远程科研总结报告

裴婷婷

2020.04.30

目录

一.　项目简介

人工智能是一个非常热门的话题，同时也是如今人们生活中不可或缺的一个重要工具。人工智能涉及的范围相当广泛，从我们日常生活中常常接触到的语音识别技术，例如大家都知道的 siri，到我们在试图查询一张图片来源时去各种以图搜图网站时用到的图像识别技术，再到之前的 alphago 在围棋上打遍天下无敌手，人工智能不仅是影响、方便着我们的生活，更是在无数科研领域实现着自己无可取代的作用。

在这个远程科研项目中，我学习了如何制作一个聊天机器人，用到的知识包括意图识别、指定实体识别、数据库和 API、多次对话及否定和状态机等。通过这些我将构建一个聊天机器人来达成查询 covid-19 在各个国家实时信息的功能。

二.　学习过程总结

1. 意图识别

意图识别的目的是在对话中通过提取关键词来分析说话人的目的，进而对整句话进行分析。具体实施起来会遇到很多问题，这就是我们需要通过不断研究和积累去解决的。因为人的话语组成十分复杂，意图也常常并不十分清晰明朗，所以针对意图识别有几种不同的方法，在这次学习过程中我们学习并了解了以下几种方法：

1）通过创建一个字典作为模板，在具体使用时通过分析语句的意图来和字典中的意图进行匹配。这种方法准确性不高，因为不同的两句话常常出现模样相近但表达意思南辕北辙的情况。

2）可以通过利用支持向量机(SVM)分析单词的向量特征，根据单词向量提取每个句子中的单词，比较两个句子之间单词向量的相似性，用数据和相应的意图来训练支持向量分类器(SVC)，然后利用支持向量分类器(SVC)预测目标语句的意图，经过比对从而确定句子的意图。最终获得的意图会更加准确，但是这种方法首先需要大量数据来对支持向量分类器来进行训练，

因此在实际操作中需要付出很多的时间去做前期准备，导致整个实现过程效率不是很高。

3）rasa_nlu 是接下来我要介绍的方法，在这种方法中我们需要利用 rasa_nlu 来对 json 数据进行机器学习的过程，从而能根据较少的数据对尽量多的语句进行精准的意图分析。训练成功后可以轻易地将其运行于程序中以实现对意图快速准确地识别，是效率高的更加优秀的方法。

2. 命名实体识别

命名实体识别(NER)就是在机器人接收到一句话时首先需要分析出其中具体的带有重要信息的词汇并进行分类。例如在接收到类似于“4 月 23 日我在中关村遇到了小李”的信息时，需要能将“4 月 23 日”、“中关村”、“小李”等重要信息识别出来并正确将其归类为日期，地点等。这样将各种不规则语句中的重要信息直接提取并分类的过程可以大大加快语义识别的过程并提高准确度。针对命名实体识别，相对应的在这次远程科研学习中有进行几种方法的介绍和学习。

1）首先是使用正则表达式，通过正则表达式我们能对单词进行精准提取，这样相对于对整句话进行对比的

方式显然是更有效率也更加灵活的，但是正则表达式相对的也具有很多限制，例如提取单词有比较严格的要求，无法轻易区分出正确的目标单词所属分类。

2）我们可以使用 spacy 提供的不同大小的语言包来进行提取命名实体，这是前人经过不断的积累和完善形成的非常方便的一种途径。不过其针对的实体相对不是特别灵活，对于一些特别的目标无法进行很好的提取，然而对于国家名称之类的固定的词语是十分适合的。

3）与上面的方法相对应的，在这里依然可以使用 rasa_nlu 作为一种方法来提取实体并进行分析。同样的，利用已有的 json 文件中的数据，在不同格式的句子中可以精确提取正确的实体。

3. 数据库和 API

如果想要设计一个查询类的聊天机器人，准确的数据来源当然是必不可少的关键一环。在此时我们就可以去网上寻找相对应的我们需要的优质数据库，通过相对应的 API 来调用数据，实现聊天机器人的数据查询功能。

首先我们需要的数据库可以自己建立，所有数据都需要从数据库中调取。针对不同的查询目标，建立自己的数据库对于一些查询目标来说也是一种选择。在本次项目中涉及到的相关的 COVID-19 的数据显然是不能轻易自己构建的，这时从网上已经建立起来的不断更新的数据库中获取数据当然是更好的选择。

当我们需要进行查询的是比较庞大的目标时，利用 API 去获得数据是非常便捷迅速的一个选择。我们首先需要找到一个合适的优质的网站，确保它能提供我们所需要的数据然后通过网站提供的 API 就可以轻松获取到非常庞大的数据。要注意的是很多网站提供的 API 并不十分稳定，也许会出现需要调换 API 的情况，所以找到一个稳定的网站也很重要。

4. 多次对话和否定

在设计聊天机器人时，还有一个重点也是难点的存在就是对话中的多次进行以及其中出现的否定语句。如何识别否定语句并进行正确回复是一个很值得研究的问题。在多次对话中我们还需要设计程序来让机器人在之后的交谈中能记住之前语句中的信息并在之后的对话中再次进行调用。例如在查询股票信息时需要记住之前提问的是哪支股票进而在之后的具体查询过程

中可以更有效率地对所需信息进行查询。

之所以说否定语句判断很困难是因为否定含义的表达方式是非常多样的，有很多委婉的表达都是否定含义，但想要让机器去识别这些隐含的否定含义是很困难的。首先拥有否定含义的词汇就有很多，但是这些词汇在不同状态下也不一定是否定的含义。其次有时候会用到的一些委婉表达例如“我觉得应该还有更好的选择”之类的话语，要想识别出其中的否定含义也是有一定困难度的，太多表达都不只是简单地用 no 或者 not 来表达否定含义，此时也可以选择通过训练数据来更好地对否定语义进行识别。

5. 状态机

状态机是非常具有实用性的一个数学模型，在游戏制作等领域有很广泛的应用。在不同的状态下行动具有不同的意图，尤其是在对话机器人中，根据不同状态去分析相似语句意图是非常重要的。只有在能分析出当下的语言环境时才能对意图进行正确判断进而进行正确有效的回应，这样的聊天机器人相应地也具有较高的智能。

三.　具体程序和结果展示

1.具体程序

首先准备好所需 module，并设计机器人的回复模板。def main（）主要用来连接具体的 telegram 机器人，需要用自己申请的 telegram bot 的 TOKEN 填入对应位置。

```
import logging
import requests
import telegram
from time import sleep
from bs4 import BeautifulSoup
from telegram.error import NetworkError, Unauthorized

update_id = None
countries = 'China\nItaly\nUSA\nSpain\nGermany\nIran\nFrance\nS. Korea\nSwitzerland\nUK\nNether
instructions = "Hi, I'm ABOT.I can offer you the current information about covid-19 in each cou
def main():
    global update_id
    bot = telegram.Bot('please input your own TOKEN')

    try:
        update_id = bot.get_updates()[0].update_id
    except IndexError:
        update_id = None

    logging.basicConfig(format='%(asctime)s - %(name)s - %(levelname)s - %(message)s')

    while True:
        try:
            echo(bot)
        except NetworkError:
            sleep(1)
        except Unauthorized:
            update_id += 1
```

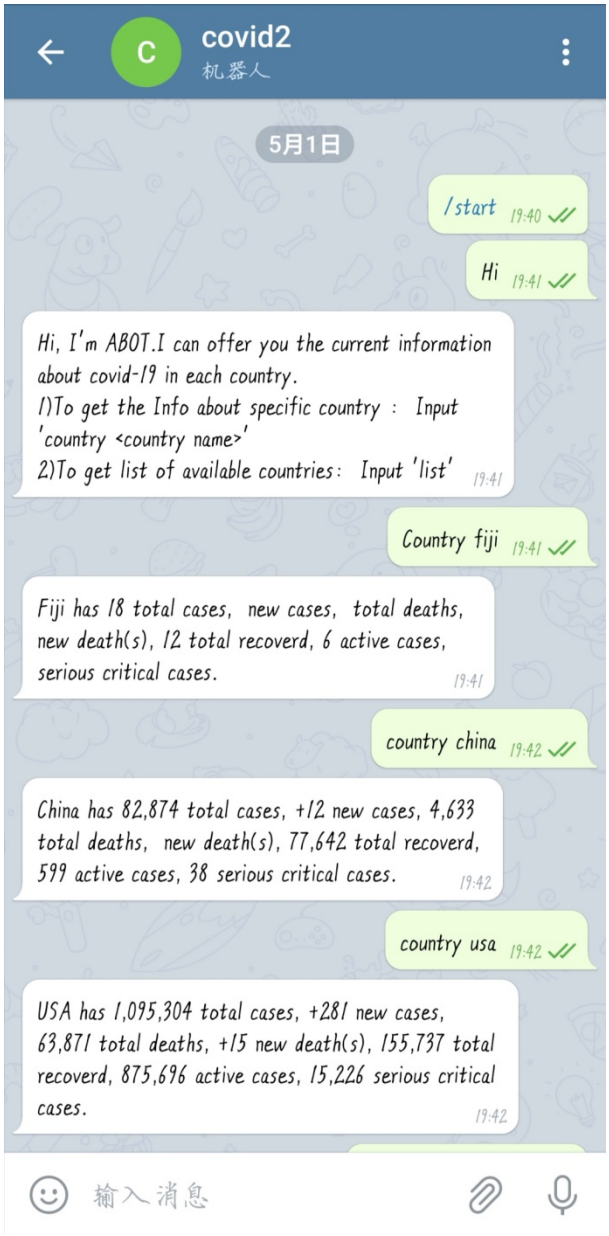
def echo(bot)设计了根据具体收到信息的内容来进行下一步的具体函数操作。例如当收到 country 内容时就会转到对应的函数来获取对应的信息并进行回复。

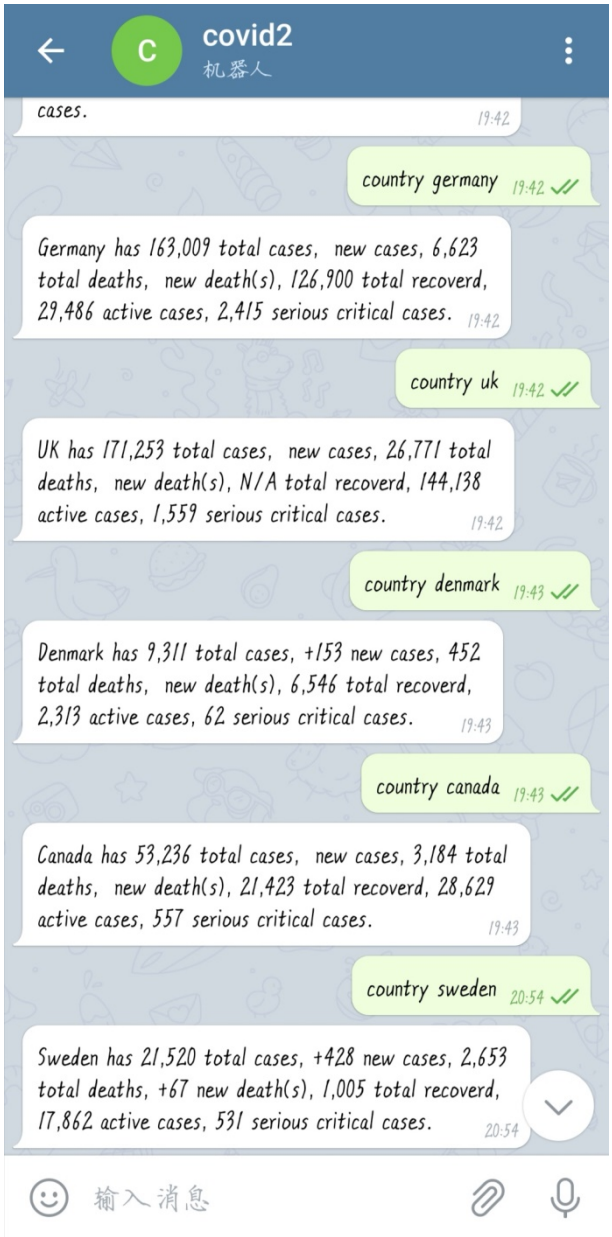
```
def echo(bot):
    global update_id
    for update in bot.get_updates(offset=update_id, timeout=10):
        update_id = update.update_id + 1

        if update.message:
            if update.message.text.lower().split(" ",1)[0] == "country":
                update.message.reply_text(data(update.message.text.split(" ",1)[1]))
            elif update.message.text.lower() == "list":
                update.message.reply_text(countries)
            else:
                update.message.reply_text(instructions)
```

def data(country)就是具体的进行 API 的调用，从对应的网址来获得所要求的具体信息并按照设计好的模板输出。

```
def data(country):
    i = 0
    page = requests.get("https://www.worldometers.info/coronavirus/")
    soup = BeautifulSoup(page.content, 'html.parser')
    table = soup.find('table')
    table_rows = table.find_all('tr')
    for tr in table_rows:
        td = tr.find_all('td')
        if i > 0:
            if td[0].text.lower() == country.lower():
                return td[0].text.strip() + " has " + td[1].text.strip() + ' total cases, ' + td[2].text.
        elif i >= len(table_rows)+1:
            return "Invalid Country"
        i =i+1
```

2.结果展示


covid2
机器人
5月1日
/start 19:40
Hi 19:41
Hi, I'm ABOT.I can offer you the current information about covid-19 in each country.
1)To get the Info about specific country : Input 'country <country name>'
2)To get list of available countries: Input 'list' 19:41
Country fiji 19:41
Fiji has 18 total cases, new cases, total deaths, new death(s), 12 total recoverd, 6 active cases, serious critical cases. 19:41
country china 19:42
China has 82,874 total cases, +12 new cases, 4,633 total deaths, new death(s), 77,642 total recoverd, 599 active cases, 38 serious critical cases. 19:42
country usa 19:42
USA has 1,095,304 total cases, +281 new cases, 63,871 total deaths, +15 new death(s), 155,737 total recoverd, 875,696 active cases, 15,226 serious critical cases. 19:42
输入消息


covid2
机器人
cases. 19:42
country germany 19:42
Germany has 163,009 total cases, new cases, 6,623 total deaths, new death(s), 126,900 total recoverd, 29,486 active cases, 2,415 serious critical cases. 19:42
country uk 19:42
UK has 171,253 total cases, new cases, 26,771 total deaths, new death(s), N/A total recoverd, 144,138 active cases, 1,559 serious critical cases. 19:42
country denmark 19:43
Denmark has 9,311 total cases, +153 new cases, 452 total deaths, new death(s), 6,546 total recoverd, 2,313 active cases, 62 serious critical cases. 19:43
country canada 19:43
Canada has 53,236 total cases, new cases, 3,184 total deaths, new death(s), 21,423 total recoverd, 28,629 active cases, 557 serious critical cases. 19:43
country sweden 20:54
Sweden has 21,520 total cases, +428 new cases, 2,653 total deaths, +67 new death(s), 1,005 total recoverd, 17,862 active cases, 531 serious critical cases. 20:54
输入消息

四．总结

COVID-19 机器人提供许多国家的具体疫情信息查询，能够在学习完课程后设计出这样一个机器人让我感觉收获很多。从之前的对机器人了解并不全面的情况下开始，一步步的从简单的回声机器人做起，过程中遇到了很多困难也尝试了很多方式，了解了很多以前没有接触过的新领域的新知识。包括在课上学习到的知识都很好地帮助我不断去学习和进步。第一次做出一个机器人并得到正确回应时的心情是激动而又倍感开心的，一个很新奇的世界开始展现在我眼前。很感谢张帆老师的授课和带领，让我在这次远程科研项目中收获良多。之后我还会继续学习更多的对话机器人相关知识，同时探索更多新的领域不断充实自己，提高自己的能力。