

聊天机器人项目报告

目录

一、项目背景-----	2
二、项目简介-----	2
三、数据收集-----	2
四、项目架构-----	3
流程图 1-----	5
流程图 2-----	7
五、学习收获-----	8

张劲帆

2020.4.16

1 项目背景

聊天机器人，是一种经由对话或文字进行交谈的计算机软件程序，能够模拟人类对话并解决不同场景下的业务需求。随着人工智能、互联网等技术的发展，聊天机器人应用多重领域，在电信、旅游、医疗航空、金融等领域均有涉及，效果明显。人工智能技术突破了聊天机器人原有的技术瓶颈，并且实践证明，聊天机器人的使用不仅能够为企业减少一大笔人力成本，而且能够明显提高工作效率，国内外多家企业纷纷布局聊天机器人行业，聊天机器人市场一片蓝海。

2 简介

本项目是基于医疗知识图谱、bert 文本相似度和 seq2seq+attention 模型实现的集成式聊天机器人，可实现问诊、开方等功能，也可根据情感分析对用户进行心理开导，查看最新疫情情况。其中，知识图谱主要为肝类疾病，规模较小，专业性较强，但其图模型的结构特点使得它能够挖掘更深层的关系。文本相似度问答的优点为储备规模大，若遇到相似问题可给出较好的回答，缺点是计算太慢。seq2seq+attention 模型的优点是可以对新的问题进行回答，泛化性好，缺点是需要大量数据进行训练，否则效果不佳。

3 数据收集

知识图谱

采用了 github 上的医疗知识图谱，该知识图谱以肝类疾病为核心，包含多种疾病，7 类规模为 4.4 万实体的知识实体，11 类约为 30 万的实体关系。

其中实体的类型有：

实体类型	中文含义	实体数量	举例
Check	诊断检查项目	3,353	支气管造影;关节镜检查
Department	医疗科目	54	整形美容科;烧伤科
Disease	疾病	8,807	血栓闭塞性脉管炎;胸降主动脉动脉瘤
Drug	药品	3,828	京万红痔疮膏;布林佐胺滴眼液
Food	食物	4,870	番茄冲菜牛肉丸汤;竹笋炖羊肉
Producer	在售药品	17,201	通药制药青霉素 V 钾片;青阳醋酸地塞米松片
Symptom	疾病症状	5,998	乳腺组织肥厚;脑实质深部出血
Total	总计	44,111	约 4.4 万实体量级

该知识图谱结构如图：

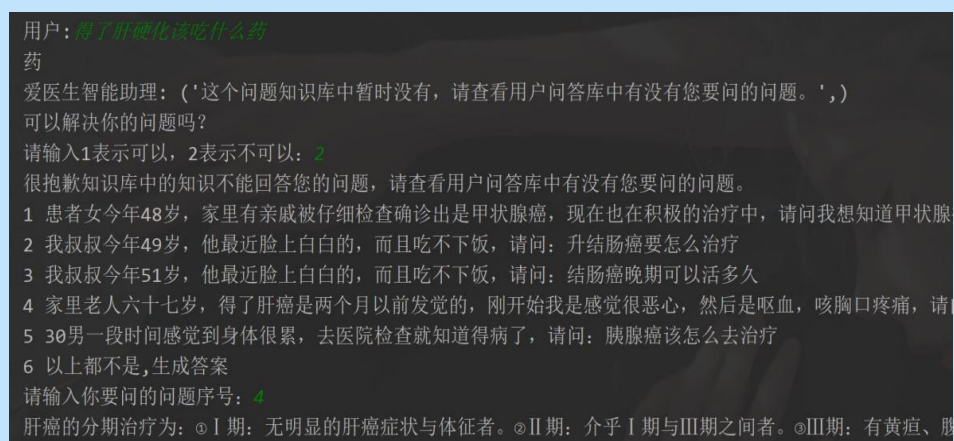


问答对

爬取了包括儿科、内科、肿瘤科等 30 万患者-医生问答对，用于文本相似度匹配和训练语言生成模型。问答对的构成为 lable, question, answer。如“‘心血管病’，‘得了个高血压’，‘要多吃清淡的食物’为一组数据。

4 项目架构

本项目分为了两块，一块是 pycharm 里运行的聊天机器人，一块是在前端界面显示的聊天机器人，原理基本一致。效果图如下：



用户：得了肝硬化该吃什么药

爱医生智能助理：('这个问题知识库中暂时没有，请查看用户问答库中有没有您要问的问题。',) 可以解决你的问题吗？

请输入1表示可以，2表示不可以：2

很抱歉知识库中的知识不能回答您的问题，请查看用户问答库中有没有您要问的问题。

1 患者女今年48岁，家里有亲戚被仔细检查确诊出是甲状腺癌，现在也在积极的治疗中，请问我想知道甲状腺

2 我叔叔今年49岁，他最近脸上白白的，而且吃不下饭，请问：升结肠癌要怎么治疗

3 我叔叔今年51岁，他最近脸上白白的，而且吃不下饭，请问：结肠癌晚期可以活多久

4 家里老人六十七岁，得了肝癌是两个月以前发觉的，刚开始我是感觉很恶心，然后是呕血，咳胸口疼痛，请

5 30男一段时间感觉到身体很累，去医院检查就知道得病了，请问：胰腺癌该怎么去治疗

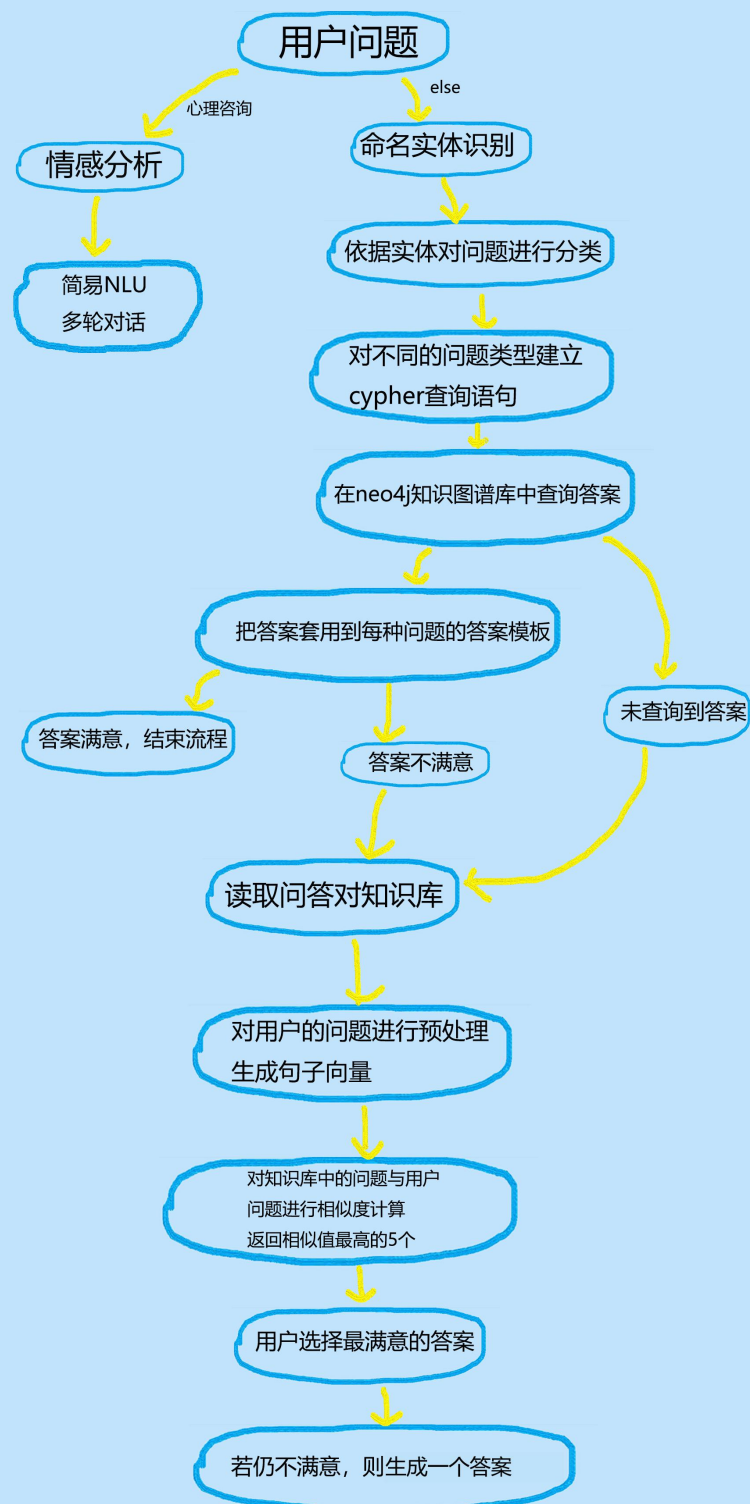
6 以上都不是,生成答案

请输入你要问的问题序号：4

肝癌的分期治疗为：① I 期：无明显的肝癌症状与体征者。② II 期：介乎 I 期与 III 期之间者。③ III 期：有黄疸、腹



下面是在 pycharm 里运行的聊天机器人的架构图：



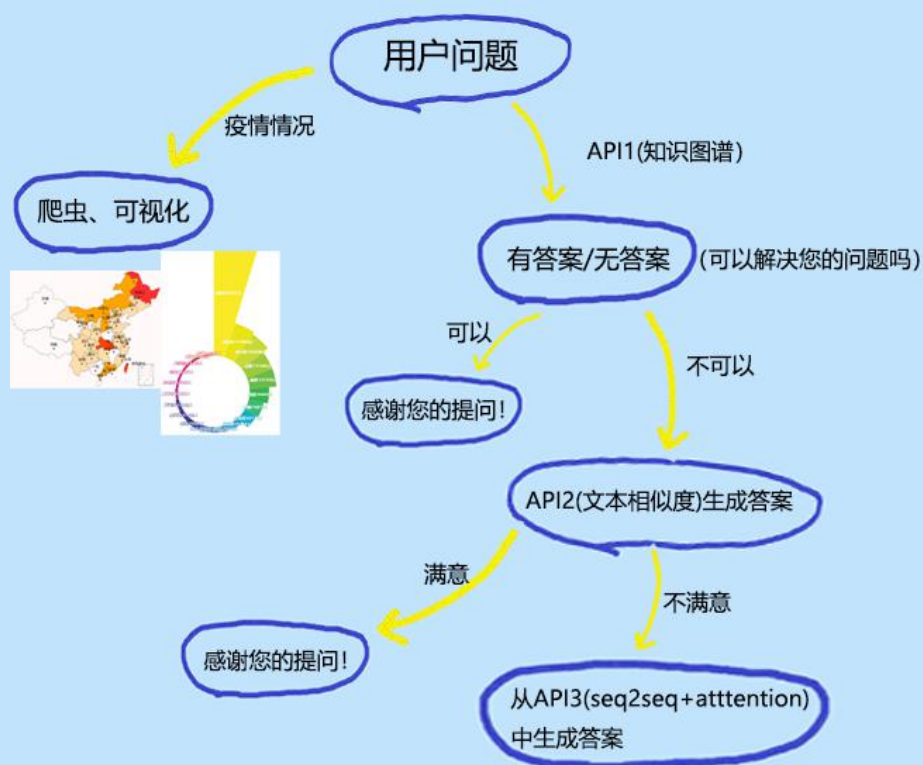
在 pycharm 里运行的聊天机器人技术架构：

1. 当用户输入‘您好，我想进行心理咨询’时，计算机会对用户输入的话进行情感分析，并给出相应回答。用户的情感分为积极、中立和消极三种状态。
2. 由于用户问题的意图较为单一、简单，当用户进行医疗咨询时，未对问题进行意图识别，而是会对问题进行命名实体识别，由于中文没有较成熟的命名实体识别库函数，这里采用的是一种多模式匹配算法：AC 自动机，以及正则表达式甄别否定实体技术来提取问句中的关键词。
3. 目前知识图谱问答的主流方法有 3 种：基于模板的方法，基于语义解析的方法和基于深度学习的方法。由于该知识图谱较小，种类较少，采用基于模板的方法效果会更好，但缺点是专业性太强，泛化性不佳。在 question_classifier.py 中，计算机会根据提取出的实体对问题进行分类。如将含有‘药品’，‘用药’的句子归为‘吃什么药？’问句类型。
4. 在 question_parser.py 而后根据不同的问句类型构造相应的图数据库查询语言 cypher。如问句类型查询疾病的原因对应的 cypher 查询语句为 "MATCH (m:Disease) where m.name = '{0}' or '{0}' in m.another_name return m.name, m.cause"
5. 在 answer_search.py 中，通过 cypher 在 neo4j 图数据库中查询答案后，将答案套用到每种问句类型的答案模板中。如“{0}通常可以通过以下方式检查出来：{1}”
6. 在 sim_seq.py 中，若用户对答案不满意，则利用 bert 文本相似度，将用户的问题与问

答库中的问题进行相似度计算，返回相似度值最高的 5 个问题，由用户选择最接近的问题，而后计算机返回对应的答案。由于数据量大，计算相似度非常慢，我采用了如下的方法来减少计算时间：(1)对用户输入的问题进行词性标注，提取出词性为专业名词(nz)的词，绝大部分情况就为疾病名称。(2)每个问答对都有一个 lable，如“得了高血压”则属于心血管病。(3)相比于数十万的问答对，lable 的总数只有数百个。通过计算问题提取的词与 lable 计算词相似度，就可缩小搜索范围，将用户的问题与标签为某个 lable 的问答对计算句相似度。

7. 若用户仍对答案不满意，则利用预训练好的 seq2seq+attention 模型，返回生成的答案。相比于传统 seq2seq，attention 机制有着参数少、速度快和效果好的优点，具体细节本文不做阐述。如此，实现了多轮对话机制。

下面是在前端界面里运行的聊天机器人的架构图：



在前端界面里运行的聊天机器人技术架构：

1. 当用户想了解中国及世界其他各国的疫情情况时，计算机会自动爬取最新疫情情况，并生成可视化图。
2. 由于需要与前端结合，基于 flask 框架把主函数分别做成了 3 个接口，分别返回基于知识图谱，文本相似度，seq2seq 三种模型的答案。
3. 其他部分原理架构与之前一致。

5 学习感受与收获

短短的一个月时间的远程科研项目，随着时间飞逝而过，但收获的知识与经验却比我想得要远大的多，超乎了我的想象。抱着不辜负老师朋友的期待，我在这次实践过程中，努力将每一件事都做到认真对待细心观察，体悟蕴含在其中无穷奥妙。

以前我总是胆怯、懒惰，觉得写代码是一件很难的事情。总是喜欢看原理，看论文，不怎么看代码。但经过这次项目实践让我明白，写代码并没有那么难，它是一个循序渐进的过程，并且通过写代码，才能让我们的理论知识得以体现运用。同时，通过阅读代码、写代码，能让我们对论文原理有着更深的理解。更重要的是，我还锻炼了心态、代码素养和个人的代码风格。

这次的大量代码训练，对我的编程能力、调试能力以及分析问题的能力有着很大的提升。我也体现到了把理论运用到实践的感觉。我还学到了比如考虑到时间复杂度、并发，缓存机制，协程管理等问题，而不是简单的让代码能运行就完事了。我同时也学到了很多代码上的技巧，比如前段报错时要看浏览器 network 部分的报错，看是前端还是后端的问题。搞不清楚怎么回事时，print 出来。比如之前有个报错我怎样都解决不了，print 出来中间量后，发现有个换行，原来是我发送问题时按了回车。

在与前端人员合作的时候，我们积极商讨应对在合作中出现的考验，齐心合力共克时艰，虽然屡屡报错，但经过不断的调试我们最终克服了苦难。这次实践活动不仅锻炼了我的团队合作能力、沟通能力还提升了协调部署的能力。

我也从前辈们的身上学到了他们的刻苦钻研，精益求精的精神，只有耐得住寂寞的人才才能守得住繁华。做科研，最重要的就是耐性。同时，我也认识到自然语言处理尤其是中文还有着很长的路要走。我作为一名自然语言处理研究者，愿在不久的将来能够为这个的行业的

进步以及发展贡献自己的一份微薄力量。我也期盼这个产业所带来的收获不仅可以便利人们的生活还可以优化甚至改变一些人的生活方式。

在聊天机器人运行成功之际,我感到由衷的高兴与激动,因为自己的付出终于有了收获,这种快乐是难以言表的。苦心人天不负,黎明的曙光已经到来,这项技术终要在未来发光发热。