本周报告

> 2020/04/24 1652792 罗吉皓

本周主要做了机器学习模型的调整以及相关项目的搭建，分为以下几个部分：

1. 数据集重新标注
2. 中国城市编码
3. 模型的重新训练以及对比实验准备
4. 标注及展示项目的前后端搭建

首先针对上次报告中提出的地理位置范畴问题，在本项目中城市为最小的地理位置单位。每个新闻事件的地理位置最后统一用城市来表示。

数据集重新标注

在上次报告的时候我发现整体机器学习的模型的准确率存在一定的问题，进行一定的分析之后发现数据集的标注存在一定的问题对于结果的影响也比较大，因此这次前一个礼拜的时间主要是对数据集进行了一个检查，完成了2000条数据，接近10w的数据量的纠正。

这项工作并没有让学弟参与的主要原因是因为每个人对于核心地点的定义还是有一些偏差，尤其是在同一个句子中出现多个相同地名的时候，究竟选取哪一个还是存在一定的误差的。这项工作由同一个人来完成也能保证整体数据的一致性。

另外在处理数据的过程中，我也对数据集的多样性进行拓展，现在数据集中主要包含以下几种数据：

1. 新闻中存在一个地理位置，且为核心地理位置
2. 新闻中存在多个地理位置，但不存在核心地点
3. 核心地理位置出现在主语的情况
4. 核心地理位置为名次组成的一部分，如澳门特区政府等

在数据集重新调整后，整体多样性和准确性都得到了一定的提升，比较适合作为机器学习的语料。

中国城市编码

中国城市编码这一块我们主要借鉴了CAMEO的标注规范以及中国邮政编码的形式。借鉴邮政编码的主要原因一个是能保证所有的编码的唯一性。另外一方面，邮政编码的精度可以达到区/县级别，精度相对比较高，能满足最后以城市为最小单位的项目要求。

数据集的重新训练以及对比实验的准备

	precision	recall	f1-score	support
P	0.7576	0.7062	0.7310	177
T	0.6319	0.6849	0.5988	73
V	0.5686	0.6042	0.5859	96
micro avg	0.6454	0.6734	0.6591	346
macro avg	0.6575	0.6734	0.6628	346

由于时间仓促，在轮数比较少的情况下先训练了一个模型查看效果，可以发现其中准确率较上次报告的基础上已经提高了10个百分点。

```
precision: 0.631300, recall: 0.687861, f1: 0.658368
```

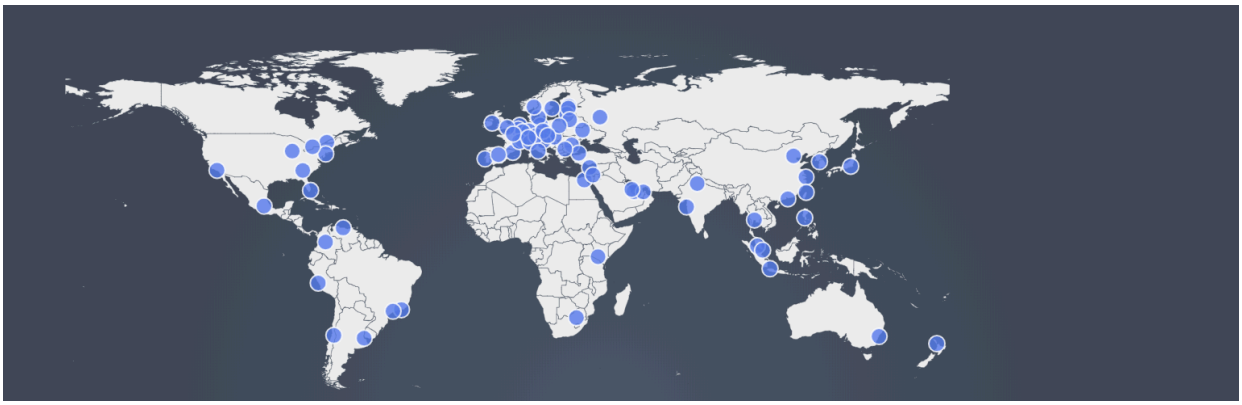
在对比实验中主要针对CNN_LSTM模型，BiLSTM模型以及BiLSTM_CRF三个模型进行对比实验，模型已经搭建完毕。由于本机训练的时间比较长，在这次的报告中还无法得到最后的对比结果。

标注及展示项目的前后端搭建

在本周标数据以及等待项目跑模型的空闲时间中，我主要着重于构思及搭建整个项目的展示平台。目前我的想法是搭建一个集标注与展示于一体的一个项目来进行相应的展示。其主要功能如下：

1. 前端界面有对应的文本框输入新闻以及展示模型预测结果
2. 用户可以调整相关模型结果
3. 调整完/确认完的结果会在地图上进行展示
4. 在用户确定以后，相关的文本以及标注信息会被传递到后端进行存储
5. 后台定期整理所有数据，进行重复训练

目前前端展示界面如下：



目前项目进度：

1. 前端搭建，目前选取了echarts作为图表展示平台，目前基本可以做到前后端联动
2. 后端搭建，目前选用的Django作为信息处理平台。包括模型预测，数据再整理等过程。

后端目前的解决方案

1. 模型处理方案
    1. 目前采用的模型处理由动词-地名树与机器学习模型共同搭建，目前采用的方案是通过动词-地名树和机器学习同时预测相关核心地理名词的位置，如果两者完全匹配则直接返回，如果匹配有问题则将两方面的结果都返回给前端，由用户进行相应的判断
    2. 遇到的坑：keras和Django不能共存的问题
        1. Django后端启动的时候会重复多次调用keras模型以至于会报告一些奇奇怪怪的参数错误

```
ValueError: Tensor Tensor("dense_2/Softmax:0", shape=(?, 6), dtype=float32) is not an element of this graph.
```

解决方案：在Django项目启动前先进行keras模型的加载

```
This could mean that the variable was uninitialized. Not found: Container localhost does not exist.
```

解决方案：保存初始session，在模型预测时采用初始session进行解决。