# Web信息处理与应用第二次实验报告

PB17151767 焦培淇

PB17151799 代弢

## 一、实验目的

本实验要求以给定的英文文本数据集为基础，实现一个信息抽取系统。

## 二、任务描述

本实验利用经过特定处理的英文数据集作为训练数据，数据集中包括训练所用的文本、文本中所包含的 实体及实体间的关系。例如

```
“The system as described above has its greatest application in an arrayed
configuration of antenna elements.”
Component-Whole(elements, configuration)
```

其中第一句是文本信息，这些句子都包含一对以上的实体，无需考虑句子中只包含单实体的情况。第二 句Component-Whole(elements, configuration)是得到的结果，提取到的实体对是(elements, configuration)，实体之间的关系是Component-Whole。实验的任务是训练关系抽取模型，即给定一个 句子输入，模型可以输出该句子中最有可能所包含的关系。

在本次实验中，你需要：

- 对于文档进行适当的预处理，例如，去除标点符号与停用词等。
- 选取合适的模型对文本进行建模，并在训练集上进行关系抽取模型训练。例如，将关系类别作为文本的标签，将问题形式化为文本分类任务，并使用相应的模型进行处理。
- 在在线平台提交结果验证关系抽取模型的准确率。

## 三、算法描述

本次实验参考了OpenNRE这个工具包(https://github.com/thunlp/OpenNRE.git)，它是由是清华大学自然语言处理实验室推出的一款开源的神经网络关系抽取工具包，包括了多款常用的关系抽取模型。我们这里采用了有监督的卷积神经网络的方法来进行关系抽取。

卷积神经网络是一种多层的监督学习神经网络，隐含层的卷积层和池采样层是实现卷积神经网络特征提取功能的核心模块。该网络模型通过采用梯度下降法最小化损失函数对网络中的权重参数逐层反向调节，通过频繁的迭代训练提高网络的精度。卷积神经网络的低隐层是由卷积层和最大池采样层交替组成，高层是全连接层对应传统多层感知器的隐含层和逻辑回归分类器。第一个全连接层的输入是由卷积层和子采样层进行特征提取得到的特征图像。最后一层输出层是一个分类器，可以采用逻辑回归，Softmax回归甚至是支持向量机对输入图像进行分类。

代码具体实现时，首先执行prepare.py文件，将test.txt和train.txt中的内容进行格式上的修改得到mytest.txt、mytrain.txt和myverify.txt，这样才能将它们作为输入来进行神经网络的训练，其中mytest.txt是测试集，

mytrain.txt是训练集，myverify.txt是验证集，mytest.txt是直接根据test.txt得到的，mytrain.txt和myverify.txt是根据train.txt得到的，并且mytrain.txt和myverify.txt的数据量的比值为3:1，即让训练集与验证集的比例大约为3:1。之后执行train_supervised_cnn.py文件，在该文件中需要指定学习速率、最大迭代次数等神经网络需要的参数，之后将mytrain.txt和myverify.txt作为输入来训练卷积神经网络，训练完毕后将mytest.txt输入到网络中，得到关系抽取后的结果，该结果是由一系列由0-9的数字组成的，每个数字对应了一个关系，映射关系写在了myjson.json文件中，最后需要执行transformResult.py，将result.txt转化为relationshipResult.txt，该文件即为最终的结果。

# 四、运行结果