

每日问答的需求推进计划书

背景介绍：

了解目前的游戏中的答题模块的现状，分析是否需要使用NLP或者其他方式进行补充，用来生成更多多样性的题目，并满足游戏的使用条件。

首先与直接负责人沟通，探讨是否有做的必要性，手游、端游分别推进。

需求了解

神武手游：

- 综合设计-文案-柯学
- 需求描述：

探讨一下，是否有使用人工智能方法(NLP算法)或者其他算法来生成每日问答题库的需求。

现状：

目前题库的来源：人工收集或编辑

目前手游的问题是题库较为老旧，玩家根据现有的内容制作了题库，可以自动答题，搜集答案。

但是更新题目的代价比较大，且时效性只有一次，保持长久的大体量更新题库是不太现实的

这里是否能让题目和答案自动生成一些变种，让玩家不能直接用题库搜索

分析：

手游题库现状：

- 1、目前题目老旧
- 2、保持大体量题库比较麻烦

需求：

生成新题库，目的：玩家不能直接搜索得到答案

综合设计-文案-柯学 观点：

难点：很多题目是常识通识文学、诗词类的知识，所以变换起来比较麻烦

题库数据：

手游与端游题目有部分相同

神武端游

- 神武4-游戏策划-司羽
- 需求分析--(与手游类似，不重复)

端游的每日答题题目类型为：

传统文化
 自然科学
 生活百科
 游戏知识

- 数据--题库

已拿到端游题库
- 关注点：需要较明确的实现时间表，担心无疾而终

端游题库分析

编号表示题目类型	
100,000~199,999	传统文化
200,000~299,999	自然科学
300,000~399,999	生活百科
400,000~499,999	游戏知识
500,000~599,999	图片题

科举答题活动抽题规则(只记录、暂不考虑细则):

1. 不抽取5开头的图片题
2. 明经固定5道游戏知识题、5道其他题
3. 进士固定8道游戏知识题、12道其他

题例:

1. 传统文化(438题)

题目	选项
“民为贵，社稷次之...”出自哪部作品？	《孟子》、《孔子》、《庄子》
佛光寺是现存的唐代哪种结构的建筑？	木结构、砖结构、石结构
三国演义中的“卧龙”是？	诸葛亮、赵云、华佗
明朝戏剧《牡丹亭》又称？	还魂记、西厢记、长生殿
古人用“廿”表示多少？	二十、四十、三十

2. 自然科学(239)

题目	选项
不移动圆饼切三刀，最多可以切出多少块？	8、10、14
蜂巢的内部结构是什么形状？	六边形、方形、圆形
雷雨天，人们先看到闪点后听到雷声是什么原因？	光速大于声速、先产生闪电后产生雷声、耳朵反应比眼睛慢
人脑中控制平衡性的部位是？	小脑、大脑、脑干
光年是天文学上的什么单位？	长度单位、光速单位、时间单位

3. 生活百科(416)

题目	选项
飞行最快的鸟是什么？	雨燕、苍鹰、海鸥
俗称的长生果是什么？	花生、腰果、核桃
“天无三日晴、地无三尺平”值我国哪个省份？	贵州、广西、宁夏
“狼毫”原料取自哪里？	黄鼠狼尾、野狼尾、马尾
诺贝尔是哪国人？	瑞典、瑞士、奥地利

4. 游戏知识(225)--（知识图谱的推理需求

题目	选项
下面哪项属性对攻击力的影响最大？	力量、魔力、耐力
“雪千寻”有可能加入哪个门派？	魔王山、天策、凌霄天宫
以下哪个NPC在青河镇？	小石头、周道人、易容师
每件装备最多可同时镶嵌几种宝石？	2、3、4
唐刀不适用于哪一个角色？	慕子白、燕归行、羽无殇

5. 图片题(略--48题)

目标

问题生成按照有无答案可分为：有答案问题生成、无答案问题生成；本项目着眼于有答案问题生成问题的解决。问题生成的主要挑战是识别答案相关的上下文单词、文辞陈述句--疑问句的转换。

目前的Seq2Seq模型认为答案近邻的单词更可能是答案相关的，通过显示编码每个词与答案的相对距离来实现。

执行计划

目前每日问答的问题生成推进计划：

1. 查找目前可用的开源中文数据语料，针对不同的问答范围对语料进行分类
 - 收集可获数据集，若不满足需求则需要后续自己制作语料数据集（不同的问题范围不同，需区分开）
2. 查找目前问题生成领域可用的较为成熟的解决方案，评估不同方法的难度并快速验证是否可用
 - 根据已有实现方法进行实验测试，看是否可满足策划需求
3. 根据评估的方法，推进问答项目并完成题目生成的需求，与策划进行沟通，针对问题质量进行调整
 - 针对策划反馈的具体细节进行调整

目前的问答题库--推进方案

1、确定评估模型性能的测试数据集，具体方案为：

- 首先根据语料的类别进行分类处理
 - 生活常识、自然科学、传统文化
 - 注意问题的分布要均衡
- 研究其他的可量化评估工具
 1. 指标-1：精准匹配率（Exact Match, EM）
 - 计算预测结果与标准答案是否完全一致
 - 一致=1分，不一致=0分
 2. 指标-2：模糊匹配率（F1-score）
 - 计算预测结果与标准答案之间的字级别（character-level）匹配程度

目前该方法如何使用，还需要进一步研究。

2、目前模型算法的推进方案

- 简介目前使用的算法

目前使用的算法基于bert-base的预训练模型，在CNMed数据集(中药数据集)上进行训练，然后再在CMRC2018数据集上进行fine-tune，目前这样的模型的性能最佳。之前使用A/B测试的方法对基础模型进行训练，实验的设计为：基于bert-base + CNMed + CMRC2018 训练出来的模型比bert-base + CMRC2018 的模型具有更好的泛化性能。可以得出的结论为：更多的高质量数据集对于提升模型的性能具有决定性作用。

- 后续可以采集的研究方向

1. 更换底层的base-model，进行更多预训练模型的尝试：

- 高规模调参的尝试
- 中文预训练模型的尝试

2. 收集或者制作更多的目前项目所需的训练数据集

- 高质量数据集的作用效果显著
- 针对后续的问题生成的定向数据集制作，比如先尝试制作一个小型数据集(自然科学方向)

然后确定多大规模的数据集，可以显著继续提升模型性能，即：提高问题生成的质量

3. 学习使用HuggingFace的Transformer库

- 使用强大的可用工具进行探索

目前的实验记录及得出的一些结论

- 时间：2021-12-31
- 内网机器：10.17.67.221:8848

QA-NLP/wobert_L-12

1. base-config

- bert_config.json
- bert_model.ckpt.data-00000-of-00001
- bert_model.ckpt.index
- bert_model.ckpt.meta
- checkpoint
- vocab.txt

2. 不同训练策略下的模型

1. 最基础的预训练模型(追一科技的bert-base中文预训练模型)

- base-model (1)

2. 基于中药CNMed数据集训练的模型

- base-model (1) + CNMed --> 模型(2)，命名：base-CNMed

3. CMRC2018数据集fine-tune的模型

1. base-model (1) + CMRC2018 ---> 模型(3), 命名: base-CMRC2018
2. base-model (2) + CMRC2018 ---> 模型(4), 命名: base-CNMed-CMRC2018
 - 目前 模型(4)的性能最佳。