# 我只是来打个酱油 队伍比赛报告

第十四名队伍

## 团队成员介绍

**高建伟**，浙江大学计算机学院研究生三年级

联系方式：1449894353@qq.com / 17706431266

**参赛原因**：以往做的NLP方面的比赛主要是在文本分类方面，刚好天池能够提供了一个好的平台和数据，想试下短文本进行语义的匹配的工作

**收获**：本比赛中，我主要是使用深度模型做Semantic textual similarity，在整个比赛过程中，为了取得比较好的试验结果，查看了很多篇近两年来宣称达到state-of-the-art的论文，包括BiMPM, ESIM, MPCNN, Siamese LSTM, StackedBiLSTMMaxout等模型。从中学习到很多如何使用深度模型更好的抽取出sequence 语义的方法。

## 结题思路和算法思路

1. 对Spanish和English数据进行预处理。预处理包括，转化成小写，替换一些低频的标点符号，连续重复的标点符号只保留一个；分词是使用Stanford的CoreNLP Tools做的
2. 训练数据去重
3. 生成Spanish和English字典
4. 对训练数据切分5 fold
5. 利用比赛提供的fasttext词向量，根据字典的顺序词向量
6. 通过随机选择参数的方法，选择不同的模型多次运行程序，最终将所有结果融合。(比赛最终提交版本是融合了32个运行结果)

### 改进模型的方法

1. 将英文语料也添加进模型中进行学习，多个语料同时学习，相当于使用Multi Task Learning Method的方法，提高了模型的泛化能力
2. 模型增加使用Character + CNN的embedding，丰富词向量
3. ~~Position Embedding（效果不显著，没有使用）~~
4. ~~Pos Tag Embedding（添加上之后，模型很难收敛，没有使用）~~
5. ~~对模型增加手动提取的特征（模型很难收敛，没有使用）~~

## 代码部分

## 运行本项目的过程：

1. 安装requirements.txt中的所有库
2. 从链接https://stanfordnlp.github.io/CoreNLP/中下载corenlp的安装包（运行需要Java），下载西班牙语的对应文件,放置到安装包主目录中
3. 修改code/build_data.py中的`corenlp_path`

## 运行环境

Ubuntu16.04

python3

Pytorch0.4

TitanXP

## 项目文件夹说明

./activations/ 激活函数

./checkponts/ 保留的权值文件

./data/ 模型加载数据的代码

./fasttext/ 词向量位置

./ml_method/ 尝试使用machine learning方法提取特征加到深度模型中，但是提取到的特征加到模型后模型很难收敛。最后没有用到

./result/ 保存结果

./script/ 放置废弃不用的代码

./utils/ 工具代码

## 运行

1. 打开code/
2. 运行 `python build_data.py` ,数据预处理
3. 运行 `generate_embedding_weights.py` , 根据vocab生成预训练好的词向量
4. 修改config.py中的参数，然后运行main-5fold.py，会进行5 fold的CV训练。
5. 通过修改main-5fold.py中的model_class来选择对应的model。
6. 训练多次，运行blending_result.py，将所有结果进行blending, 保存到submits中
7. 提交结果