**基于置信度的自学习实验**

**---------------------2019/05/21-----------------**

**【方法】**

选择策略：置信度

停止策略：在开发集上FB1值收敛

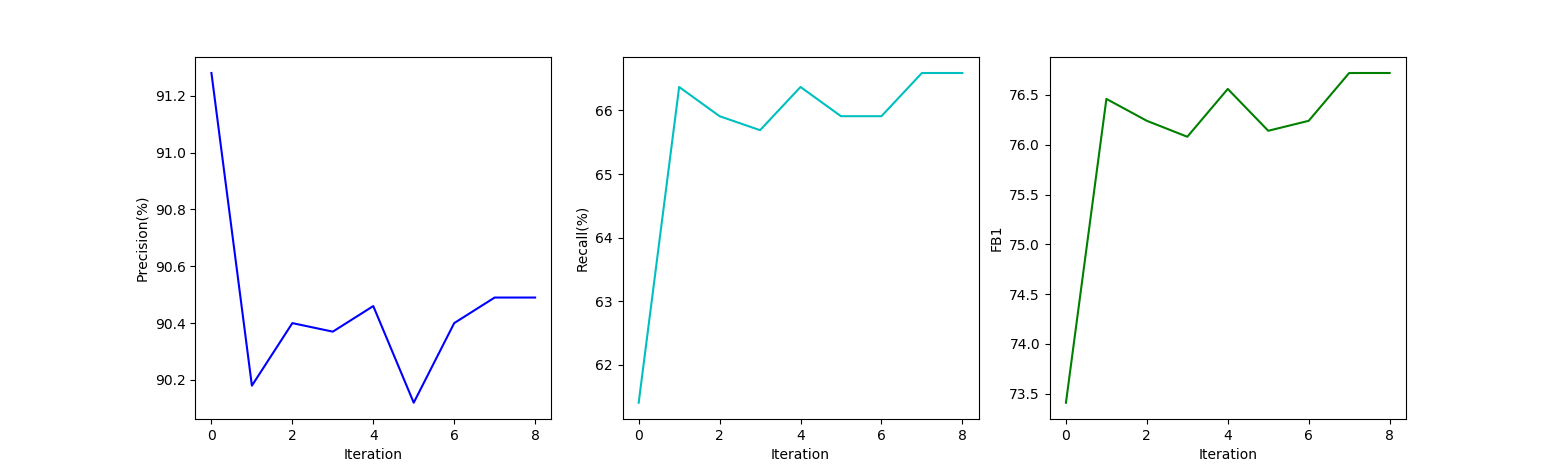
**【参数】**

初始训练集：2.91MB 开发集：1.04MB

选择置信度：0.95 停止差异值：0.005

**【结果】**

迭代次数：10 终止训练集：5.02MB 模型选择的数据：3.11MB



**【结论】**

1. 合理的参数设置可以保证基于置信度的自学习模型上升且收敛；
2. 未加入命名实体特征的基线模型识别效果较差。

**---------------------2019/11/20-----------------**

**【策略】**

（1）阈值；（2）TopN；

**【结论】**

（1）首次筛选的数据量较大且质量很差，导致后续学习曲线下降。

（2）选择策略只能改善模型识别出的实体质量，但对于模型未能识别的实体，一旦把数据加入到训练集，将造成召回率很低。

基于多模型投票：

2个学习器：一致加入，不一致判断置信度

3个学习器：少数服从多数

特征增强：

词向量→须测试覆盖率。

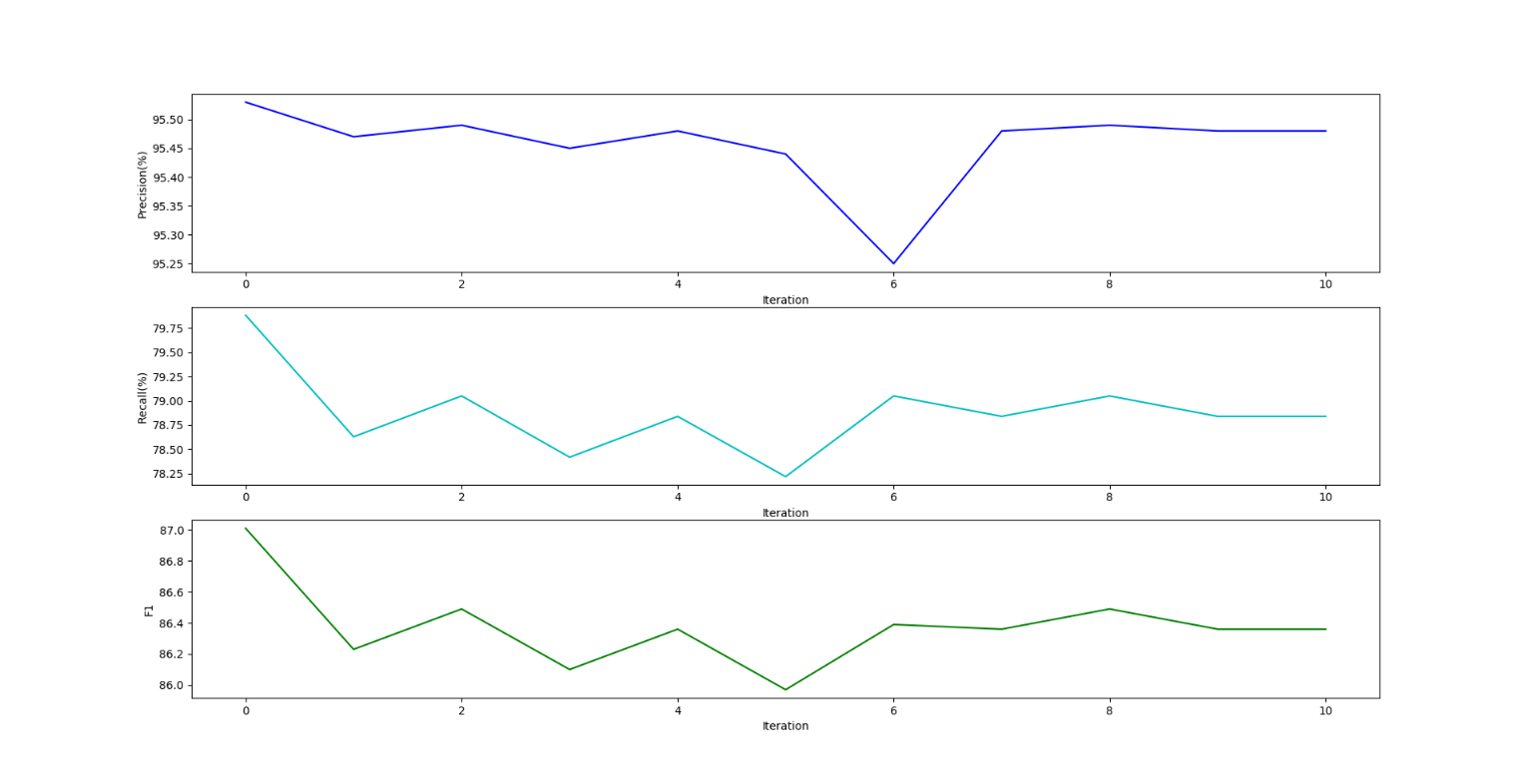
提高召回率：

从现有的训练集直接提取实体表，直接匹配待标注数据。

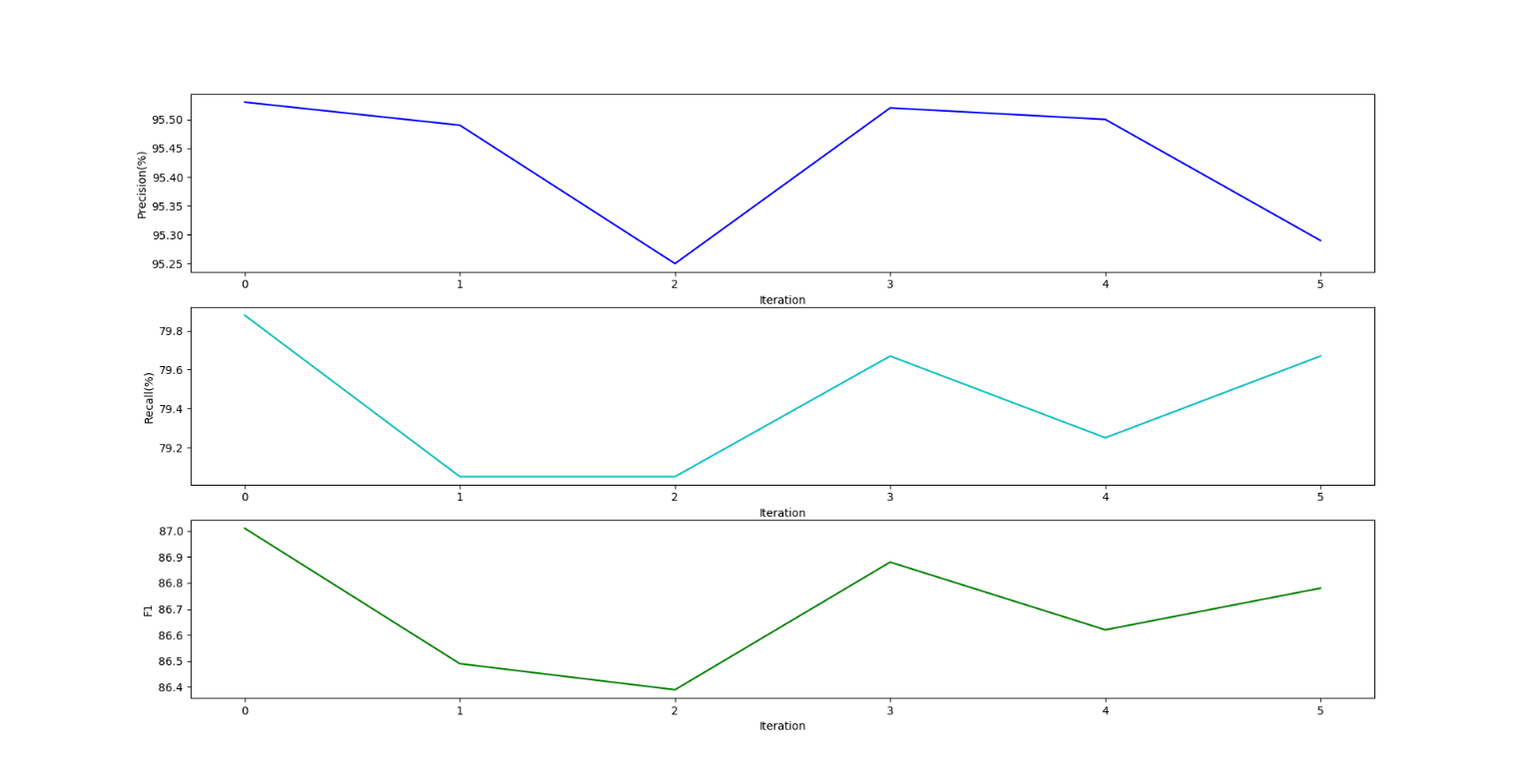
TopN方法：高准确率，低召回率

**【学习曲线】**

**置信度：0.9，迭代次数：10**



**置信度：0.95，迭代次数：5**



**置信度：0.98，迭代次数：12**

