预处理

- 数据集问答对不规范[TODO]
  - adj + n VS. n(形容词的有无, 找出实例)
- 问题分类[TODO]
  - 关键字匹配分类/翻译后关键字分类
- 数据增强[Done]
  - 回译: 中文-有道->英文-搜狗>中文, 降低1个百分点; 分析翻译句, 需要做进一步筛选[Ref]
  - EDA: 不规范, 效果差[Ref]
- start/end确定
  - 解决方案1, 制作字表, 规定答案长度为n, Ref
  - 解决方案2, 提取训练集中所有的词作为词表, 答案为词的下标Ref
  - 解决方案3, 为数据集答案生成下标, 多个匹配的答案, 使用最长匹配原则
- embedding
  - context和query分开使用Bert编码
  - [CLS]query[SEQ]context[SEQ]形式, 之后使用query/context mask信息进行attention(masked-softmax[Ref Standford final project starter code]), 主要是考虑到的Bert预测下一句的训练方式, 这样可以包含更多信息

模型尝试

- Bert baseline
  - [见结果比较]
  - 12层transformer + fully-connnected
- Bert + PointerNetwork
  - 主要解决答案边界选择问题
  - [见结果比较]
- Bert + R-net
  - attention+match+pointer-network
  - [见结果比较]

训练及相关问题(Bert + R-net)

- loss = nan, 输入样本进行了长度限制切分, 部分样本会没有答案, 导致之后loss计算出现$-y * log(0)$
  - 直接删除没有答案的样本[Done]
  - 修改loss函数[TODO]
  - 添加额外片段[?]
- 过拟合, loss = 1e-5
  - 各层加入dropout(keep_prob=0.5/0.6/0.75)

结果分析

[scores.xlsx]

- Bert baseline
- 实验结果(train:test=5:1)
  - Bert baseline
  - ■ 回译[略降]
  - ■ EDA[pass]
  - Bert + pointer-network[略升]
  - Bert + rnet[略降]
- 实验分析
  - 回译

        - 需要对翻译进行进一步的选择
    - Bert + rnet, 后续添加没有答案
    - 查看回答错误的问题情况

可视化应用

- 可视化[Done]
    - 文章文件选择
        - 准备文章[TODO]
    - 语音询问(科大讯飞接口，语音识别和语音合成)
    - 回答在原文中不同颜色标注(nbest)
        - 解决方案
            - 使用Red Here Here Here
            - 重叠问题使用区间重叠方法[]
        - ~~combox按钮，可以选择查看不同的问题进行答案查看~~
    - 考虑网页形式或PyQT内嵌网页

PPT

[Ref]

Report

Latex