

Maths+Comp

COC255

B620005

CHATBOT

by

Matthew P. Rankin

Supervisor: Dr A. Soltoggio

*Department of Computer Science
Loughborough University*

April 2020

Abstract

This project aims to create a conversational software agent that uses Natural Language Processing as well as Machine Learning techniques to understand a user and respond accordingly. Non-verbal communication has amplified massively over the recent decade with instant messaging applications and texting being at the forefront of communication and encouraging knowledgeable virtual assistants, or chatbots, to increase in popularity. This paper will discuss the different techniques on Natural Language Processing (NLP) in previous chatbots created, and highlight the importance of AI and machine learning models when dealing with large amounts of data.

Contents

1	INTRODUCTION	- 3 -
1.1	BACKGROUND	- 3 -
1.2	PROJECT AIMS	- 3 -
1.2.1	<i>Goals</i>	- 3 -
1.2.2	<i>Objectives</i>	- 3 -
1.2.3	<i>Time plan</i>	- 3 -
2	LITERATURE REVIEW	- 4 -
3	NATURAL LANGUAGE PROCESSING	- 5 -
3.1	GOOGLE CLOUD NLP	- 5 -
3.2	SENTIMENT ANALYSIS	- 5 -
4	DESIGN AND DATA STRUCTURES	- 6 -
4.1	DEVELOPMENT	- 6 -
4.2	DATA	- 6 -
4.2.1	<i>Datasets</i>	- 6 -
4.2.2	<i>Knowledge base</i>	- 6 -
4.3	SYSTEM DIAGRAM	- 7 -
5	PLAN/PROGRESSION	- 7 -
5.1	REQUIREMENTS	- 9 -
5.2	CHANGES IN DIRECTION	- 9 -
5.3	SOFTWARE ENGINEERING METHODOLOGY	- 10 -
5.4	TESTING	- 10 -
5.5	WORK PLAN	- 10 -
6	EVALUATION	- 11 -
6.1	PERSONAL DEVELOPMENT	- 11 -
6.2	SYSTEM EVALUATION	- 11 -
6.3	PROJECT EVALUATION	- 11 -
6.4	FUTURE IMPROVEMENTS	- 11 -
7	CONCLUSION	- 12 -
8	BIBLIOGRAPHY	- 13 -

1 Introduction

1.1 Background

1.2 Project aims

1.2.1 Goals

- Conversational/helpful chatbot
-

1.2.2 Objectives

1.2.3 Time plan

2 Literature Review

Reviewing existing literature on this topic is essential as it forms a basis for further progression within the project. This review backs up that chatbots are a technology with increasingly high demand within all types of industries, and highlights key examples of previous successful products.

Research into Natural Language processing began in the late 1940s, with workings that have been split out from the 40s to the present now. (Jones, 2001). Many people may think of virtual assistants being something that has only become available in recent times. However, this conception is mistaken as software applications which engage in dialog with a human, otherwise known as ‘chatbots’, go back many more years. With one of the earliest NLP responding systems arising in the 60s with Joseph Weizenbaum’s Eliza (Dale, 2016). The revolution of Eliza sparked advancements in the chatbot community, Pandorabots which claims to be one of the leading chatbot platforms stated that over 300 thousand chatbots have been created by developers, implying that they’re extremely popular in today’s tech craved society.

Research has gathered lots of insight into chatbots showing how they can be so diverse to all ranges of industries, nowadays chatbots are used to solve a number of business tasks within E-Commerce, Insurance, Banking, Healthcare, Finance, Legal, Telecom, Logistics, Retail, Auto, Leisure, Travel, Sports, Entertainment, Media and many others (Kumar et al, 2019). This extensive audience would suggest that developing a chatbot would see the rise in investment opportunities for your product.

The vast growth in interest of natural language interfaces for both data and service providers suggest that industries are very much aware of the innovation advancements that a chatbot can provide for them. However, with all the interest no studies have investigated the factors driving the popularity of chatbots. An online questionnaire directed at chatbot users in the US proposed to question the users what their motivations were for using chatbots. The study gathered responses from 146 users aged 16-55 to convey that the key motivational factors sparking this interest and usage of chatbots was “productivity” (Brandtzaeg P.B., Følstad A., 2017).

Chatbots provide the patience and politeness that many humans cannot, remaining calm in high traffic demands and repeated requests. Research has shown that the use of chatbots is by far the greatest form of digital reference for many reasons (McNeal and Newyear 2016). One key reason being the anonymity of a chatbot allowing the user to no longer be afraid of asking foolish questions, as they are aware that they are only facing a computer. This provides scope for recent relevant topics such as mental health to be viewed and analysed in a new way through chatbot technology, meaning those struggling with mental health may open up more.

In order for a chatbot to train itself to imitate human like conversation it must be filled with numerous conversational exchanges. This is in the form of extracting datasets from the internet and matching up dialog between sets and subsets of utterances. There are many locations to which datasets can be extracted from, popular datasets include Cornell Movie database. This corpus contains over 220,000 exchanges of metadata-rich fictional conversations extracted from raw movie scripts involving 9,035 characters. In one example; The persona of your favourite movie star was

extracted from the data to become your own personal assistant (Nguyen et al). There are many popular datasets used for all different types of chatbot, a recent example from the Guardian saw the Zuckerberg files being exploited as a dataset to create a 'zuckerbot' and used to feed interview questions through to get a Mark Zuckerberg like response (seeing as he'd never accept an interview with the Guardian himself!) (Wong, 2019)

The motivation behind this project is amplified by the successes of previous projects within this topic. In a tech hungry culture that today's society suggests, research has shown that the demand in technology and chatbots is not only massive but is also increasing.

3 Natural Language Processing

3.1 Google Cloud NLP

Google offers API for Natural Language Processing, within this project this is the desired NLP of choice.

3.2 Sentiment Analysis

4 Design and Data structures

Within this section, development techniques and tools used will be summarised.

4.1 Development

4.2 Data

4.2.1 Datasets

As mentioned in the literature review, one of the most important resources for creating a chatbot is the dataset that the chatbot has access to, this is what determines how the chatbot learns and what personas it is given. Having browsed over Cornell's Movie database, Twitter's extensive dataset and plenty more one dataset really stood out, that being the Reddit files. A .zst file of over 7 million reddit comments was accessible online producing astonishing amounts of conversational exchanges from an almighty range of topics

4.2.2 Knowledge base

The chosen host location for the database is in AWS (Amazon Web Server) using the Amazon Relational Database Service, this is due to being able to use the cloud for running the database rather than from my own machine, creating a serverless application. Amazon, being the most prominent cloud computing provider today offers a full stack of easy to integrate different services. The aim of the relational database is to reduce users' complexity in database management by automating the common administrative tasks and therefore reduce costs through time saved.

PostgreSQL is a free open source software database management system, it is used exclusively within businesses and is the preferred database for this project. After making the database publically accessible and changing the security group inbound rules to allow requests from anywhere the Amazon RDS is connected via a simple API call from the script. Following this tutorial on connecting postgres to python: <https://pynative.com/python-postgresql-tutorial/>

```
def create_connection():
    connection = psycopg2.connect(user="",
                                   password="",
                                   host="fyp.czhkaabceukg.us-east-2.rds.amazonaws.com",
                                   port="5432",
                                   database="postgres")
    return connection
```

A modern easy to use interface allowing Postgres to be more accessible is Postico. From the client, viewing the database in real time becomes effortless due to the familiar interface of a database.

Since the connection was successful, tests were carried out in populating the database beginning with, from the command line asking a question, storing the question asked into the database. This became a success and the knowledge base could begin to be populated.

4.3 System Diagram

5 Plan/Progression

The beginning of this project saw the creation of an initial work plan to allow scope and structure within the project. As knowledge on chatbots was rather limited as well as python skills the ultimate goal for the exploration stage was set out for research into chatbots and teachings for the basic fundamentals of python.

After reading copious amounts of literature and watching several tutorials the tasks began more clear on how the chatbot infrastructure was going to be built.

Research into which Natural Language Processing API to use required some digging around until landing on the desired choice of Google Clouds Natural Language API. This came to be the NLP of choice as it offers insight into features of interest including sentiment analysis, entity analysis, entity sentiment analysis, content classification, and syntax analysis.

Connecting to the API had its complications, from following the instructions given, a Google Cloud account was created and the NLP API was enabled. The next steps were to be able to hit the NLP from terminal with the Google Cloud CLI (Command Line Interface), this stage required a lot of attention as terminal threw errors, such as not recognising the google-cloud CLI. After some installation tweaks, working with brew proved to grant recognition of the CLI and the NLP was being reached via terminal.

From here the next task was to get it firing from the python script, once installing the google-cloud-sdk to the python project, after noticing that the directory was not in the active virtualenv environment in which the google-cloud-sdk was installed to, there was success in hitting the NLP from the project.

Now access to all of Google Clouds NLP models could be accessible and used within the script and eventually implemented into the knowledge base.

The analyse_text() function allows for any parsing string to be analysed using the models that Google Cloud NLP has to offer:


```

from google.cloud import language
from google.cloud.language import enums
from google.cloud.language import types

# Instantiates a client
client = language.LanguageServiceClient()

def analyze_text(text_content):
    # The text to analyze
    document = types.Document(
        content=text_content,
        type=enums.Document.Type.PLAIN_TEXT)

    # Detects the sentiment of the text
    sentiment = client.analyze_sentiment(document).document_sentiment
    response = client.analyze_entities(document)
    sent_score = sentiment.score
    sent_magnitude = sentiment.magnitude
    # this = client.annotate_text(document)

    for e in response.entities:
        print(u"Representative name for the entity: {}".format(e.name))
        print(u"Entity type: {}".format(enums.Entity.Type(e.type).name))
        # Get the salience score associated with the entity in the [0, 1.0] range
        print(u"Salience score: {}".format(e.salience))
        # Loop over the metadata associated with entity. For many known entities,
        # the metadata is a Wikipedia URL (wikipedia_url) and Knowledge Graph MID (mid).
        # Some entity types may have additional metadata, e.g. ADDRESS entities
        # may have metadata for the address street_name, postal_code, et al.
        for metadata_name, metadata_value in e.metadata.items():
            print(u"{}: {}".format(metadata_name, metadata_value))

        for mention in e.mentions:
            print(u"Mention text: {}".format(mention.text.content))
            # Get the mention type, e.g. PROPER for proper noun
            print(u"Mention type: {}".format(enums.EntityMention.Type(mention.type).name))

    return sent_score, sent_magnitude

```

Being able to produce sentiment and entity analysis on any given string, the next phase of development was to populate the database with this information when a user inputs a question.

Currently the program stores the sentiment values for each question through calling the returned values from this function.

```

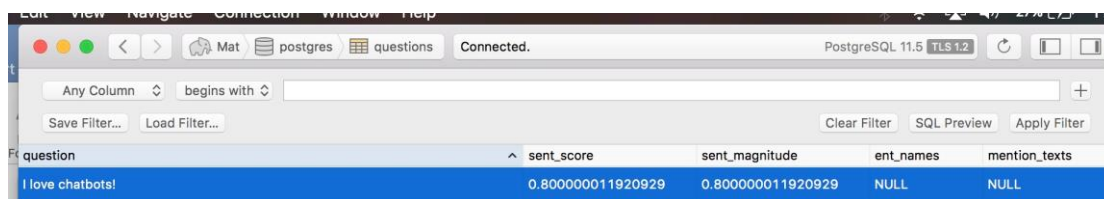
connection = create_connection()

chatting = input('Would you like to ask a question? Y/N?')
while chatting == 'Y':

    question = input('What would you like to know?: ')
    analysis = analyze_text(question)
    postgres_insert_query = "INSERT INTO questions(question,sent_score,sent_magnitude) VALUES (%s,%s,%s)"
    record_to_insert = (question,analysis[0],analysis[1])
    run_sql_query(connection, postgres_insert_query, record_to_insert)

```

As seen above the input gets parsed through the `analyze_text()` function to produce the analysis. Once done I have also created a postgres function that connects to the database and the values 'question', 'analysis[0]' and 'analysis[1]' are populated into their respected columns of the database.



question	sent_score	sent_magnitude	ent_names	mention_texts
I love chatbots!	0.800000011920929	0.800000011920929	NULL	NULL

The above picture shows when inputting 'I love chatbots!' into the program the question and sentiment scores are entered into the database.

5.1 Requirements

5.2 Changes in direction

Initial ideas

- MovieBot – A conversational tool used to help pick you choose what you wanted to watch on TV/Netflix/Prime
- CryptoBot – Users would be able to ask CryptoBot what the recent price gaps between any cryptocurrency was and ask about a future forecastings for which cryptocurrency would be worth investing. CryptoBot would have learned from trends in the past as well as live current data.
- ChitChat – Gives out non-advisory information to those struggling with mental health, ChitChat would offer 24/7 instant response times for those who need to open up to someone about how they were feeling. Potentially could work very well with those afraid to talk to a real human. However, would encounter serious problems if the bot responded in a way that triggers one's mental health state of mind.

As seen above the project product plan has seen ideas come and go, this mere brainstorming and research into practically and implications phase allowed scope of what was doable and what would help impact our society today.

5.3 Software Engineering Methodology

5.4 Testing

5.5 Work plan

Going forward in the project I need carry out several tasks,

- Firstly, carrying on from having populated the database with the sentiment scores, I need to get the entity information into the database.
- Produce a test model from extracting data from a webpage FAQs that analyses the answers and matches entities when a user asks a question
- Get the test application to choose best answer for matching entities from question and answer.
- Following this I will need to extract the reddit .zst file I downloaded and get all data inputted into a database table that I will create
- Use the matching entity process in the above point to be able to match up questions asked with reddit data
- Integrate my program with Telegram

Machine learning scope:

As well as the above work plan, I'd also like to introduce a Neural Networking system in which the bot can train itself on the decided topic.

6 Evaluation

6.1 Personal Development

6.2 System evaluation

6.3 Project evaluation

6.4 Future Improvements

7 Conclusion

8 Bibliography

Brandtzaeg P.B., Følstad A. (2017) *Why People Use Chatbots*. In: Kompatsiaris I. et al. (eds) Internet Science. INSCI 2017. Lecture Notes in Computer Science, vol 10673. Springer, Cham

Dale, R (2016) 'Industry watch' *The return of chatbots* pp. 811-817. doi: 10.1017/S1351324916000243

Ina 2017, *The History of Chatbots*, Onlim, Ina, <<https://onlim.com/en/the-history-of-chatbots/>>

Jones K.S. (1994) *Natural Language Processing: A Historical Review*. In: Zampolli A., Calzolari N., Palmer M. (eds) Current Issues in Computational Linguistics: In Honour of Don Walker. *Linguistica Computazionale*, vol 9. Springer, Dordrecht

McNeal, M and Newyear, D (2013) Streamlining Information Services Using Chatbots, *Introducing Chatbots in Libraries* pp. 5-10

V. Mateljan, D. Čišić and D. Ogrizović, "Cloud Database-as-a-Service (DaaS) - ROI," *The 33rd International Convention MIPRO*, Opatija, 2010, pp. 1185-1188. <<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5533641&isnumber=5533310>>

Wong (2019) *Interview 'I am going to say quiet words in your face just like I did with Trump': a conversation with the Zuckerbot*, The Guardian, 22 Dec 2019 <https://www.theguardian.com/technology/2019/dec/22/zuckerbot-mark-zuckerberg-facebook-botnik?CMP=Share_iOSApp_Other>