | 0.78 |
| MISC | 0.83 | 0.69 | 0.75 |
| ORG | 0.61 | 0.48 | 0.54 |
| PER | 0.76 | 0.59 | 0.67 |
| Overall | 0.77 | 0.61 | 0.68 |

从上表可以看到，3-阶隐式马尔科夫模型是全部好于2-阶隐式马尔科夫模型的。

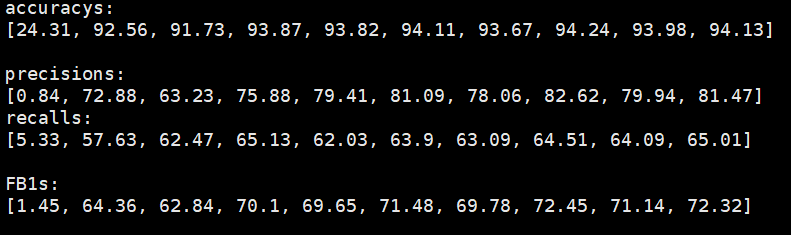
### 基于循环神经网络的序列标注

下面的神经网络的训练集为ner\_train.data; 测试集为ner\_dev.key; 深度神经网络我选择了在这一块SOTA的工作LSTM+CRF；训练均为10epoch；最后我就只展示最终的整体标注效果。

具体代码可见我的github: <https://github.com/nightzero123/SequenceTagging>

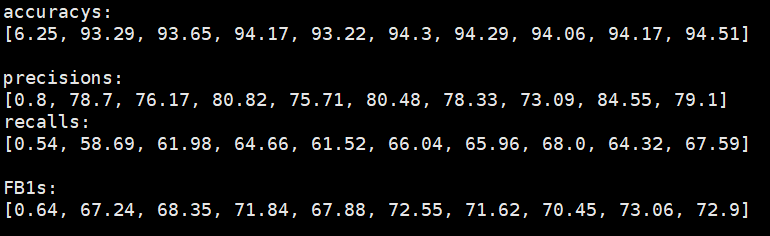
LSTM + CRF的标注效果

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | P | R | F1 |
| Overall | 81.47 | 65.01 | 72.32 |



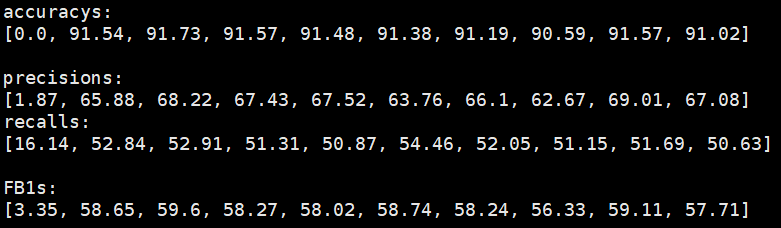
BiLSTM + CRF的标注效果

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | P | R | F1 |
| Overall | 79.1 | 67.59 | 72.9 |



3层BiLSTM + CRF的标注效果

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | P | R | F1 |
| Overall | 67.08 | 50.63 | 57.71 |



在我的实验中，由上面的结果可知，3层BiLSTM就已经过拟合了，而我们双向BiLSTM的结果也没有单项LSTM的效果好。