# 命名实体识别

### 实验目的

* 了解神经网络中的基础模型

神经网络是指一系列受生物学和神经科学启发的数学模型。主要是通过对人脑的神经元网络进行抽象，构建人工神经元，并按照一定拓扑结构来建立神经元之间的连接。前馈神经网络（FNN）是最早发明的简单人工神经网络；卷积神经网络（CNN）是一种具有局部连接、权重共享等特性的深层前馈神经网络；循环神经网络（RNN）是一类具有短期记忆能力的神经网络。

* 了解深度学习框架Pytorch的使用

PyTorch使用python作为开发语言，近年来和TensorFlow, keras, caffe等热门框架一起，成为深度学习开发的主流平台之一。PyTorch的基本元素包含张量(Tensor)、变量(Variable)、神经网络模块(nn.Module)等。

* 了解使用深度学习解决序列标注任务基本流程

序列标注是NLP领域的基础问题之一，涵盖范围非常广泛，如分词、词性标注、命名实体识别、关系抽取等等，本质上是对线性序列中每个元素根据上下文内容进行分类的问题。命名实体识别是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等，以及时间、数量、货币、比例数值等文字。

以PyTorch为例，一个常规的序列标注任务代码开发流程是：安装并导入相关的深度学习库、定义标签集合(Label set)、数据获取和预处理、定义神经网络、定义损失函数(loss function)和优化器(optimizer)、训练网络和测试网络。

### 实验环境

python3 + jieba + pytorch1.0.0 + numpy

pip install -r requirement.txt

* python3

除了高性能外，拥有NumPy、SciPy等优秀的数值计算、统计分析库。TensorFlow、Caffe等著名的深度学习框架都提供了Python接口。

* jieba

jieba是一款优秀的Python第三方中文分词库，支持三种分词模式：精确模式、全模式和搜索引擎模式。

* PyTorch

PyTorch是一个针对深度学习，并且使用GPU和CPU来优化的tensor library，它是一个以Python优先的深度学习框架，不仅能够实现强大的GPU加速，同时还支持动态神经网络。

* NumPy

NumPy是Python语言的一个扩展程序库，支持大量的维度数组与矩阵运算，此外也针对数组运算提供大量的数学函数库。

### 实验数据

数据集用的是论文ACL 2018Chinese NER using Lattice LSTM (<https://github.com/jiesutd/LatticeLSTM>) 中收集的简历数据，数据的格式如下，它的每一行由一个字及其对应的标注组成，标注集采用BME，B表示实体开头，M表示实体中间，E表示实体结尾，句子之间用一个空行隔开。该数据集就位于目录下的ResumeNER文件夹里。如下为数据及标签示例：

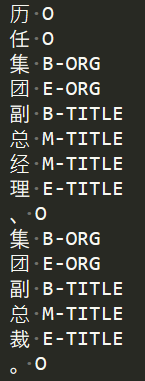


图 1 数据集示例

### 快速开始

* 配置模型相关超参数及预训练词向量路径：

/models/config.py



图 2 配置文件

* 模型训练及测试：



其中：

[模型]：cnn 或lstm

--crf ：增加crf层

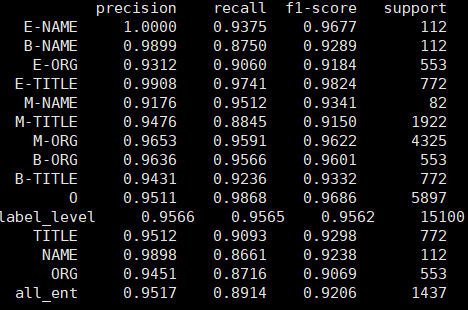
--use\_w2v ：使用预训练词向量

### 实验步骤

* 分别使用BiLSTM和BiLSTM + CRF进行NER任务，对比实验结果。

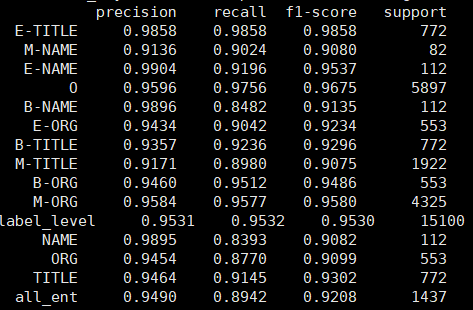
1. 阅读LSTM模型代码，运行并得到结果





1. 向模型添加CRF层，运行并得到结果

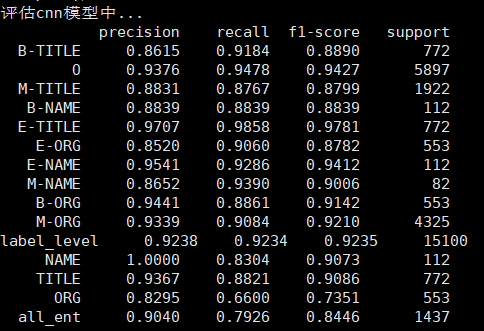


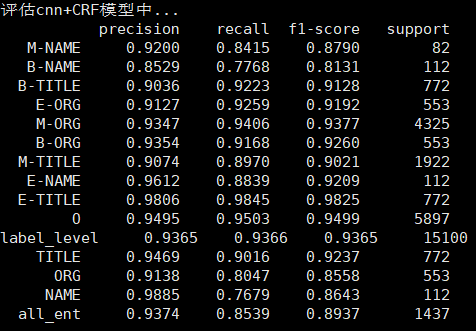


* 分别使用BiLSTM+CRF和CNN + CRF进行命名实体识别（Named Entity Recognition, NER）任务，对比实验结果。

1. 将原先的BiLSTM模型替换为CNN模型







2. 调整模型的卷积核尺寸和filter/channel参数，运行并得到结果

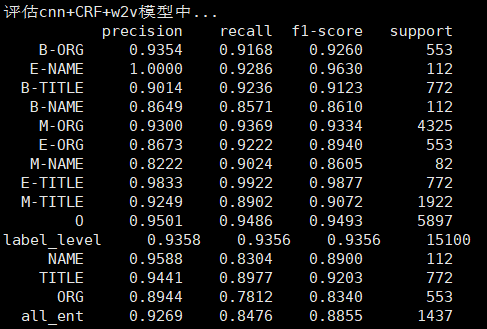
* 分别在使用预训练词向量和不使用预训练词向量的情况下进行NER任务，对比实验结果。

1. 下载预训练词向量文件，或者自己使用语料训练词向量

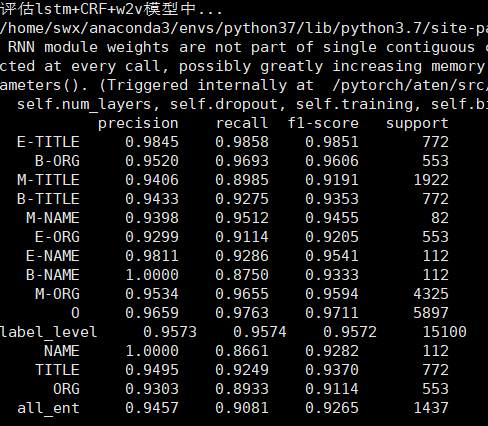
2. 编写对应的词向量初始化函数，载入已有词向量文件/随机初始化

3. 修改词向量使用参数，分别运行并得到结果









### 提交时间

**11月24号截止**

由吴雨欣同学负责收集，姓名-学号-第3次实验

### 实验要求

* 完成所有实验内容
* 良好的代码风格
* 完整的实验报告

### 参考资料

1. [Bidirectional LSTM-CRF models for sequence tagging](https://arxiv.org/pdf/1508.01991.pdf)
2. [**A survey on deep learning for named entity recognition**](https://ieeexplore.ieee.org/abstract/document/9039685/)