项目4 – 面向服务的对话机器人的构建

慧科集团开课吧人工智能学院

2019-10-28

目录

01. 问题背景描述

对话机器人的常见的NLP与自然语言处理的应用方向，本课题将使用脱敏的某真实常见下的QA数据集，构造一个用于银行系统的对话机器人。

如图所示，国内现在很多地方已经有了这种“人形”的对话服务机器人，这种机器人可以处理常见的很多业务，例如开卡，例如挂失补办等等业务。

- 银行对话机器人用到的这些技术，还可以使用在其他领域，例如保险，理财，购物等等。

- 例如阿里小蜜，就承接了诸多阿里巴巴本来需要人工客服干预的功能。

(a) Assistance Service (b) Customer Service (c) Chatting Service

Figure 6: A demonstration of *AliMe Assist*

2. 所需要的数据与环境

- 对这个问题来说，在算法层面，输入输出很简单：

- 输入： String
- 输出：String

- 例如：如何办理开卡业务？
- 输出：办理开卡需要你的XXX

- 所需要的数据集：
  - 本课题需要的数据集为脱敏的3万多条Question-Answer对，数据存放在服务器: student@39.100.3.165/dataset/ 中，请大家下载或者直接在远程服务器进行使用
- 本次项目用到的环境为：
  - Python3.6， Pycharm, Jupyter Notebook
  - Keras， Tensorflow, Pandas
- Linux Ubuntu 服务器
- Ubuntu服务器的用户名和密码为：
  - student@39.100.3.165
  - AI@2019@ai

3. 相关的实现路线

面向任务的对话机器人

- 业务相关问题
  - 意图识别
  - 语义相似度判断
  - 布尔搜索
- 基于爬虫的开防御问题
  - 爬虫技术
  - 语义相似度判断
- 相关问题的自动生成
  - 基于语法树与词向量的句子生成

Part-1

意图识别

- 用户输入的问答，首先经过该判断，判断是否为业务相关的问题，如果不是业务相关的问题，则调用爬虫，在互联网上查询答案，若找不到答案，则生成建议的问题
- 核心技术点：文本的无监督分类，首先使用无监督方法，将文本聚类，然后使用输入的句子和其中的每个kernel进行对话，观察其语义距离是否大于某个阈值

语义相似度判断

- 如果判断输入的句子属于某个无监督聚类之后的类别，则需要将这个类别中的每个句子的向量与输入的句子的向量进行对比，快速找到最相似top10个句子
- 注意：1. 本步的向量化和上一步的向量化未必一样；2. 如何快速减速向量化，这是一个重点，需要重点突破

布尔搜索

- 对于输入的问题，可以使用布尔搜索的方法，快速从字词层面获得相似top10个句子。获得的这些句子结合上一步获得的句子，作为一个20个参考候选集。在这个候选集合中，设计更加精确的排序方法，完成排序
- 完成排序之后，将问题对于的答案给出即可

Part-2

爬虫技术

- 如果之前我们找不到在QA库总比较相似的问题，那么我们就要借助互联网，在百度知道，搜狗问问中，使用爬虫的方式，获得相关问题的搜索。
- 例如问道“AI和区块链什么区别”，这种问题，QA库中最相思的问题也是低于某个阈值的，那么我们就可以使用第二课中介绍的爬虫技术，在网上进行爬取，获得答案。

基于语法树与词向量的句子生成

- 如何我们在QA库中和互联网爬虫中都找不到该客户问的问题，那么我们就要根据他提出的问题，生成我们的建议问题
- 该方法可以结合我们第15课讲过的方法，包括基于Syntax Tree和Langauge Model, Word Embedding结合的语句生成方法

4. 数据可视化建议

- 该可视化，建议做成类似于Web微信的样式，一问一答。
- 也可以使用wxpy等，集成到微信里边。
  https://github.com/youfou/wxpy

wxpy: 用 Python 玩微信

pypi package 0.3.9.8 | python 2.7 | 3.4 | 3.5 | 3.6 | docs passing

微信机器人 / 可能是最优雅的微信个人号 API

wxpy 在 itchat 的基础上，通过大量接口优化提升了模块的易用性，并进行丰富的功能扩展

Attention!

强烈建议仅使用小号运行机器人!
从近期 (17年6月下旬) 反馈来看，使用机器人存在一定概率被限制登录的可能性。
主要表现为无法登陆 Web 微信 (但不影响手机等其他平台)。

用来干啥

一些常见的场景

- 控制路由器、智能家居等具有开放接口的玩意儿
- 运行脚本时自动把日志发送到你的微信
- 加群主为好友，自动拉进群中
- 跨号或跨群转发消息

5. 如何评测？

Precision at Top10

- 1. 我们可以把数据分成training和validation两个集和
- 本课题中，我们系统给出的10个候选句子，我们将10个句子中包含该正确的对应回答的句子，定义为命中，则，precision = #命中的句子个数 / #总测试对话数；
- 2. 对话机器人的衡量本身就是一个业界难题，因为可能会牵扯到开防御问题。可以参考： https://arxiv.org/pdf/1801.03625.pdf
- 3. 为此，我们开发了一个公开的评测系统，根据输入的问题，批量产生答案，然后人工进行评测打分，源代码可参考：
  https://github.com/fortyMiles/EvaluationPlateform
- 值得说的是，在真正的使用环境下，人工打分本身就是一种非常重要的评测方式。

6. 总结与建议

该问题比较综合，牵涉到的知识点比较多，基本上把我们从第一课到第15课的内容，除了文本分类，别的都涉及到了。对于每个知识点，不要追求很深，而是要考虑如何能把这些点结合起来；

早做网页版本的对话系统，这样对自己的调试起来很有帮助，而且能在找工作的时候，发布出来，让面试官看到；

及时做评价，做反馈。只有不断跟踪结果，才能不断优化。

7. 如何提交项目

- 提交项目应该是一个压缩包，该压缩包包含以下内容：
  - 1. 项目源代码（不需要包含数据）
  - 2. 项目的PPT效果展示
  - 3. 你的参数调整记录表
  - 4. 该项目能够访问的网站链接
  - 5. 该项目的优缺点和模型分析报告，包含模型的 top10 precision

- **之后将该Zip压缩包以发送至** ai-college@kaikeba.com　抄送至：**ynzhang@kaikeba.com**
- 邮件主题：项目+所有小组成员姓名（如“项目三-张三，李四”）
- 邮件内容：应包含需要反馈的邮箱地址（如 张三：**11111@qq.com**；李四：**2222@qq.com**）
- 项目接受截止日期：2019年12月28日

8. 优秀学员奖励

- 我们按照 1 – 5分给每位同学每个项目打分，4个项目的占比分别是：
  - 20%， 20%， 30%， 30%
  - 团队成员以团队整体成绩为计算

- 对综合排名前6名的同学，我们提供以下奖励：
  - 第 1 名： Kindle Oasis 阅读器（或同等价值奖品）
  - 第 2， 3 名： Kindle Paperwhite 阅读器 （或同等价值奖品）
  - 第4， 5， 6 名：Lamy钢笔 + 精选图书一册
  - 此6名同学可直接获得阿里巴巴（蚂蚁金服），百度，字节跳动，微软，IBM内推机会，

好了，大家加油吧！