Quora 问题等价性案例学习

极客时间

扫码试看/订阅

《NLP 实战高手课》视频课程

Quora 问题等价性案例学习

- Quora 问题等价性比赛简介
- 第四名 Solution 解释
- 如何在 BERT 基础上提升等价性模型效果

Quora 问题等价性比赛简介

- Kaggle最著名的比赛之一
- 训练集 40 万条，测试集 200 万条（有机器生成的）
- 标注质量较高
- 区分难度较大

第四名队伍

Niu Guocheng
Master
@Baidu

Researches: Natural Language Processing and Machine Learning

Pang Liang
PhD. candidate
@ICT

Researches: Information Retrieval, Matching Learning, Deep learning

Hou Jianpeng
Master
@Google

Researches: Machine Learning and Distributed Computing

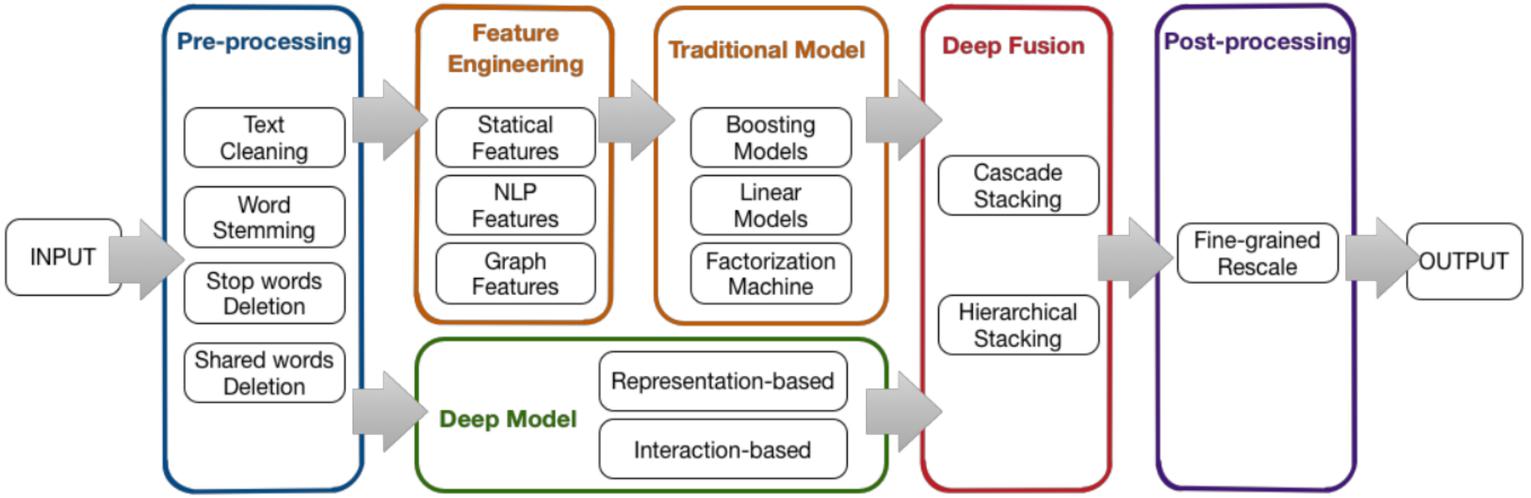
Fan Yixing
PhD. candidate
@ICT

Researches: Information Retrieval, Matching Learning, Deep learning

Yue Xinyu
Master
@ICT

Researches:Recommendation, Data Mining, Machine Learning

整体方案架构


INPUT
Pre-processing
Text Cleaning
Word Stemming
Stop words Deletion
Shared words Deletion
Feature Engineering
Statical Features
NLP Features
Graph Features
Traditional Model
Boosting Models
Linear Models
Factorization Machine
Deep Model
Representation-based
Interaction-based
Deep Fusion
Cascade Stacking
Hierarchical Stacking
Post-processing
Fine-grained Rescale
OUTPUT

预处理

- 目的：构建不同版本的数据集供模型平均使用
- 对于人工特征可采取该种方案
- 对于深度学习模型，可能需要考虑到算力均衡
- 一些注意事项：清理的未必是错误的数据
  - 例子：$->dollar
  - 例子：Machine Learning -> machine learning

特征来源

- 基础特征（无需和问题相关）
- 数据特征
  - 来源于统计特征
  - 来源于直接观察
- 和问题相关的特征->区别
- 文本之外的特征

极客时间

思考：如何衡量区别

思考：文本之外的特征有哪些

其他特征

- Topic-model
  - TF-IDF 矩阵进行降维
- 关键词提取
  - 讲关键词后置于原文档中

Quora 问题等价性案例学习：深度学习模型

Quora 问题等价性案例学习

- 深度学习模型分类
- 引入 BERT 的过程
- 如何提升

深度学习模型分类

- Representation Based：BIMPM（https://arxiv.org/pdf/1702.03814.pdf）
- Interaction Based：MatchPyramid（https://arxiv.org/pdf/1602.06359.pdf）

思考：如何在其中引入 BERT

进一步提升方法

- 网络架构的改变
- 采用更多层作为输入
- 不同语言模型 concat
- Adversarial Training
- ...

文本校对案例学习

文本校对案例学习

- 英文和中文校对的区别
- 翻译校对、语音转文字校对和文本校对的区别
- 如何解决样本分布不合理的问题
- 如何解决语言模型预测能力不足的问题
- 提升

问题一：英文校对和中文校对的区别

极客时间

问题二：翻译校对、文字校对和文本校对的区别

问题三：如何定义“正确”的错误

问题四：如何解决语言模型预测不准确的问题

提升

- 语言模型的 post-train
- 问题的分类和分步提升
- 针对高频错误的纠正
- 其他可能性（语法）

极客时间

扫码试看/订阅

《NLP 实战高手课》视频课程