实习成果汇报

张益宁

第一次汇报包含了前两周的汇报，因为最开始确定选题是周四，所以5.11~5.14算作第零周，就不再单独进行汇报了，整合在第一周的汇报中。

考虑到尽可能不占用您日程安排，我这边就直接简短的写了，也不做什么规范格式之类的。

实习方向确定

第零周主要还是确定选题，然后查阅资料，了解命名实体识别方向的一些完成方法和策略，然后制定接下来主要的完成计划。因为实习时间比较有限，所以我在看了很多命名实体的工作后对早期传统的基于词典和规则的方式产生了一些兴趣，就想着对这个方向进行深挖。

最早期的词典规则其实都是通过人工来编制的，本次实习我主要的方向就是基于深度学习来训练这样的规则，完成命名实体识别任务。这次工作也是吸取了您之前对我的指导，如果确定科研方向的话肯定是先思考能解决什么样的问题。我思考之后想着这次实习的时间比较短，目标也是一个大作业性质的，所以我的思路其实是主要没太考虑深远的创新点，想通过这次实习也是进一步锻炼一下自己的能力（相当于是解决自己编写代码不熟悉问题哈哈哈）。

因此本次实习前期简单的构建方面，我不打算参考GitHub或一些论文提供的开源代码，想着完全靠自己写一个雏形（后续写LSTM可能要参考一下，LSTM我现阶段还不会手写），主要还是为了加强一下对代码编写的能力。

日程规划

|  |  |  |
| --- | --- | --- |
| **周** | **任务** | **完成情况** |
| 5.11~5.14 | 确定选题 | 完成 |
| 了解命名实体识别基础知识 | 完成 |
| 下载并初步分析数据集 | 完成 |
| 制定日程规划 | 完成 |
| 5.15~5.21 | 了解基于规则和词典的传统方法 | 完成 |
| 了解深度学习相关方法 | 完成 |
| 手动编写代码框架 | 完成 |

主要思路：

1、先把数据集扫一遍统计出高频词（497个），构建高频词字典

2、然后对每一个词构建一个500维的onehot向量（500\*1）

0~496分别对应高频词，497表示不在字典里的其他词，498表示句首，499表示句尾。

3、然后对每一个词都链接前面一个词向量和后面一个词向量组成一个1500\*1的向量。

4、后面就针对这些词向量进行训练。（待扩展思路）

5、生成预测

总体的思路其实是类似于一个encoder+decoder的方式

其中encoder就是构建词向量的规则

Decoder目前还没开始（我先放了个MLP尝试效果），想着各种都试试看看哪个比较好

辅助介绍词向量构建：

假设我的onehot是5维的，以地名为例：

我 去 天 安 门

O O B I I

我的词典是

0 1 2 3 4

我 天 其他 句首 句尾

那么这五个字分别单独的onehot向量为：

我（1,0,0,0,0）

去（0,0,1,0,0）

天（0,1,0,0,0）

安（0,0,1,0,0）

门（0,0,1,0,0）

扩展后的词向量为

我（0,0,0,1,0，1,0,0,0,0，0,0,1,0,0）

去（1,0,0,0,0，0,0,1,0,0，0,1,0,0,0）

天（0,0,1,0,0，0,1,0,0,0，0,0,1,0,0）

安（0,1,0,0,0，0,0,1,0,0，0,0,1,0,0）

门（0,0,1,0,0，0,0,1,0,0，0,0,0,0,1）

目前思路就是这个，后续实习过程的对比实验阶段可能还会对这个词向量的构建方式进行一个更改，比如扩展词向量的窗口大小，前后看两个那种，或者更改一下其他encode模式

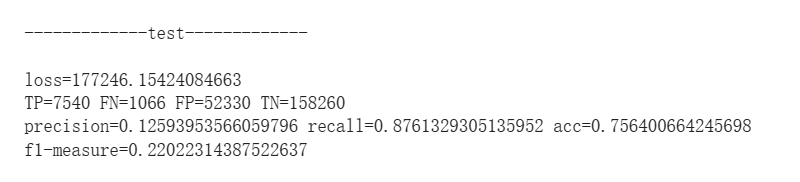
目前成果及遇到的困难：

完成了第一版本代码的编写，decoder直接就是个MLP，然后做了一个简单的三分类，先在人民日报数据集上试验了一下LOC的BIO分类。

效果非常的差，正常训练直接pre 除0异常，一看发现直接把所有词全都预测为O了

分析数据集看到在单词个数上LOC的B I占比约为1%，数据集严重不均衡。

各种调参，甚至修改loss稍微给BI两个正例求导的时候更多的权值，目前最好的结果是在测试集上f1——0.2



下周工作

接下来一周的任务就是调参+训练

首先是开展数据集预处理的工作，第一周课外时间基本都花在词向量构建的想法上了，之后可能考虑稍微清洗一下，毕竟三分类的情况下1:100确实太难预测了。

然后就是把MLP换成更好的更适合处理NLP系列问题的模型，如传统RNN、LSTM、transformer。目前我还不清楚哪个有用，下周都会试试

目前遇到的问题都可以自己解决，就想着少麻烦您一点，每周出具这样一个报告给您审阅。