ELIZA Plus

Implementing and evaluating a long-term memory mechanism in Weizenbaum's ELIZA chatterbot

by

Pritam Sangani

Supervised by: Dr. Jim O'Shea

Manchester Metropolitan University

School of Computing, Mathematics and Digital Technology
Manchester Metropolitan University

A Dissertation submitted to Manchester Metropolitan University, as part of a final year project, in accordance with the requirements of the degree of Bachelor of Science in the Faculty of Science and Engineering.

April 2019

Word count: ten thousand two hundred and seventy-six words

Declaration

I Declare,

No part of this project has been submitted in support of an application for any other degree or qualification at this or any other institute of learning. Apart from those parts of the project containing citations to the work of others, this project is my own unaided work.

Signed

Pritam Sangani

Date

05/04/2019

Acknowledgements

I would like to thank my supervisor, Dr Jim O'Shea, for his guidance and constructive suggestions during the planning and development of this research work and writing of this report. Finally, I wish to thank my friends and family for their support and encouragement throughout my studies.

Abstract

This project will look at how a long-term memory network can be implemented in the ELIZA chatterbot. To do this a literature review was done in the areas of long-term memory, chatterbots and Natural Language Processing. Named Entity Recognition was found to be a useful field to look at during this project, because of the way it can extract key pieces of information, such as names of people, from textual data. A product was implemented to test the hypotheses defined at the start of this project and an evaluation was held with participants at the end of the implementation phase. However, it was found that there were some shortcomings in the dialogue management area of the project, but Named Entity Recognition was evaluated to be a useful technique in extracting key pieces of information that people are likely to talk about.

Keywords: ELIZA, chatbot, chatterbot, NLP, NER, Natural Language Processing, Named Entity Recognition, Long term memory, dialogue management

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence. 1, 12, 33
AIML Artificial Intelligence Markup Language. vi, 6

NER Named Entity Recognition. 2, 3, 15, 16, 24, 30, 31, 33
NLP Natural Language Processing. 1–6, 9, 12–16, 19, 33, 34

Chapter 1

Introduction

This chapter introduces the project and gives a statement of the aims and objectives, as well as the research question and hypotheses, defined before the start of this project. It will also give a brief overview of what the reader should expect from the rest of this report.

1.1 Research Background

In 1950, Alan Turing published a paper called 'Computing Machinery and Intelligence' (Turing 1950), in which he posed the question, "Can machines think?". Alongside this question, he also proposed the Turing Test - a test of a computer's or a machine's ability to replicate or exhibit the intelligence of a human. Since then, computer scientists have been researching how to develop chatterbots that converse convincingly with humans. Research, in this field, intensified when in the mid-1960s, Joseph Weizenbaum, developed ELIZA (Weizenbaum 1966) - known to be the first chatterbot ever created. ELIZA used an early-form of Natural Language Processing (NLP), where it would match patterns in the text input and substitute it with phrases, to create the illusion of understanding. In recent years, chatterbot developers have been trying to win the Loebner Prize - a modern-day version of the Turing Test. This contest has been held since 1991 in which judges converse with chatterbots not knowing whether they are talking to a human or a chatterbot (Zdenek 2001). Chatterbots have also become very popular commercially, as businesses look to make their customer services more efficient, innovative and, most importantly, more personal, due to the rising demand of customers looking for fast resolutions to their problems and the increase in time they are spending online (Bakhshi et al. 2018). This report by Deloitte, also found that one of the market forces driving chatterbot development is the "technological advances in Artificial Intelligence (AI) and NLP".

NLP is a branch of AI concerned with the research of interactions between humans and computers through natural language. Natural Language has been defined in a white paper on NLP as being "the most natural means of communication between humans, and the mode of expression of choice for most of the documents they produce" (Weischedel et al.

1989). Entity Recognition (also known as Named Entity Recognition (NER)) is a branch of NLP concerned with labelling "sequences of words in a text which are the names of things, such as person and company names, or gene and protein names." (*Stanford Named Entity Recognizer (NER)* n.d.)

A big part of developing chatterbots is actually evaluating the quality of existing chatterbots and how convincing they are at mimicking the functions of the human brain. Two recent studies have tried to answer two questions that must be answered when evaluating the quality of a chatterbot. Firstly, the current uses of chatterbots in society must be examined. Brandtzaeg and Følstad have published their study of the uses of chatterbots on various platforms and across various categories of uses. The vast majority of participants in the study reported using chatterbots to increase their productivity, by quickly retrieving information or accessing assistance (Brandtzaeg and Følstad 2017). Rather surprisingly, the study also found that 12% of the participants reported using chatterbots for social or relational use. The study found that the human nature of chatterbots drove people to use them for this purpose - the responses stated that chatterbots were a way of "avoiding loneliness" and "improving their social and conversational skills". Another study tried to gather together all the attributes and features that can be used to assess the quality of chatbots. (Radziwill and Benton 2017) The study found that two attributes that measure the effectiveness of the chatterbot are the ability to "maintain themed discussion" and to deliver "convincing, satisfying and natural interaction". The study also found that one of the attributes that measure the satisfaction of the chatterbot is the ability to "detect meaning or intent".

A big part of this project is to add a "long-term memory" mechanism to the core of the ELIZA chatterbot. Long-term memory is the ability to refer back to earlier conversations and to bring the information back at relevant points in the current conversation. This would make the chatterbot more human-like and, as discussed, would improve the quality of the chatterbot. This project aims to achieve this by using NLP, more specifically NER, to store relevant and linked information into a database to retrieve at a later point in the conversation.

1.2 Aims

The aims that were defined before the start of this project are to:

- Integrate Long-Term Memory mechanism to the ELIZA chatterbot core.
- Evaluate how convincing the long-term memory mechanism is, by way of research methods determined by research done in literature review.

1.3 Objectives

The objectives that were defined for this project from the above aims are:

1. Complete a Literature Review by reading research papers and articles on the topic of NLP and NER from the Internet and textbooks.

2. Find source code for the ELIZA chatterbot in Python and start understanding how the code works.

3. Compare the different tools for NER readily available for use in Python.

4. Design an overview of the software and a plan to implement.

5. Implement the software according to the plan, whilst incrementally testing where appropriate.

6. Carry out a final testing of the software.

7. Evaluate the quality of the software by carrying out a study with participants.

8. Write up the research findings in a report.

1.4 Research Question

The research question that has been proposed, as part of this research project, is:

Is it possible to implement convincing long-term memory into an existing chatterbot, such as Weizenbaum's ELIZA?

1.5 Hypotheses

The following hypotheses have been derived from the above research question:

- **H_0 - it is not possible to add long-term memory to a chatterbot, at this stage, which is convincing enough to be comparable to human memory mechanism.**

- **H_1 - it is possible to add convincing long-term memory at a significant statistical level, if the number of participants in the research is high enough.**

1.6 Report Structure

The rest of this report is structured as follows:

Chapter 2 - Literature Review
- A review of current research in the areas of chatterbots, NLP and the psychology of conversation.

Chapter 3 - Design
- A discussion of the considerations and requirements taken into account when selecting tools and services to use to implement the product.

Chapter 4 - Implementation and Testing
- A discussion of the steps and decisions made when implementing and testing the different parts of the product.

Chapter 5 - Evaluation
- A discussion on how well the product performed against the hypotheses set out in this chapter, using the findings from the usability and functionality evaluation tests as evidence.

Chapter 6 - Conclusion
- A final review of the project, including a personal assessment of the author's achievements and failures throughout the duration of the project. This chapter also includes recommendations of future work relating to the product.

1.7 Summary

This chapter introduced the reader to the background of the project and gave an overview of what this project aims to achieve.

The next chapter will aim to give the reader a detailed understanding of the current research in the research areas related to this project.

Chapter 2

Literature Review

This chapter will aim to provide a critical evaluation of the current relevant research that was investigated as part of this project. The research areas that are relevant to this project are in chatterbots, psychology of conversation and Natural Language Processing (NLP).

2.1 General research in chatterbots

Chatterbots are computer programs that try to simulate logical and intelligent conversations with users (De Angeli, Johnson, and Coventry 2001). In 1950, Alan Turing published 'Computing Machine and Intelligence', in which he proposed the 'Imitation Game' (commonly known as the Turing Test) to help break down the proposed question, "Can machines think?"(Turing 1950). This is a test that was described as a game. The game is played as follows:

There are 3 players: Person A (a male), Person B (a female) and Person C (the interrogator - either sex). Persons A and B are in a different room from the interrogator and the aim of the game is for the interrogator to correctly classify the genders of A and B. The interrogator does this by asking questions of A and B in a written form as to not give away gender from the tone of voice. The example question in the paper can be seen in Figure 2.1 below.

```
C: Will X please tell me the length of his or her hair?
```

Figure 2.1: Example question in Turing's 'Imitation Game'. (Turing 1950)

As part of this paper, Turing also proposed that a question and answer system, as seen in Figure 2.2, might be a suitable criteria to answer the proposed question, "Can machines think?". This is the first notion of the idea of chatterbots to enable conversation between man and machine.

```
Q: Please write me a sonnet on the subject of the
   Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give as answer) 105621.
Q: Do you play chess?
A: Yes.
Q: I have K at my K1, and no other pieces. You have only
   K at K6 and R at R1. It is your move. What do you play?
A: (After a pause of 15 seconds) R-R8 mate.
```

Figure 2.2: Example question and answer dialog proposed by Turing. (Turing 1950)

The 'Turing Test' forms the basis of the 'Loebner Prize', which is a contest held annually since 1991, in which judges hold the role of the interrogator and converse with the chatterbots on computers, not knowing whether they are speaking to a human or the chatterbot itself (Zdenek 2001). Two of the most successful entrants to the 'Loebner Prize' are called 'Mitsuku' and 'A.L.I.C.E'. Both have been implemented using a common design technique known as Artificial Intelligence Markup Language (AIML), which is a language derived from XML and was first conceived as part of the implementation of the 'A.L.I.C.E' chatterbot (Abdul-Kader and Woods 2015). The aim of AIML was to make it easier to model conversations and responses. AIML uses tags as part of the language that correspond to snippets of code, which dictates the commands sent into the chatterbot. Each command sent to the chatterbot is written as an AIML object and the general structure of these objects can be seen in Figure 2.3, whilst the most important object structure can be seen in Figure 2.4.

```
<command> List of parameters </command>
```

Figure 2.3: General structure of AIML objects (Abdul-Kader and Woods 2015)

```
<category>
        <pattern>User Input</pattern>
        <template>
                Corresponding Response to input
        </template>
</category>
```

Figure 2.4: Most important AIML object structure (Abdul-Kader and Woods 2015)

2.2 ELIZA

Ever since Turing published 'Computing Machinery and Intelligence', computer scientists have been researching how to develop chatterbots that can convincingly converse with humans. Research intensified when Joseph Weizenbaum, in 1966, published the ELIZA chatterbot (Weizenbaum 1966). ELIZA is known to be the first ever chatterbot created and it used an early-form of NLP - a pattern matching and substitution technique. This technique analyses text input and searches for patterns that match patterns in a predefined dictionary. The dictionary contains templates for responses that each pattern returns to the user. ELIZA uses scripts, containing these dictionaries, which dictate the types of responses that the chatterbot responds with. Weizenbaum's ELIZA took on the form of a Rogerian psychotherapist, which is a type of therapy developed by Carl Rogers, in the 1940s. The therapy aims for patients, who are seeking counselling to be a self-expert on themselves (Brooker 2003). The therapist facilitates this and the patient's search for self-actualisation by responding to the patient's conversation in a way that it makes the patient question his/her actions and reflect on why they are feeling a certain way. In Rogers' theory, self-actualisation meant for the patient to self-reflect and re-interpret experiences to allow for recovery and personal growth.

ELIZA doesn't understand the context of the conversation, instead, it uses the pattern matching and substitution technique, discussed above, to search for appropriate responses and phrases the responses in a way to simulate a Rogerian psychotherapist (Shum, He, and Li 2018). This also gives the illusion of the chatterbot being able to understand the conversation, which is one of the aims of creating a chatterbot that can beat the 'Turing Test'. Weizenbaum discussed the five technical problems that are associated with the implementation of ELIZA - these can be seen in Figure 2.5.

```
1. The identification of keywords
2. The discovery of minimal context
3. The choice of appropriate transformations
4. Generation of responses in the absence of keywords
5. The provision of an editing capability for ELIZA "scripts"
```

Figure 2.5: Five technical problems associated with the implementation of ELIZA (Weizenbaum 1966)

The first problem is concerned with identifying important keywords in the user input, used to trigger decomposition rules, which are patterns that the input is filtered down to. The second problem is concerned with identifying minimal context, locally within just the sentence that is being analysed. For example, if the keyword is "I", if it is followed by the word "will", this suggests that the user is going to do some action or event in the future. The third problem is concerned with choosing the most appropriate transformation of the user input to provide a appropriate response. Each of the decomposition rules, have a number of corresponding reassembly rules, which transform the user input to generate the responses that are provided back to the user. The fourth problem is concerned with generating appropriate and intelligent responses if no keywords are identified in the user input. This is to ensure that conversation can continue for longer. The final problem is concerned with providing a way for people to extend the script or provide a new script that ELIZA uses to analyse and respond to user input.

The ELIZA chatterbot is constrained to the scope of knowledge it has and the domain it has knowledge in, which is defined by the decomposition and reassembly rules, the script it is using contains. However, when ELIZA was released many people were fooled into thinking that they were speaking to a human when conversing with ELIZA (Shum, He, and Li 2018). An example of a typical conversation with ELIZA that Weizenbaum provided can be seen in Figure 2.6. This figure shows the way that ELIZA composes its responses to give the illusion of understanding.

```
Human: Men are all alike.
ELIZA: In what way
Human: They are always bugging us about something or other.
ELIZA: Can you think of a specific example
Human: Well, my boyfriend made me come here.
ELIZA: Your boyfriend made you come here
Human: He says I’m depressed much of the time.
ELIZA: I am sorry to hear you are depressed
Human: It’s true, I am unhappy.
ELIZA: Do you think coming here will help you not to be unhappy
Human: I need some help, that much seems certain.
ELIZA: What would it mean to you if you got some help
```

Figure 2.6: Typical conversation with ELIZA (Weizenbaum 1966)

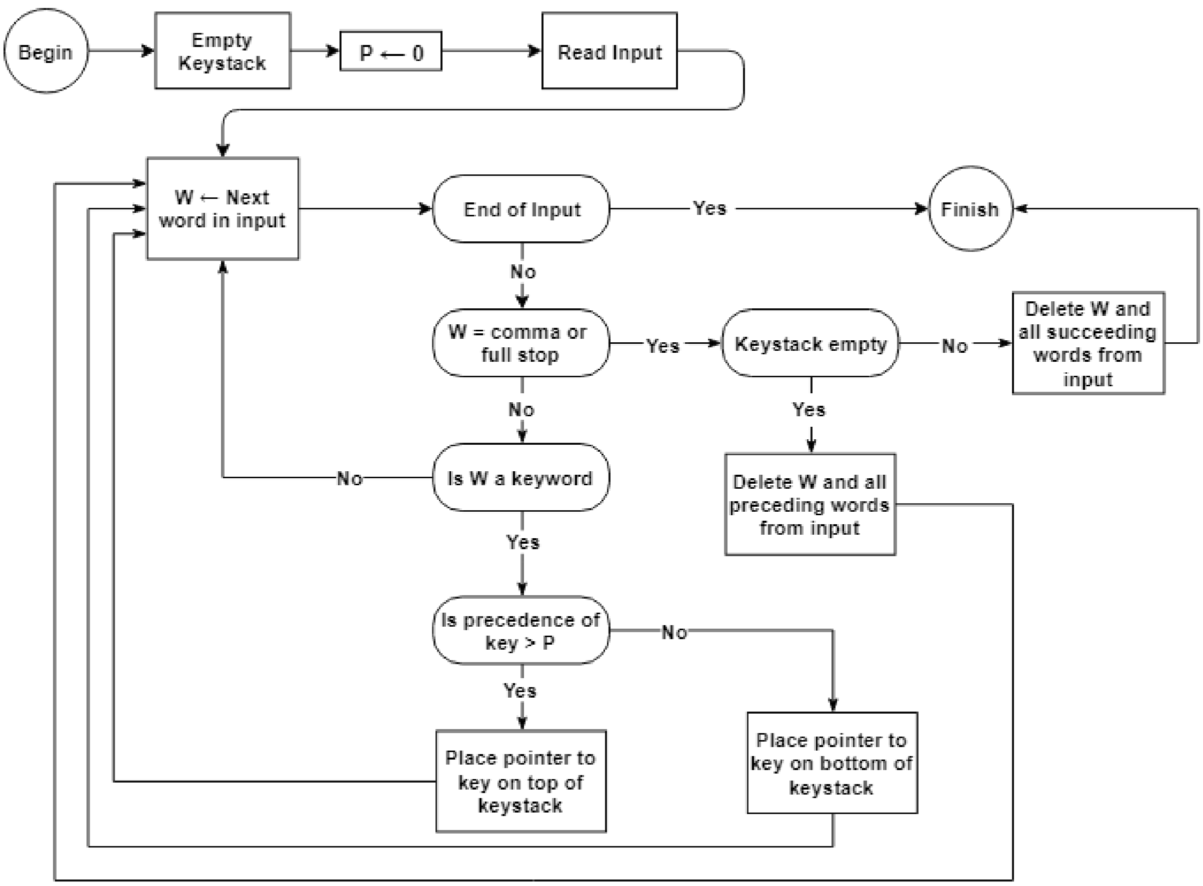


Figure 2.7: Flow diagram of keyword detection in ELIZA (Weizenbaum 1966)

Figure 2.7 above shows a simplified flow diagram of the process of keyword detection. The keystack is a stack data structure that holds all of the keys that are identified in the input. P is a pointer to a variable that holds the value of the highest key precedence that has been identified, whilst, W is a pointer to a variable that holds the current word in memory. In Weizenbaum's ELIZA, a comma or full stop was recognised as a delimiter and if the key stack was not empty, the rest of the input was discarded and the process of keyword identification ended. This process only cares about the key with the highest precedence, so if a key is found with a precedence value lower than P, then that key is placed on the bottom of the keystack.

2.3 Psychology of conversation with chatterbots

The popularity and usability of chatterbots have increased with the technological advancements in NLP, processing power available, machine learning models and the amount of data that is available to learn from (Bakhshi et al. 2018). To understand the psychology of conversation with chatterbots, the reasons why people use them and the types of conversations people have with them must be considered, as does the characteristics of the human brain a chatterbot must mimic to be considered successful.

2.3.1 Uses of chatterbots

The increase in engagement could be partly down to large companies, such as Microsoft, Facebook and Google, investing heavily into providing resources to developers to help them create chatterbots, quickly and more easily, and integrate them into social media platforms. For example, within a year of Facebook opening up their messaging platform, Facebook Messenger, to allow for chatterbot integration, more than 30,000 chatterbots had been developed and deployed (Brandtzaeg and Følstad 2017). In a recent study researching why people use chatterbots, the majority of participants commented that they use chatterbots to increase their productivity. The other main categories the responses fell under were: entertainment, social/relational and novelty/curiosity (Brandtzaeg and Følstad 2017). The responses that fell under the category of productivity felt that chatterbots allowed them to retrieve and access information more quickly and easily, which is one of the main applications of NLP, known as information retrieval, discussed in subsection 2.5.2. Perhaps, most surprisingly, was the amount of people who responded that they often used chatterbots for entertainment and social/relational use. As the quality of chatterbots increase and become more human-like (which is the goal), people are interacting with chatterbots to alleviate boredom or as a outlet to relax, "avoid loneliness" and "improve social and conversational skills". Commercially, chatterbots are becoming popular as businesses look to make their customer services more efficient, innovative and, most importantly, more personal, due to the rising demand of customers looking for fast resolutions to their problems and the increase in time they are spending online (Bakhshi et al. 2018).

2.3.2 Types of conversation with chatterbots

A recent study found that the way people converse with chatterbots differs to the way people converse with another person (Hill, Ford, and Farreras 2015). It measured the differences in conversation between those with chatterbots and those with other humans among a number of categories, including use of profanity, length of messages and the use of shorthand and emojis. The study found that in general people conversed with chatterbots for a longer period of time, but on average the length of the messages was shorter. It also found that people used a much larger vocabulary with other humans as there was less diversity, in terms of the number of unique words used in a conversation with a chatterbot. The amount of profanity used in a conversation with a chatterbot was far greater than that with humans, which suggests that people feel safe to voice their true feelings with a chatterbot than with a human, which correlates to the research done by Brandtzaeg & Følstad in their study on why people use chatterbots, which found that many people used chatterbots as a safe place to chat about their feelings (Brandtzaeg and Følstad 2017).

2.3.3 Characteristics of a chatterbot

Different parts of the brain carry out different functions and so chatterbots can have different functionality mimicking different parts of the brain. Figure 2.8 shows the different functions that the brain performs and how they relate to the architecture of a chatterbot. The three parts of the figure that this dissertation focusses on is: dialogue management (see section 2.4), natural language processing (see section 2.5) and entity recognition (see section 2.6). To be considered human-like, a chatterbot must try and mimic all parts of the brain as humans use all parts of the brain when conversing with other humans. From a commercial point of view, there are a number of characteristics that are important to consider when developing chatterbots (Bakhshi et al. 2018). Firstly, humans tend to engage more with chatterbots if they show human-like characteristics. One way this could be done is by developing the chatterbot to show emotion to the person it is conversing with. Weizenbaum's ELIZA tried to mimic emotion and the illusion of understanding by constructing responses in a way that it makes the person think that the chatterbot is understanding what they are saying. Another important characteristic is being able to reliably recognise the intent of the person chatting with the chatterbot. The chatterbot needs to be able to recognise what the user is asking for even if the message is phrased unusually, for example with spelling or grammatical errors. This is especially important in customer service chatterbots as you do not want to annoy the user.

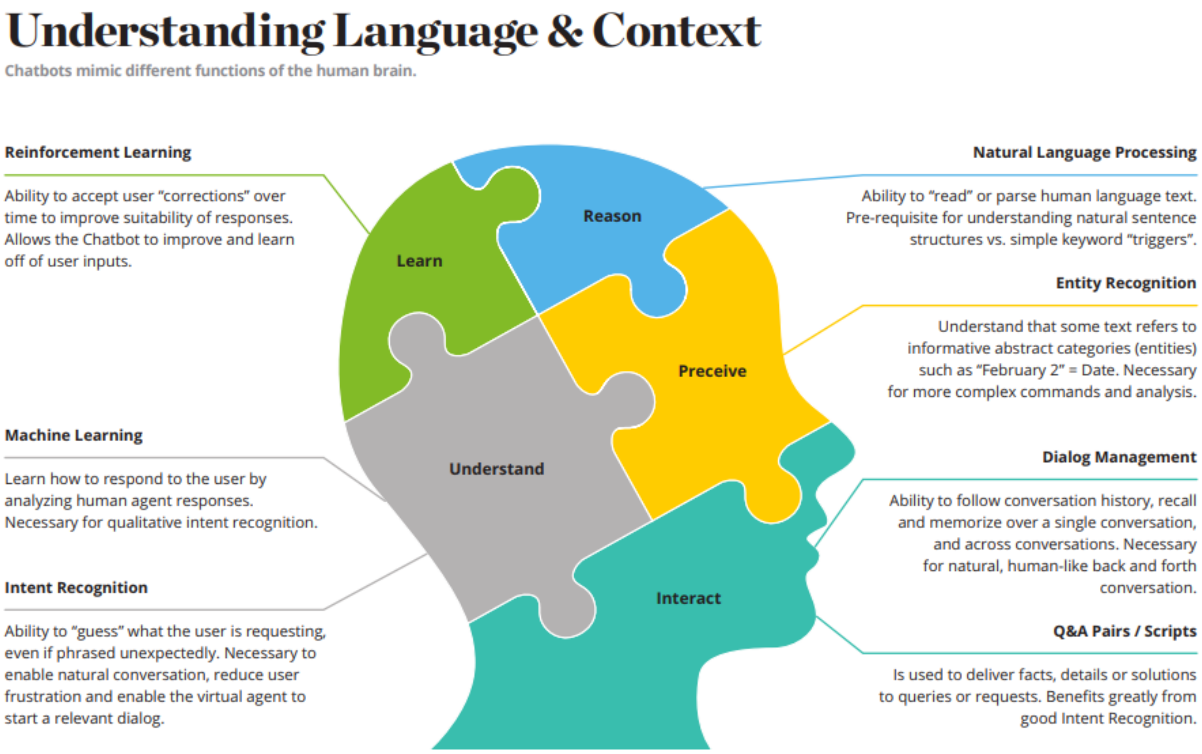


Figure 2.8: Infographic showing the different functions of a chatterbot architecture (Bakhshi et al. 2018)

2.4 Long-term memory and dialogue management

As you can see in Figure 2.8, one of the techniques that a chatterbot can utilise to mimic the brain function to interact with other people is "dialogue management". Dialogue management is concerned with building a mechanism to give the chatterbot the ability to control the flow of the conversation by interrupting the conversation flow at appropriate times to bring in knowledge that is outside of the current conversation. One way of managing the dialogue is by having a memory mechanism that stores the important parts of the conversation in a data structure and then refer back to this data structure at appropriate times to see if there is any knowledge that can be appropriately referred to in the current conversation flow. There are two types of memory mechanisms that can be used: long-term memory and short-term memory. Long-term memory is a way of recalling information into the current conversation from previous conversations. For example, if you were talking about going on a holiday with a friend, if you remember a memory of going on a holiday in the past, you can recall that information and talk about it in the current conversation. Short-term memory, on the other hand, is the ability to recall information that you have talked about in the current conversation.

In 1972, Tulving released an influential psychological insight into the different types of long-term memory (Tulving et al. 1972). Three distinctive parts of long-term memory were identified: procedural, semantic and episodic memory (Mcleod 2010). Procedural memory is the part of the long-term memory responsible for remembering how to perform certain tasks (i.e. procedures). Semantic memory is a declarative form of memory responsible for remembering worldly and general knowledge, such as meanings of words or facts such as names of places. Finally episodic memory is responsible for remembering events that have occurred in our lives in the past.

A technique that has been used with various degrees of success is Hidden Markov chain models, which is a statistical model based on the probability of a word occurring in the input. For example, given the input, "agggcagcgggcg", the markov model of order 1 predicts that the letter 'g' will occur with a probability of 8/13 (Bradeško and Mladenić 2012). This technique is not suitable to be used with the ELIZA chatterbot due to the pattern matching and substitution technique the chatterbot employs. Markov models are much more suited to a chatterbot that has a more freely built conversation flow and chatterbots that perform certain actions when triggered by certain input.

Research that was found after the implementation of the product found another technique that is being used to solve the problem of long-term memory. This technique aims to solve the problem of mimicking the decentralised nature of the human brain, which with computers storing data in a centralised and structured manner was not possible. This technique uses a deep learning algorithm known as Recurrent Neural Networks, which can help memorise and retrieve data that has been processed. To process the data a algorithm called Neural Stack Machines is used to store the textual data in a suitable data structure.

2.5 Natural Language Processing

NLP is a branch of Artificial Intelligence (AI) concerned with the research of natural language interactions between humans and computers. It has been defined in a number of ways, but the common aspects of all definitions is that NLP involves the analysis and representation of naturally occurring language, at a number of levels of linguistic analysis, for the sole purpose of achieving the accuracy of language processing, on-par with humans, for a range of tasks (Liddy 2001). Natural language has been defined in the white paper on NLP as being "the most natural means of communication between humans, and the mode of expression of choice for most of the documents they produce" (Weischedel et al. 1989) and this could be represented by written texts or spoken language.

2.5.1 Levels of linguistic analysis

Below are the 7 levels of linguistic analysis, from lowest to highest level of analysis, that a NLP system can utilise (Liddy 2001):

- Phonology
- Morphology
- Lexical
- Syntactic
- Semantic
- Discourse
- Pragmatic

Phonology deals with interpreting different spoken sounds within and across words. This type of analysis takes into account sounds when words are spoken together as a sentence, individual word sounds, as well as, sounds when words are spoken with emphasis or when a person is stressed. When speech input is passed into an NLP system, the sound waves are analysed and encoded into a digital signal, which is then interpreted by the rules employed by the system.
Morphology deals with breaking down words into its individual components. For example, the word "preregistration" can be broken down into the following components: "pre"(prefix), "registra"(root) and "tion"(suffix). This level of analysis is especially useful as the NLP system can understand the context of conversation. For example, the word "played" can be broken down into "play" and "ed". This provides information that this event happened in the past because of the suffix, "ed".
Lexical analysis is concerned with understanding the meaning of individual words, but is not concerned with whether the words make sense in the context of the whole sentence. At this level of analysis, if a word can only have a single possible meaning in the context of the sentence, a semantic representation of this word can be made. Combining semantic representations across multiple words, can help provide complex insight into the meaning of sentences, as a whole, just as humans can produce this in their brains. See Figure 2.9 for an example of a possible semantic representation of the word, "launch".

```
launch (a large boat used for carrying people on rivers, lakes harbors, etc.)
((CLASS BOAT) (PROPERTIES (LARGE)
(PURPOSE (PREDICATION (CLASS CARRY) (OBJECT PEOPLE))))
```

Figure 2.9: Example semantic representation of the word, "launch" (Liddy 2001)

Syntactic analysis is concerned with analysing whole sentences to identify the grammatical structure that makes up that sentence and identify the dependencies between the words. This is important in most languages as the order words in a sentence can change dependencies and therefore the meaning of the sentence.
Semantic analysis is concerned with making decisions about the meanings of whole sentences by considering how the different semantic representations of the words, as dealt with in the lexical level, fit together. This level also deals with reducing ambiguity of words with multiple meanings. It does this by considering the context, in which the word appears in the sentence and/or considering domain knowledge if the scope of the NLP system is limited.
Discourse analysis is concerned with analysing whole texts of documents to understand the meaning of the text as a whole. It does this by interpreting the interactions between the different sentences that makes up the text. The two most common discourse processing techniques that happen at this level are: anaphora resolution and discourse/text structure recognition. **Anaphora resolution** is concerned with replacing entities, such as pronouns, with the appropriate corresponding entity to which it refers. **Discourse/-text structure recognition** is concerned with interpreting the purpose of a sentence in a document. This can help identify parts of a document that are meaningful or important or separate the document into different meaningful sections.
Pragmatic analysis is concerned with identifying the intentions of the author of the text and the wider context of the sentence in the document. This analysis requires a lot of understanding of world knowledge and intentions of speech in day-to-day conversations. For example, if someone asks another person to "Draw the curtains", if the curtains are open you expect that person who is present to close the curtains, likewise, if the curtains are closed you expect that person to open the curtains (Briscoe n.d., example taken from p. 5).

2.5.2 Applications of Natural Language Processing

The most common applications that involve NLP can be grouped into the following categories (Liddy 2001):

- Information Retrieval
- Information Extraction
- Question-Answering
- Summarisation
- Machine Translation
- Dialogue Systems

Information retrieval is the task of finding material of an unstructured nature that satisfies a need of finding information from within large collections (Manning, Raghavan, and Schütze 2009, pp. 1-4). An example of an application of information retrieval is when you search for something on a search engine, such as Google, the query will return a list of links to pages that match your search term(s) closest.

Information extraction is the task of recognising, tagging and extracting key pieces of information into a structured format from large, unstructured collections of text (Liddy 2001). Examples of the key pieces of information that this task will try and extract are: names of people, names of companies, geographical locations and names of products. A prominent technique that falls under this category is named entity recognition, which is discussed in section 2.6.

Question-answering goes one step further than information retrieval by trying to provide a direct answer to a query rather than provide a list of relevant documents. An example of this is if you ask a question to a smart speaker, such as Amazon's Alexa-enabled devices, if it can provide a direct answer to your question it will use a third-party website to provide an answer to your question.

Summarisation is the task of reducing a large document into a smaller document. An example of an application of summarisation is that news aggregator apps or websites could summarise news articles to display a short summary of the news article, before users can click on the summary to see the full article.

Machine translation is the oldest task of NLP and the goal is automatic translation of a text in one natural language into another, for example, English to Arabic, preserving the context and meaning of the original text as close as possible (*Stanford NLP Group - Machine Translation* n.d.).

Dialogue systems is an area of NLP with a lot of active research and the goal is to achieve human-like dialogue or conversation, whether that be using typed conversation or oral communication.

2.6 Named Entity Recognition

Named Entity Recognition (NER) is a sub-task of NLP concerned with identifying information units, know as entities, such as names of people, places, products and companies, as well as, numerical units, such as dates, times and monetary expressions (Nadeau and Sekine 2007), within textual documents. Figure 2.10 shows a visualisation of entities being extracted from a body of text.

Figure 2.10: Figure showing a visual representation of entities being extracted from a body of text

2.6.1 Named Entity Recognition Tools

A study was completed comparing the leading tools and libraries for NLP and NER across a number of programming languages in 2015 (Choi, Tetreault, and Stent 2015). It found that overall spaCy had the best accuracy and speed when training and using their models. spaCy is also written in Cython, with a very high level wrapper written in Python making it easy to get started and apply an NLP model. Python is also a simpler language to learn than a language like C++, where performance will be optimal, but it would take a long time to learn the language and implement the product. There are other libraries, like CoreNLP and StanfordNLP, but they are written in Java, which would require much more effort to write an application as it is a lower level wrapper. For prototyping, Python is one of the best languages as it has a simple syntax and a rich ecosystem of libraries, which are often high-level making it easy for beginners to get started.

2.7 Summary

This chapter introduced the reader to research that has been undertaken in the areas of chatterbots, psychology of conversation and NLP. The next chapter will discuss the requirements of the product and the design decisions made prior to the start of the implementation phase.

Chapter 3

Design

This chapter covers the decisions made and the reasoning behind these decisions during the design and planning stage before the implementation phase of this project. It will first cover what was decided what the requirements of the product were going to be, before a high-level discussion and comparison of design methodologies that would be suitable for this project. This chapter will also discuss and compare all the languages, tools and services that could have been chosen to be used for implementation and provide all the plans and diagrams used for planning the project.

3.1 Product Requirements

When thinking about and documenting the requirements of the product, two well-known and popular frameworks are, the user stories framework and the jobs to be done framework. The user stories framework allows requirements to be written from the perspective of a user/persona and the jobs to be done framework allows requirements to be written to set out actions that are executed when a specified event is triggered. From a high level, the statement for the overall requirement for the product is as follows:

"A chatterbot that allows a user to login/signup and converse with the chatterbot. Additionally, the chatterbot will be connected with a database and have a long-term memory mechanism that allows it to remember conversation."

The jobs to be done framework was chosen, for writing the requirements as there aren't really multiple user groups and this framework is useful for documenting, at a high level, all the events that could occur and their expected outcomes. Additionally, the requirements were broken down into two categories, functional and non-functional requirements. Functional requirements are those that are concerned with features of the system, such as logging into the app, whilst non-functional requirements are concerned with how well the system should perform, such as system security and ease of use.
Listed below are all of the functional and non-functional requirements for the system.

Each requirement has a corresponding priority, with Priority 1 being of the highest priority and Priority 4 being the lowest.

3.1.1 Functional Requirements

1. When a user visits the chatterbot for the first time, they should be able to signup, so they can interact with the chatterbot. **Priority 1**

2. When a user returns to the chatterbot, they should be able to login, so that they can interact with the chatterbot again. **Priority 1**

3. When a user logs in, the chatterbot should display all the previous messages in the chat, so that the user can see and refer back to them. **Priority 1**

4. When a user is logged in, they should be able to send a message to the chatterbot, so that they can converse with the chatterbot. **Priority 1**

5. When the chatterbot receives a message from a user, the chatterbot should send a message back, so the user can receive a response to their user and continue the conversation. **Priority 1**

6. When the chatterbot receives a message from a user, the chatterbot should access the past conversations with the user, so that the chatterbot can see if it can refer back to any topics or entities in the current conversation and change the current topic of the conversation accordingly. **Priority 1**

7. When the chatterbot responds to a message from the user, both the user's message and the chatterbot's response should be displayed, so the user can see and follow the conversation. **Priority 1**

8. When any messages are sent to/from the chatterbot, both the user's message and the chatterbot's response should be stored, so the user can see the past conversation and the chatterbot can access the past conversation. **Priority 1**

9. When a user logs in, the chatterbot should send a generic greeting to the user, so that the user feels welcome and it serves as a conversation starter. **Priority 2**

10. When a user logs in, the chatterbot should send a personalised greeting, based on the user's sentiment from the last time the user conversed with the chatterbot, so that the chatterbot can gauge the user's current mood. **Priority 3**

11. When a user is logged in, the user should have an option to sign out, so that the user can safely exit the program. **Priority 3**

12. When a user sends a message, there should be a short delay before the response is displayed, so that the user thinks that they are talking to a human. **Priority 4**

3.1.2 Non-Functional Requirements

1. The system should have an easy-to-use and intuitive user interface. **Priority 1**
2. The system should keep personally identifiable user data secure. **Priority 1**
3. Users should not have access to other users' conversation data with the chatterbot. **Priority 1**
4. The system should securely authenticate users. **Priority 2**
5. The system should quickly retrieve and send messages from the message storage facility. **Priority 3**
6. The system should not crash if too many messages are sent to the chatterbot by the user. **Priority 4**

3.2 Design Methodologies

It has been decided that the best methodology for this project is the Agile design framework. This is because it will allow for quick feedback for features and allow for changes to be made quickly if more useful literature can be found, which requires changes to be made to the implementation to improve the performance of the long-term memory. A close contender was the Waterfall method, as the author is very familiar with this methodology and so if the Agile methodology doesn't suit the project, the methodology will probably be changed to the Waterfall method. One thing to note is that tests will have to be run frequently at the end of each 2-week sprint and so if features are not being implemented quickly enough then the timeframe could be jeopardised.

3.3 Languages, Tools and Services

After reviewing the languages, tools and services that would be suitable for this project, the following were chosen:

User Interface:
Languages - HTML5, CSS3, JavaScript, Python
Tools - Socket.io (for web sockets)
Services - Google Firebase (for secure user authentication and real-time database)

Chatterbot:
Languages - Python 3.6
Tools - Flask server (to host and run web application), Socket.io (for web sockets), spaCy (Python package for NLP)

Services - None used

3.4 Database Schema

Google Firebase provides a service called Firestore, for real-time database storage. It uses a NoSQL data model and so is much more flexible than an SQL data model, as complex objects can be stored in a scalable and hierarchical structure. Firestore organise objects (known as documents) into individual collections and the database schema, shown in Figure 3.1 reflects this.

There are two main collections in the database: **chats** and **users**. There is also a subcollection for **messages**. The messages and users collections both contain an embedded JSON object that holds the data for entities that have been found in the user's messages. Firestore doesn't support foreign keys so the users collection holds a key for the id of the chat the user is associated with so the messages can be retrieved for the specific user.

The entities object shown is an embedded JSON object, known as a map, and is not its own collection as there isn't a need for retrieving just the entities, as they are related to a specific user, and unlike the messages collection, all of the entities will have to be retrieved every time a user logs in to the application.

3.5 System Overview

Figure 3.2 shows an overview of the different components of the system and how they communicate and transfer data. The web application, which is going to be implemented in HTML, CSS and JavaScript, will communicate with the chatterbot, which will be implemented in Python, via web sockets. There are libraries, available both in Python and JavaScript which implement the same standard of web sockets making this functionality simple to implement and transfer small amounts of data. The web application will make use of Firebase Authentication, to securely authenticate users, and Firestore, as the database service. Both services will communicate to/from the web application using http calls via JavaScript libraries. Finally the web application will be hosted on a Flask server, which can be run on a localhost or can be deployed on a hosting platform, such as Heroku. One limitation of Heroku is that the free tier shuts down the server after 1 hour of inactivity and so there may be a short delay before the application becomes live again.

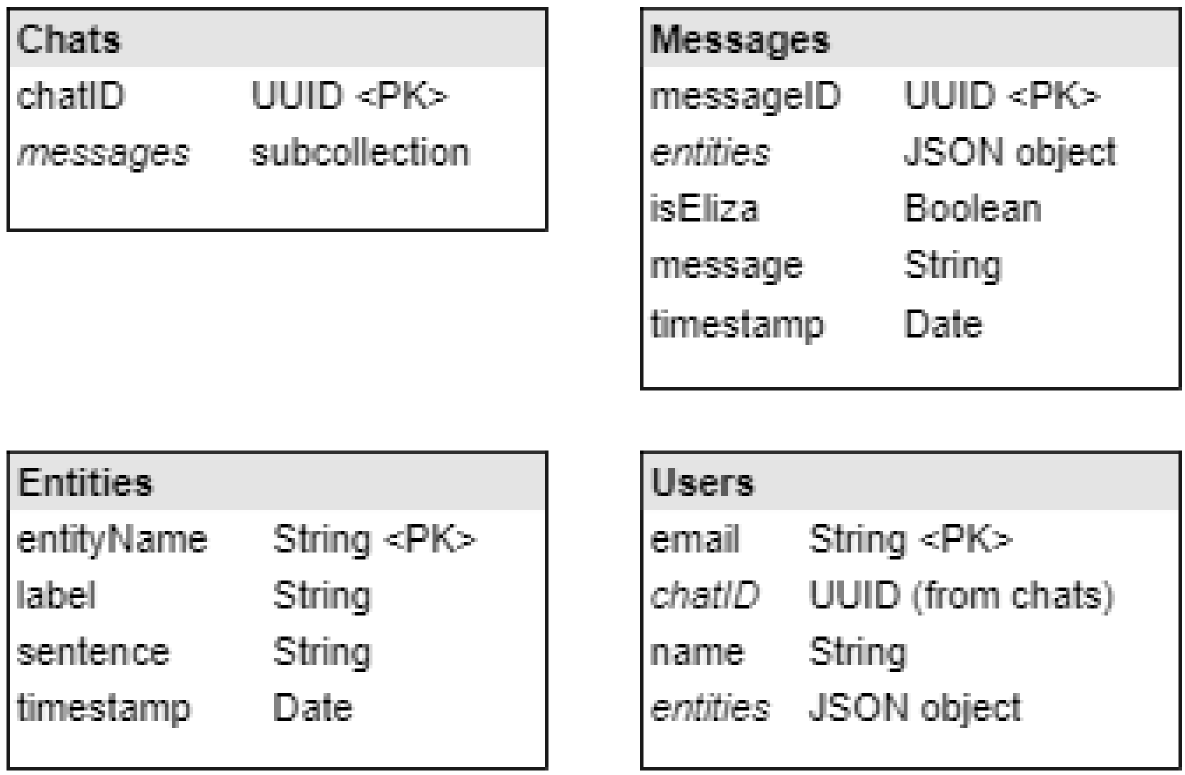


Figure 3.1: Diagram showing the schema used for the database structure

3.6 Summary

This chapter gave the reader a detailed look at the planning that was done prior to the implementation phase of this project. It first outlined the functional and non-functional requirements, including the level of priority. The functional requirements were written using the popular jobs to be done framework to show the different functional events and their expected outcomes. The design methodologies, programming languages, tools and services used to implement the product were then introduced and finally the database schema and system overview diagram were discussed.

The next chapter will discuss the work carried out during the implementation phase of this project.

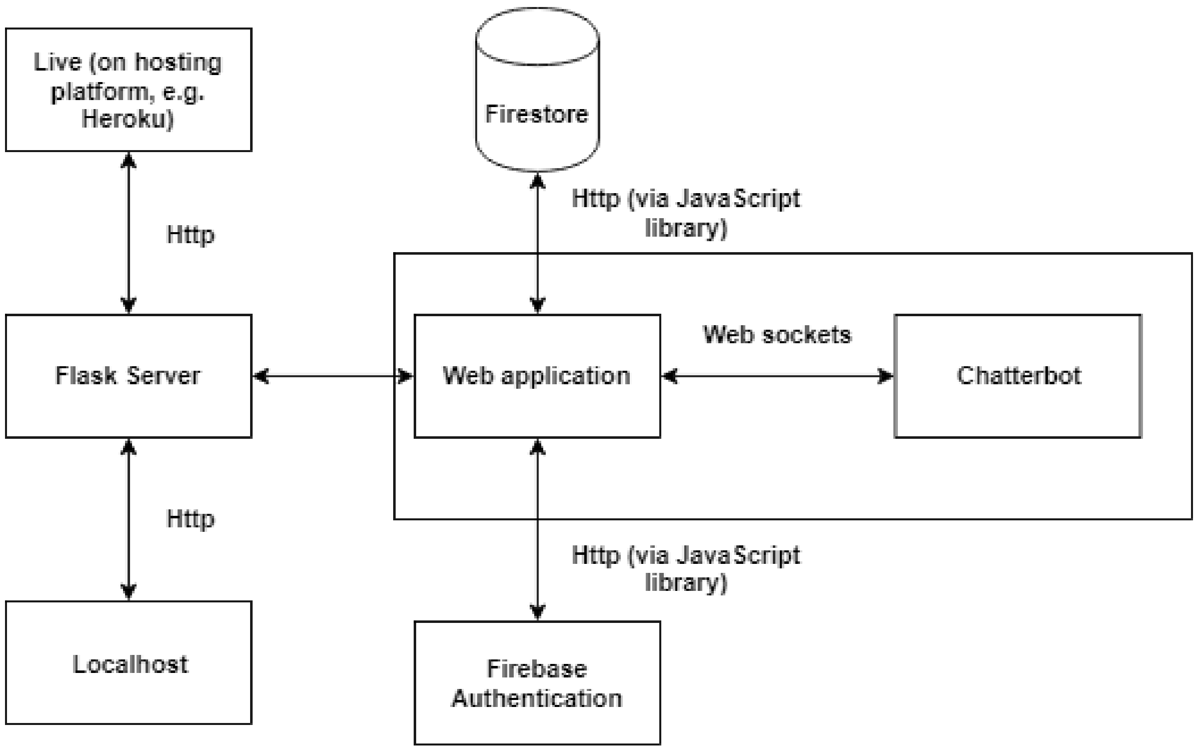


Figure 3.2: Diagram showing an overview of how the different components of the system are connected together

Chapter 4

Implementation and Testing

This chapter discusses the tasks that were completed during the implementation phase of this project. It is broken down into the different sections/components of the system and includes the testing strategies that were used to test the completeness and functionality of each feature of the system. To see the full source code, visit the link provided in Appendix C.

4.1 Core chatterbot

The core of the chatterbot was implemented in Python. It was adapted from two repositories on GitHub at the following links:

`https://github.com/jerryxu178/Chatbot-Eliza`
`https://github.com/jezhiggins/eliza.py`

The chatterbot is mainly made up of a dictionary of patterns and corresponding responses, which will be used to match a message from the user to a pattern in the dictionary and respond with a random response from the dictionary. In psychology, the type of language used in this dictionary is known as psychobabble. Figure 4.1 shows a small snippet of what the dictionary looks like. The dictionary includes a pattern as the key for each item in the dictionary. The 'r' at the front of each pattern denotes that the phrase uses regular expressions, so the string is not parsed as a normal string, but is parsed as a regular expression by the in-built regular expressions library. The value associated with each key is a dictionary of responses and most of the responses includes a '0' in the string. This will be replaced by the text that falls within the brackets in the pattern associated with the response. For example, given the input "I need to go meet my mother today", the pattern that will be matched will be "r'I need (.*)'". If the response is "Why do you need 0?", the response displayed to the user will be "Why do you need to go meet your mother today?". The reason why the response doesn't say "meet my mother" is because a process to reflect personal pronouns occurs to change the pronoun used in the response. This happens by going through each word in the response and checking if the pronoun is in a pre-defined dictionary of pronouns and if it does replacing it with the corresponding reflected pronoun.

```
psychobabble = [
    [r'I need (.*)',
     ["Why do you need {0}?",
      "Would it really help you to get {0}?",
      "Are you sure you need {0}?"]],

    [r'Why don\'?t you ([^\?]*)\??',
     ["Do you really think I don't {0}?",
      "Perhaps eventually I will {0}.",
      "Do you really want me to {0}?"]],

    [r'Why can\'?t I ([^\?]*)\??',
     ["Do you think you should be able to {0}?",
      "If you could {0}, what would you do?",
      "I don't know -- why can't you {0}?",
      "Have you really tried?"]],

```

Figure 4.1: A small snippet of code showing what the psychobabble dictionary looks like.

4.2 Long-Term memory mechanism

The long-term memory mechanism utilises NER and to implement this the spaCy library is used to quickly pass in input to a in-built model. The 'en_core_web_md' model is used, which is trained on a web corpus from Wikipedia. Entities are extracted by passing in the input into the model as shown in Figure 4.2. The entities extracted are added to a JSON object, which includes the entity extracted from the text as the key and the full sentence and the entity type (label) as the values. These entities are passed back to the user interface once the response has been formulated, which can then be added to the database.

The following steps take place in the process of interrupting the conversation flow to access the memory:

1. The chatterbot core finds a match for the user input from the psychobabble and returns a random response.

2. The response is then checked to see if the response is appropriate enough to access the memory. It does this by using a regular expression to check for square brackets in the response. As seen in Figure 4.3, some responses contain square brackets, which include some placeholder text for the entity labels accounted for.

```python
current_entities = {}
doc = nlp(statement)
print("Doc: ")
ents = {}

for e in doc.ents:
    ents[e.text] = {
        "sentence": f"{doc}",
        "label": e.label_
    }

print(ents)
current_entities.update(ents)
```

Figure 4.2: Figure showing how entities are extracted from user input

3. If there isn't a match for the regular expression, the original response is returned. If there is a match, the match is split into the individual labels so that it is easier to know which entity types can be substituted into the response.

4. If any entities have just been extracted, the dictionary containing the entities will be checked to see if there are any with the same entity type as one of the placeholders. If there are the placeholder will be replaced by the entity and if the response is not in the recent memory, the response will be returned.

5. If no entities have been extracted recently, the past entities will be accessed and if the entity has not been talked about recently, the chatterbot will return the following response, where "[ENTITY]" will be replaced by an actual entity:
   "Let's change focus. Previously you have talked to me about [ENTITY]. Please tell me more."

```python
[r'(.*)',
 ["Please tell me more [''/about PRODUCT/about PERSON/about GPE]",
  "Let's change focus a bit... Tell me about your family.",
  "Can you elaborate on that?",
  "Why do you say that {0}?",
  "I see.",
  "Very interesting.",
  "{0}.",
  "I see.  And what does that tell you?",
  "How does [that/PERSON/going to GPE] make you feel?",
  "How do you feel when you say that?"]]
```

Figure 4.3: Figure showing some entities containing placeholder text for some entity labels.

4.3 User Interface

The user interface was adapted from the following GitHub repository:

`https://github.com/keithweaver/eliza`

It is written in HTML5, CSS3 and JavaScript, using the Bootstrap framework for the front-end. For user authentication, Firebase Authentication service was used and the FirebaseAuth-UI library was used to make it easier to implement the authentication flow and so the front-end of the authentication service did not have to be implemented.

When the user signed up, a chat node was created first, with a UUID acting as the chat id before the user node was created in the database with the email being the unique id and the chat id of the chat node created being added to the user node to make a reference to the chat the user is associated with. When the user logs in, all the messages, which are stored in the messages sub-collection of the chat node, and entities, which are stored in the users node, are retrieved from the database. The chat messages are then displayed on the screen before the interface connects to a web socket, which emits a message to the chatterbot notifying it to send a greeting. When the user sends a message, the message is displayed and the socket which the chatterbot is connected to, emits the message along with all the entities. When the response is returned, the response is displayed and then sent to the database along with any entities that have been extracted.

4.4 Testing

To test the authentication flow, messages were printed to the console, to debug where in the flow the program had been executed. This was useful as initially, there were problems with signing up and creating the chat and user node correctly. This testing strategy helped see which functions were being executed and in which order.

To test the chatterbot, again messages were printed to the console to see the responses, see what matches were being made and also what entities were being extracted. To test the memory mechanism, the different dictionaries that were being used as memory, such as the past entities, current entities and recent responses, were logged to the console to see if the logic that was implemented was correct and if they were being used as expected.

To test the overall system at the end of the implementation phase, the chatterbot was used as intended and the chat logs were viewed to see if the responses being formulated by the chatterbot were as expected and if the memory mechanism was picking up any topics of conversation in the past.

4.5 Summary

This chapter gave the reader a detailed insight into the implementation details of this project. It looked at how the core of the chatterbot, long-term memory and user interface were implemented and subsequently pieced together. This chapter also gave the reader information about how the system was tested for completeness and functionality. Where appropriate, snippets of code were given to explain exactly how a certain feature was implemented.

The next chapter will discuss how the system was evaluated to test how far the system goes in achieving the aims and objectives of this project.

Chapter 5

Evaluation

This chapter will discuss in detail what methodologies were used to evaluate the quality of the product and details on exactly how the product was evaluated and the materials given to the evaluators, before discussing the results of the evaluation.

5.1 Evaluation Methodologies

To evaluate the product, it was decided that for this project both a usability and functionality evaluation should be undertaken. As there were three specific entity types that had been accounted for (names of people, names of products and names of places), it was also decided that there should be three scenarios written to guide the participants to talk about one of the three entity types - these scenarios can be found in Appendix D.

There were five participants in total and they were each given a scenario to follow. As this project was evaluating a long-term memory mechanism, the participants interacted with the chatterbot for two 15-20 minutes sessions. For the first session, the participants were told to stay within the bounds of the scenarios they had been given for the full session and for the second session they were told to stay within the scenario for the first 10 minutes, with the rest of time dedicated to the participant being able to freely interact with the chatterbot. Before the participants interacted with the chatterbot, they were given the questions that were going to be asked in the questionnaire after they finished evaluating the chatterbot. The questionnaire consisted of 9 questions with a range of questions relating to the functionality and usability evaluation criteria. The questionnaire questions can be found in Appendix E.

5.1.1 Functionality Evaluation

The functionality evaluation was evaluating the functional requirements of the system listed in sub-section 3.1.1. The main features evaluated as part of this evaluation were:

- Login/Signup flow - make sure that messages are retrieved when logging in and only a greeting is displayed when signing up.

- Sending messages - make sure that when the user sends a message, a response is displayed from the chatterbot.
- Memory mechanism - make sure that the chatterbot can pick up topics of past conversations and refer to them in the current context.

The requirements with priorities of 3 and 4 were not implemented and so were not part of this evaluation.

5.1.2 Usability Evaluation

The usability evaluation was evaluating the non-functional requirements of the system listed in sub-section 3.1.2. The main features evaluated as part of this evaluation were:

- Ease of use - the chatterbot should be easy to use and require minimal instructions to be able to use.
- Security - personally identifiable data should be securely stored in a database and users should not have access to other users' data.
- Reliability - the chatterbot should not crash unexpectedly and should always return messages to the user interface.

All non-functional requirements were implemented and so will all be part of the evaluation.

5.2 Findings

The full findings can be viewed in Appendix E - this section will discuss the summary of the findings for both the functionality and usability evaluation.

5.2.1 Functionality Evaluation

Questions 3, 4, 5, 6 and 7 were related to the functionality evaluation. All participants responded that the chatterbot always responded to a question they asked, whilst just over half of the participants found the chatterbot to be welcoming. More participants could have found the chatterbot to be more welcoming if the greeting message wasn't generic and included their personal name. Unsurprisingly, most participants found the unique nuances of ELIZA to be engaging and entertaining, but it was disappointing to learn that only just over 50% of participants found that the chatterbot was picking up topics of conversation that was talked about in the first session. Even more disappointing was that most participants found that the chatterbot was unable to hold themed conversation for a long period of time.

These results, although disappointing demonstrate the complex nature of conversation and memory. Positives can be found, in the sense that the findings also show that NER can indeed extract key pieces of information that people talk about and if more work is done in the area of memory management and efficient data structures, NER can be a useful technique in a long-term memory mechanism.

5.2.2 Usability Evaluation

Questions 1, 2, 8 and 9 were related to the usability evaluation. All participants responded saying that the user interface was easy to navigate and intuitive so instructions were not needed to show how to use the application. Every participant also said that the application met the reliability criteria as they all responded saying that the application did not crash or break unexpectedly. For the concluding questions of the questionnaires evaluating the application as a whole, most people said that they were somewhat likely to interact with the ELIZA chatterbot again, given the chance.

Question 9 was a free-text question that enabled the participants to give feedback that was not already given as part of the preceding questions. A common negative theme across the responses to this question was that the chatterbot was moving on to new topics too quickly and that it wasn't probing the participant enough for more information about a topic of conversation, ultimately annoying the user. On the other hand, a positive theme was the interesting nature of the pattern matching and substitution technique that the chatterbot employs, which gripped the user.

This shows that overall the usability of the chatterbot was of a high quality and that many of the reasons why there was negative feedback was down to the functionality of the chatterbot, rather than the usability, which was to be expected given the naive approach that was taken to keep the authenticity of the ELIZA chatterbot.

5.3 Summary

This chapter firstly discussed what evaluation methodologies were used and the criteria for the functionality and usability evaluations. It also discussed what information was given to the participants prior to them interacting with the chatterbot. Finally the summary of the findings from the evaluation were discussed, which found that the chatterbot's memory mechanism was not the most accurate or of the highest quality, but positives can still be taken from the work undertaken.

Chapter 6
Conclusion

This chapter will review the project as a whole and summarise the findings from the evaluation and will include a personal evaluation of how well the author of this report completed this project before making some recommendations on further work that can improve this area of study.

6.1 Project Review

At the start of the project, the aims and objectives were set out to help define what the author wanted to achieve from the project. The aim of the project was to implement a long-term memory mechanism in the ELIZA chatterbot and evaluate the memory's performance. This formed the basis of the research question and hypotheses, which were:

RQ: Is it possible to implement convincing long-term memory into an existing chatterbot, such as Weizenbaum's ELIZA?

H_0 - it is not possible to add long-term memory to a chatterbot, at this stage, which is convincing enough to be comparable to human memory mechanism.

H_1 - it is possible to add convincing long-term memory at a significant statistical level, if the number of participants in the research is high enough.

Once the project aims and objectives had been decided, a literature review was completed to synthesise current research and decide on the most appropriate techniques and tooling to implement the product with that would help answer the proposed research question. The most important areas of research was in the areas of ELIZA, the psychology of conversation, including long-term memory and Named Entity Recognition, a popular Natural Language Processing technique. The research in the psychology of conversation helped provide guidance in the most prominent topics of conversation that people talk about and the research in the area of Named Entity Recognition helped in learning about how NER works and the types of entities the technique can extract from textual data.

The design phase of the project involved planning and visualising how the system would look like. The first task was to list all the requirements of the project and categorise them into functional and non-functional requirements, which helped in defining the questions for the usability and functionality evaluation. Diagrams were also drawn up to visualise the schema for the NoSQL database and a system overview diagram to visualise how the different components of the system communicated with each other. The design phase helped to save a lot of time during the implementation phase as all of the requirements had been identified and the best design of the system had been researched, as well as the technologies and libraries that were most suitable for this project.

During the implementation phase, the design plans were followed closely, however, the methodology changed from Agile to Waterfall because it turned out to be difficult to implement and debug some features, which meant that enough features were not being implemented in the agile time frame of two weeks. Another reason why the Waterfall method was used is because it was quite hard to test individual features as a lot of features depended on other features. Therefore, it made sense to test the whole system as a whole at the end of the implementation phase, with some minor testing done along the way. The main difficulties in the implementation was learning how to use the libraries that were used, such as spaCy and socket.io, and also learning how to program efficiently in Python. It was also challenging to work out how to best approach and implement the interruption in conversation flow to access the memory, which contained the entities that had been extracted from the user's previous chat messages. Another key point to mention is that originally the mode of communication between the web interface and the chatterbot was to be via an API that was going to be built using the Express.js package in the Node.js runtime environment. However, this was taking too long to implement and a easier way, but less efficient and secure, was found in the form of web sockets - the design diagrams shown in Chapter 3 reflect this change.

To answer the research question and evaluate the hypotheses, a usability and functionality evaluation was held with 5 participants. The participants were give one of three scenarios, which tested each of the three entity types that were accounted for in the implementation. The entity types accounted for were names of products, names of places and names of people. The participants had two sessions with the chatterbot to test the accuracy and quality of the long-term memory mechanism. The overall feedback for the chatterbot was that the chatterbot was engaging, but the long-term memory mechanism didn't pick up every key piece of information that was talked about and when the chatterbot referred to topics talked about in the past, sometimes the chatterbot moved on too quickly to other topics and didn't probe fully about key topics. These limitations are both down to the accuracy of models that implement NER and the naive approach that was taken to implement the conversation flow, due to the complex nature of conversation.

From the evaluation completed, it can be evaluated that hypothesis H_0 is the most true and more work is needed to move towards a more human-like memory mechanism.

6.2 Personal Evaluation

This section is dedicated to allow the author to personally evaluate his performance during the project and allow him to identify his shortcomings and successes to better his performance in subsequent projects.

I felt my performance on this project had a lot of positives, but I acknowledge that I had some shortcomings along the way, which meant that the quality of the product was not as high as it could have or should have been. I feel that I did not plan the project with enough detailed plans, which meant that I had to change my approach a couple of times during the implementation period. I also should have synthesised my literature study, when I did it at the start of the project as I had to look back at the same papers during the implementation. Doing this would have saved me a lot of time, both during the implementation and during the writing of this report. Another shortcoming is that I didn't follow the implementation plan and so quickly fell behind on my work for this project.

However, there are a lot of positives to take out of this experience. Firstly, I can say that I am proud of learning how to program in Python in such a short amount of time and also I have improved my project management skills, which I can transfer and use throughout my career. I have also learnt a lot about the interesting and fast growing area of NLP and can see the importance of the research that has already been done and will be done in the future.

Some skills that I have developed are project management, programming and problem solving through managing this project, building the long-term memory mechanism in Python and I developed my problem solving through learning a new language and working with a new AI technique that I hadn't used before. Some skills that I have identified that need work are time management and project management. I plan to work on these 2 skills by working on a few projects after my exams are over and taking the time to plan out each project properly and setting out timelines as to when I should finish certain tasks. I also need to work on working out how to keep myself motivated when I run into technical or implementation difficulties, as I lost motivation when I ran into problems during this project.

6.3 Recommendations for further work

To take this project further, more work should be done in researching more suitable data structures to hold and connect different pieces of data together. This is because at the moment, the product that has been implemented doesn't implement a suitable data structure to hold multiple pieces of information about a single topic and so this hindered the dialogue manager to bring in suitable information about a topic that was previously talked about. Another area of research that must be looked into further is in dialogue management and how to probe for more information about a topic and then link the newly found information with the topic as that was one of the shortcomings of this project. Finally, a topic that wasn't looked at during the project but could help identify the sentiment that users are feeling is semantic analysis, which is another technique that falls under the area of NLP. This could help in showing and understanding emotion, which can help make the chatterbot more human-like.

6.4 Summary

This chapter reviewed the project as a whole, including what went well and where improvements could be made. It first reviewed the aims and objectives defined at the start of the project and how the project helped achieve them. The author then personally evaluated his performance during the project and where there were shortcomings as well as successes. Finally, some recommendations were made for further work that could be carried out to take this project further.

References

Abdul-Kader, Sameera A and JC Woods (2015). "Survey on chatbot design techniques in speech conversation systems". In: *International Journal of Advanced Computer Science and Applications* 6.7.

Bakhshi, Naser et al. (2018). *Chatbots Point of View*. Tech. rep. Deloitte The Netherlands.

Bradeško, Luka and Dunja Mladenić (2012). "A survey of chatbot systems through a loebner prize competition". In: *Proceedings of Slovenian Language Technologies Society Eighth Conference of Language Technologies*, pp. 34–37.

Brandtzaeg, Petter Bae and Asbjørn Følstad (2017). "Why people use chatbots". In: *International Conference on Internet Science*. Springer, pp. 377–392.

Briscoe, Ted (n.d.). *Introduction to Linguistics for Natural Language Processing*. Available at `https://www.cl.cam.ac.uk/teaching/1314/L100/introling.pdf` (Accessed: 24/03/2019).

Brooker, Dawn (2003). "What is person-centred care in dementia?" In: *Reviews in Clinical Gerontology* 13.3, pp. 215–222. DOI: `10.1017/S095925980400108X`.

Choi, Jinho D, Joel Tetreault, and Amanda Stent (2015). "It depends: Dependency parser comparison using a web-based evaluation tool". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Vol. 1, pp. 387–396.

De Angeli, Antonella, Graham I Johnson, and Lynne Coventry (2001). "The unfriendly user: exploring social reactions to chatterbots". In: *Proceedings of The International Conference on Affective Human Factors Design, London*, pp. 467–474.

Hill, Jennifer, W Randolph Ford, and Ingrid G Farreras (2015). "Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations". In: *Computers in Human Behavior* 49, pp. 245–250.

Liddy, Elizabeth D (2001). "Natural language processing". In:

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2009). *Introduction to information retrieval*. Cambridge University Press.

Mcleod, Saul (2010). *Long-Term Memory — Simply Psychology*. URL: `https://www.simplypsychology.org/long-term-memory.html`.

Nadeau, David and Satoshi Sekine (2007). "A survey of named entity recognition and classification". In: *Lingvisticae Investigationes* 30.1, pp. 3–26.

Radziwill, Nicole M and Morgan C Benton (2017). "Evaluating quality of chatbots and intelligent conversational agents". In: *arXiv preprint arXiv:1704.04579*.

Shum, Heung-Yeung, Xiao-dong He, and Di Li (2018). "From Eliza to XiaoIce: challenges and opportunities with social chatbots". In: *Frontiers of Information Technology & Electronic Engineering* 19.1, pp. 10–26.

Stanford Named Entity Recognizer (NER) (n.d.). Available at `https://nlp.stanford.edu/software/CRF-NER.shtml` (Accessed: 17/10/2018).

Stanford NLP Group - Machine Translation (n.d.). Available at `https://nlp.stanford.edu/projects/mt.shtml` (Accessed: 24/03/2019).

Tulving, Endel et al. (1972). "Episodic and semantic memory". In: *Organization of memory* 1, pp. 381–403.

Turing, A. M. (1950). "I.—COMPUTING MACHINERY AND INTELLIGENCE". In: *Mind* LIX.236, pp. 433–460. DOI: `10.1093/mind/LIX.236.433`.

Weischedel, Ralph et al. (1989). "White Paper on Natural Language Processing". In: *Proceedings of the Workshop on Speech and Natural Language*. HLT '89. Cape Cod, Massachusetts: Association for Computational Linguistics, pp. 481–493. ISBN: 1-55860-112-0. DOI: `10.3115/1075434.1075526`.

Weizenbaum, Joseph (Jan. 1966). "ELIZA - a Computer Program for the Study of Natural Language Communication Between Man and Machine". In: *Commun. ACM* 9.1, pp. 36–45. ISSN: 0001-0782. DOI: `10.1145/365153.365168`.

Zdenek, Sean (2001). "Passing Loebner's Turing Test: A Case of Conflicting Discourse Functions". In: *Minds and Machines* 11.1, pp. 53–76. DOI: `10.1023/a:1011214808628`.

Bibliography

Bang, J. et al. (2015). "Example-based chat-oriented dialogue system with personalized long-term memory". In: *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*, pp. 238–243. DOI: 10.1109/35021BIGCOMP.2015.7072837.

Collobert, Ronan et al. (2011). "Natural Language Processing (almost) from Scratch". In: *Journal of Machine Learning Research* 12, pp. 2493–2537.

Coniam, David (Sept. 2014). "The linguistic accuracy of chatbots: Usability from an ESL perspective". In: *Text & Talk* 34. DOI: 10.1515/text-2014-0018.

Ferland, Libby et al. (2018). *Assistive AI for Coping with Memory Loss*. URL: https://www.aaai.org/ocs/index.php/WS/AAAIW18/paper/view/17360/15619.

Lucassen, Garm et al. (Mar. 2018). "Jobs-to-be-Done Oriented Requirements Engineering: A Method for Defining Job Stories". In: pp. 227–243. ISBN: 978-3-319-77242-4. DOI: 10.1007/978-3-319-77243-1_14.

Mitkov, Ruslan (2014). *Anaphora resolution*. Routledge.

Roy, Nicholas, Joelle Pineau, and Sebastian Thrun (2000). "Spoken Dialogue Management Using Probabilistic Reasoning". In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. ACL '00. Hong Kong, pp. 93–100. DOI: 10.3115/1075218.1075231.

Shah, A. et al. (2018). "Problem solving chatbot for data structures". In: *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 184–189. DOI: 10.1109/CCWC.2018.8301734.

Shawar, Bayan Abu and Eric Atwell (2007). "Chatbots: are they really useful?" In: *Ldv forum*. Vol. 22. 1, pp. 29–49.

Appendix A

Feasibility Study

This appendix attaches the feasibility study, which was carried out prior to the start of this project. The feasibility study document includes the ethics and research insurance forms.

Feasibility Study

ELIZA Plus - Long-Term Memory

by

PRITAM SANGANI

SUPERVISED BY: DR. JAMES O'SHEA

School of Computing, Mathematics and Digital Technology
MANCHESTER METROPOLITAN UNIVERSITY

A Feasibility Study document submitted to Manchester Metropolitan University, as part of a final year dissertation, in accordance with the requirements of the degree of BACHELOR OF SCIENCE in the Faculty of Science and Engineering.

OCTOBER 2018

Course-Specific Learning Outcomes

Below are the Course-Specific Learning Outcomes for BSc (Hons) in Computer Science. Please note that only those outcomes which are relevant to this dissertation have been listed.

On completion of the course students will be able to:

- use knowledge, abilities and skills for further study and for a range of employment in areas related to scientific and technical computing;
- interpret legislation appropriate to computer professionals and also be aware of relevant ethical issues and the role of professional bodies;
- analyse, design, and implement algorithms using a range of appropriate languages and/or methodologies;
- demonstrate effective communication, decision making and creative problem solving skills, and identify appropriate practices within a professional, legal and ethical framework;
- critically appraise and apply suitable artificial intelligence techniques for a variety of software systems.

Background

In 1950, Alan Turing published a paper called 'Computing Machinery and Intelligence' (Turing 1950), in which he posed the question, "Can machines think?". Alongside this question, he also proposed the Turing Test - a test that would use conversation to answer this question. Since then, Computer Scientists have been researching, for decades, how to develop chatterbots that converse convincingly with humans. Research, in this field, intensified when in the mid-1960s, Joseph Weizenbaum, developed ELIZA (Weizenbaum 1966) - known to be the first chatter bot. ELIZA used an early-form of Natural Language Processing, where it would match patterns in the text input and substitute it with phrases, to create the illusion of understanding. In recent years, chatterbot developers have been trying to win the Loebner Prize - a modern day version of the Turing Test. This contest has been held since 1991 in which judges converse with chatterbots not knowing whether they are talking to a human or a chatterbot (Zdenek 2001). Chatterbots have also become very popular commercially, as businesses look to make their customer services more efficient, innovative and, most importantly, more personal, due to the rising demand of customers looking for fast resolutions to their problems and the increase in time they are spending online (Bakhshi et al. 2018). This report by Deloitte, also found that one of the market forces driving chatterbot development is the "technological advances in AI and NLP".

Figure 1 below shows an infographic, also from the Deloitte report, displaying the "different functions of the human brain" that chatterbots try to "mimic". For my thesis, I will be concentrating on the Natural Language Processing (NLP) and Entity Recognition parts of the infographic for Dialog Management. NLP is a branch of AI concerned with the research of interactions between humans and computers through natural language. Natural Language has been defined in a white paper on NLP as being "the most natural means of communication between humans, and the mode of expression of choice for most of the documents they produce" (Weischedel et al. 1989). Entity Recognition (also known as Named Entity Recognition[NER]) is a branch of NLP concerned with labelling "sequences of words in a text which are the names of things, such as person and company names, or gene and protein names." (*Stanford Named Entity Recognizer (NER)* n.d.)

A big part of developing chatterbots is actually evaluating the quality of existing chatterbots and

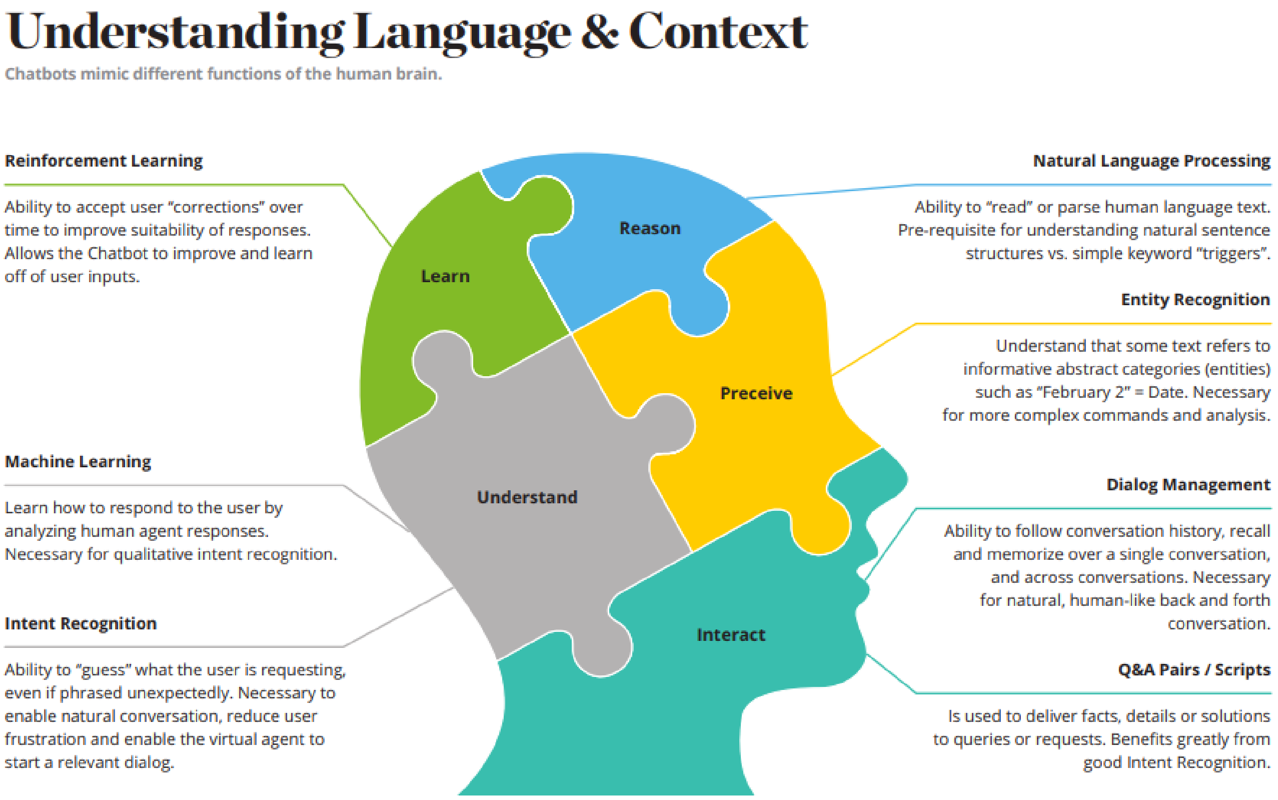


Figure 1: Infographic showing the different functions of a chatterbot architecture (Bakhshi et al. 2018)

how convincing they are at mimicking the functions of the human brain. Two recent studies have tried to answer two questions that must be answered when evaluating the quality of a chatterbot. Firstly, the current uses of chatterbots in society must be examined. Brandtzaeg and Følstad have published their study of the uses of chatterbots on various platforms and across various categories of uses. (Brandtzaeg & Følstad 2017) The vast majority of participants in the study reported using chatterbots to increase their productivity, by quickly retrieving information or accessing assistance. Rather surprisingly, the study also found that 12% of the participants reported using chatterbots for social or relational use. The study found that the human nature of chatterbots drove people to using them for this purpose - the responses stated that chatterbots were a way of "avoiding loneliness" and "improving their social and conversational skills". Another study tried to gather together all the attributes and features that can be used to assess the quality of chatbots. (Radziwill & Benton 2017) The study found that two attributes that measures the effectiveness of the chatterbot are the ability to "maintain themed discussion" and to deliver "convincing, satisfying and natural interaction". The study also found that one of the attributes that measures the satisfaction of the chatterbot is the ability to "detect meaning or intent".

A big part of this project is to add a "long-term memory" mechanism to the core of the ELIZA chatterbot. Long-term memory is the ability to refer back to earlier conversations and to bring the information back at relevant points in the current conversation. This would make the chatterbot more human-like and, as discussed, would improve the quality of the chatterbot. I plan to achieve this using NER and NLP to store relevant and linked information into a database to retrieve at a later point in the conversation.

Research Question

The question I propose to answer with my research is as follows:

Is it possible to implement convincing long-term memory into an existing chatterbot, such as Weizenbaum's ELIZA?

Hypotheses

From the research question above, I propose that the following hypotheses could be derived from my research:

- **H_0 - it is not possible to add long-term memory to a chatterbot, at this stage, which is convincing enough to be comparable to human memory mechanism.**
- **H_1 - it is possible to add convincing long-term memory at a significant statistical level, if the number of participants in the research is high enough.**

Note: The degree of convincingness would be determined by the analysis of conversation log files, to see how often the memory mechanism appeared to access what the user was referring to, and by a subjective analysis of a questionnaire put to the participants of a research experiment.

Aim

The aims of this thesis are as follows:

- **Integrate Long-Term Memory mechanism to the ELIZA chatterbot core.**
- **Evaluate how convincing the long-term memory mechanism is, by way of research methods determined by research done in literature review.**

Objectives

The problem is a complex one to be solved and so I have broken the problem down into small steps and approaches that I will take to solve the problem. These are as follows:

1. Complete a Literature Review by reading research papers and articles on the topic of Natural Language Processing and Named Entity Recognition from the Internet and textbooks.
2. Find source code for the ELIZA chatterbot in Python and start understanding how the code works.
3. Compare the different tools for Named Entity Recognition readily available for use in Python.
4. Design an overview of the software and a plan to implement.
5. Implement the software according to the plan, whilst incrementally testing where appropriate.
6. Carry out a final testing of the software.
7. Evaluate the quality of the software by carrying out a study with participants.
8. Write up the research findings in a report.

As well as the objectives listed above there are a number of interim deliverables that need to be met. These are as follows:

1. Prototype Report
2. Prototype Software
3. Final Software
4. Report Outline
5. Showcase Event
6. Final Report

Problems

As with any project, problems can arise and it is important to resolve them quickly. To mitigate the effect of problems during this project, I am considering the problems that could arise and how I would overcome each problem.

Firstly, a problem that is likely to occur is in the implementation of the product. As I have not worked with Natural Language Processing before, I am likely going to have some trouble, initially, implementing the Named Entity Recognition algorithms. To resolve any issues in implementation, I will first look at the documentation for any modules I use. If this does not solve the problem, I can use sites, such as Stack Overflow, to search for possible solutions.

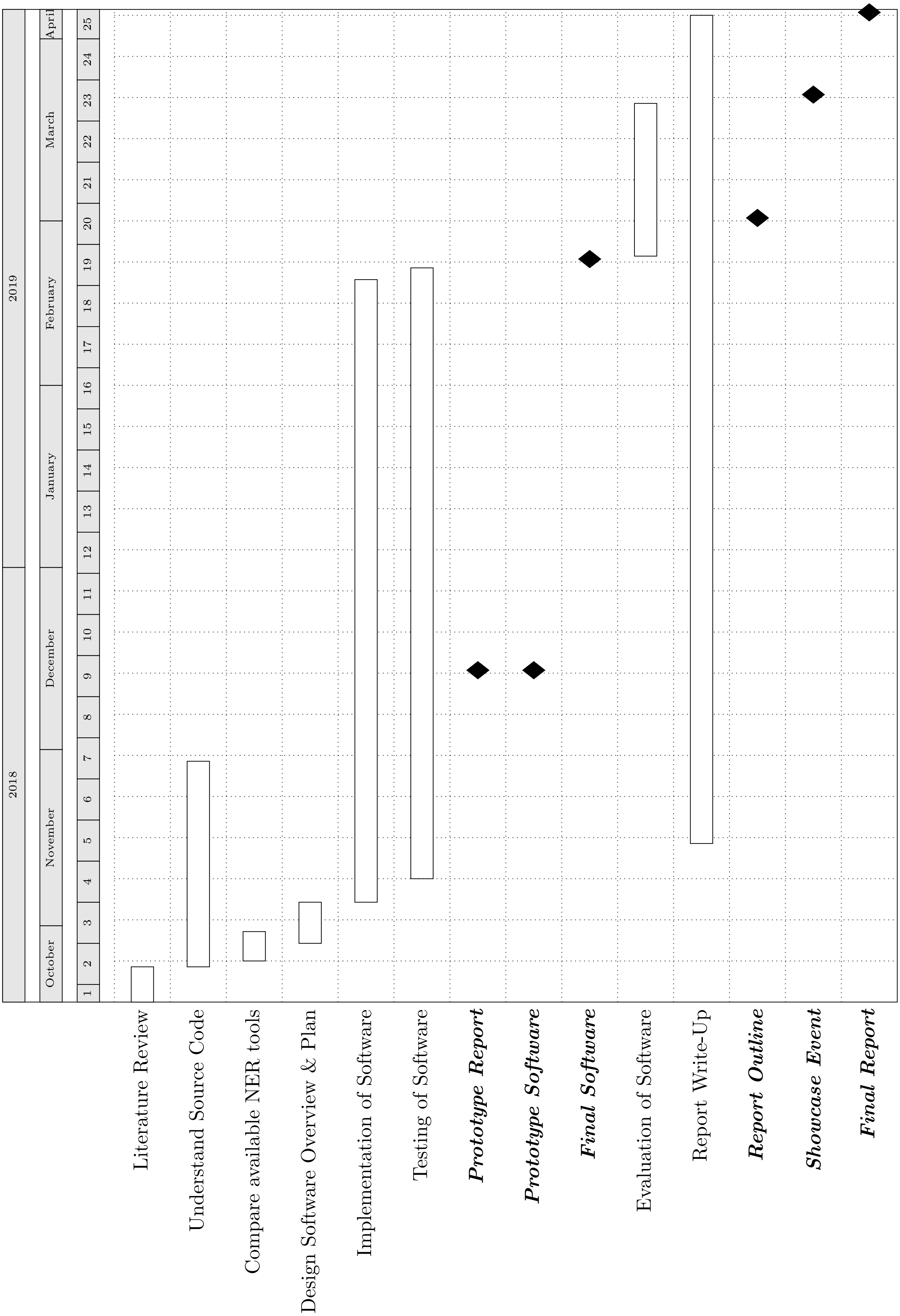
A second problem, that is not likely to occur but could still occur, is that my code could be lost, either by my laptop becoming corrupt or stop working. To mitigate this problem, I will use a version control system called Git and I will host my code on GitHub, which is a cloud-based storage service for code repositories.

Evaluation

As with all projects of this nature, an evaluation of the produced software must be undertaken. I intend to do this by running an experiment where participants would converse with the chatterbot for two sessions. There will be two sessions as this project involves implementing long-term memory. Therefore, in order for the participant to properly evaluate the quality of the chatterbot, in particular its ability to retrieve information from previous conversations, I would need the participant to have two sessions with the chatterbot. Once the participant has concluded the experiment, I would have them answer a questionnaire to gauge their responses and feelings about the chatterbot's quality. I would then subjectively analyse the responses, as well as, look at the conversation logs to determine the degree of convincingness of the chatterbot.

Schedule

Below is the proposed week-by-week schedule for the project as outlined by the tasks/objectives outlined in the Objectives section. I have displayed the objectives as bars, whilst the deliverables are classed as milestones, which are displayed using diamond symbols.


2018
2019
October
November
December
January
February
March
April
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
Literature Review
Understand Source Code
Compare available NER tools
Design Software Overview & Plan
Implementation of Software
Testing of Software
Prototype Report
Prototype Software
Final Software
Evaluation of Software
Report Write-Up
Report Outline
Showcase Event
Final Report

Required Resources

The resources required to carry out this research project is as follows:

- A laptop or PC capable of smoothly running:
  - Python 3.x
  - Google Chrome
  - PyCharm - an IDE, by JetBrains, to write and compile Python code
  - Visual Studio Code - an IDE to write code in any programming language (HTML, CSS and JavaScript for this project)

All resources listed above are all available to download for free. (Note: PyCharm has a paid version, but, for the scope of this project, the free community edition will suffice.) Furthermore, I will not be relying on the University Laboratory PCs as I can run all required resources on my personal laptop.

References

Bakhshi, N., van den Berg, H., Broersen, S., de Vries, D., El Bouazzaoui, H. & Michels, B. (2018), Chatbots point of view, Technical report, Deloitte The Netherlands.

Brandtzaeg, P. B. & Følstad, A. (2017), Why people use chatbots, *in* 'International Conference on Internet Science', Springer, pp. 377–392.

Radziwill, N. M. & Benton, M. C. (2017), 'Evaluating quality of chatbots and intelligent conversational agents', *arXiv preprint arXiv:1704.04579* .

Stanford Named Entity Recognizer (NER) (n.d.), Available at `https://nlp.stanford.edu/software/CRF-NER.shtml` (Accessed: 17/10/2018).

Turing, A. M. (1950), 'I.—computing machinery and intelligence', *Mind* **LIX**(236), 433–460.

Weischedel, R., Carbonell, J., Grosz, B., Lehnert, W., Marcus, M., Perrault, R. & Wilensky, R. (1989), White paper on natural language processing, *in* 'Proceedings of the Workshop on Speech and Natural Language', HLT '89, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 481–493.

Weizenbaum, J. (1966), 'Eliza - a computer program for the study of natural language communication between man and machine', *Commun. ACM* **9**(1), 36–45.

Zdenek, S. (2001), 'Passing loebner's turing test: A case of conflicting discourse functions', *Minds and Machines* **11**(1), 53–76.

Manchester Metropolitan University

Undergraduate and PGT Application

START HERE - Basic Information

This form must be completed for all student projects.

Before you proceed

Some activities inherently involve increased risks or approval by external regulatory bodies, so a proportional ethics review is not recommended and a full ethical review may be required.

These may include:

i. Approval from an external regulatory body (including, but not limited to: NHS (HRA), HMPPS etc.);
ii. Misleading participants;
iii. Research without the participants' consent;
iv. Clinical procedures with participants;
v. The ingestion or administration of any substance to participants by any means of delivery;
vi. The use of novel techniques, even where apparently non-invasive, whose safety may be open to question;
vii. The use of ionising radiation or exposure to radioactive materials;
viii. Engaging in, witnessing, or monitoring criminal activity;
ix. Engaging with, or accessing terrorism related materials;
x. A requirement for security clearance to access participants, data or materials;
xi. Physical or psychological risk to the participants or researcher;
xii. The project activity takes place in a country outside of the UK for which there is currently an active travel warning issued by the authorities (see info button);
xiii. Animals, animal tissue, new or existing human tissue, or biological toxins and agents.

If any of these activities are fundamental to your project, please contact your supervisor to determine if a full application is required.

This form must be completed for each research project which you undertake at the University. It must be approved by your supervisor (where relevant) PRIOR to the start of any data collection.

In completing this form, please consult the University's ACADEMIC ETHICAL FRAMEWORK for ethical research.

A1 Please confirm that you will abide by the University's Academic Ethical Framework in relation to this project.

- (●) Yes
- ( ) No

A2 Are you submitting this application as a learning experience, for a unit which already has ethical approval? (please confirm with your supervisor)

- ( ) Yes
- (●) No

A3 Student details

Title	First Name	Surname
	Pritam	Sangani

Email: pritam.sangani@stu.mmu.ac.uk

A4 Supervisor

Title	First Name	Surname
Dr	James	O'Shea

Faculty: Science and Engineering

Telephone: 0161 247 1546

Email: j.d.oshea@mmu.ac.uk

A5 Which Faculty is responsible for the project?

Science and Engineering

A6 Course title

6G6Z1101 - Project

A7 Project title

Eliza - Long Term Memory

A8 What is the proposed start date of your project?

19/10/2018

A9 When do you expect to complete your project?

23/08/2019

A10 Please describe the overall aims of your project (3-4 sentences). Research questions should also be included here.

This project has two aims. The first aim is to "Integrate Long-Term Memory mechanism into the ELIZA chatterbot core." The second aim is to "Evaluate how convincing the long-term memory mechanism is, by way of research methods determined by research done in the literature review." I also have a research question that I aim to answer. This is "Is it possible to implement convincing long-term memory into an existing chatterbot, such as Weizenbaum's ELIZA?"

A11 Please describe the research activity

I will start my project by completing a short literature review by reading research papers and articles on the topic of Natural Language Processing and Named Entity Recognition, which are the topics my project is based upon. I will obtain these from the internet and textbooks. At the same time, I will also find open-source code for the ELIZA chatterbot and start looking through the code to gain an understanding of how the software works. I will obtain the code from GitHub, a cloud-based storage service for code repositories. I will then research and compare the different tools available for implementing Named Entity Recognition. Once this is done, I will begin planning for the implementation stage, by designing an overview of the software and creating a plan of features that need to be implemented, thus breaking down the implementation into small tasks. During the implementation stage, I have decided to incrementally test the software, where appropriate, by conversing with the chatterbot - I will personally be doing the testing. Once the implementation stage is complete, I will be carrying out a final testing of the software to make sure it is ready to be evaluated by participants - again I will personally be doing the testing. Once testing is complete, I will be ready to ask participants to evaluate the quality of the chatterbot by having them converse with the chatterbot. I will be keeping a log of the conversation history for each participant, which I will analyse once the study is complete. I will also ask the participants to fill out a questionnaire to get their responses and feelings about the quality of the chatterbot. I will subjectively analyse the questionnaires and then write up my findings in a report.

A12 Please provide details of the participants you intend to involve (please include information relating to the number involved and their demographics; the inclusion and exclusion criteria)

I will try and involve a minimum of 5 participants to a maximum of 15. The participants will all be students and I will try and have a mixed demographic of participants. (a mix of gender and race)

Project Activity

B1 Are there any Health and Safety risks to the researcher and/or participants?

- ○ Yes
- ◉ No

B2 Please select any of the following which apply to your project

- ☑ Aspects involving human participants (including, but not limited to interviews, questionnaires, images, artefacts and social media data)
- ☐ Aspects that the researcher or participants could find embarrassing or emotionally upsetting
- ☐ Aspects that include culturally sensitive issues (e.g. age, gender, ethnicity etc.)
- ☐ Aspects involving vulnerable groups (e.g. prisoners, pregnant women, children, elderly or disabled people, people experiencing mental health problems, victims of crime etc.), but does not require special approval from external bodies (NHS, security clearance, etc.)
- ☐ Project activity which will take place in a country outside of the UK
- ☐ None of the above

B2.4 Is this project being undertaken as part of a larger research study for which a Manchester Metropolitan application for ethical approval has already been granted or submitted?

- ○ Yes
- ◉ No

Informed Participation/Consent

C1 Will participants be given accessible information about:

a) the general purpose of the project

b) what is expected from them in the project

c) their right to refuse or withdraw at any time

d) how their data will be used and managed, and their relevant legal rights

- (x) Yes
- ( ) No

C1.2 Please describe how you will do this

I will give the participants all information regarding the project and what is expected of them in written form (probably a leaflet). I will also outline how they can withdraw from the project at any time, as well as how their conversation logs will be used and kept strictly private.

C2 Will you ask for informed consent from all participants?

- (x) Yes
- ( ) No

C3 Will any participants be legally unable to consent, and require you to obtain informed consent from a legal representative?

- ( ) Yes
- (x) No
- ( ) Not Applicable

C3.1 Please give details why this will not be necessary or relevant

Study will only involve day-to-day conversation and will not involve any information which is confidential of requires any legal consent from authorities.

C4 Will participants have an opportunity to ask questions prior to agreeing to participate?

- (x) Yes
- ( ) No

C4.1 Please describe how participants will be able to ask questions prior to agreeing to participant

I will provide possible participants with my student email which they can use if they have any questions to ask.

C5 Will participants receive any payments, reimbursements of expenses or any other benefits or expenses for taking part in this project?

- ( ) Yes
- (x) No

C6 Have relevant authorities (gatekeepers) given their permission for project activities to take place on their premises (e.g. shop managers, service managers, head teachers, classroom lecturers)?

- ( ) Yes
- ( ) No
- (x) Not Applicable

C7 Could your past or present relationship with the potential participants give rise to a perceived pressure to participate?

- ○ Yes
- ◉ No

C9 Will any participants be identified through posters, leaflets, adverts, social media or websites?

- ○ Yes
- ◉ No

Anonymity and Confidentiality

D1 Please describe how you will protect participants anonymity

> I will remove all identifying data, such as name, from the conversation logs and also I will inform the participants that they should not enter any personal information when conversing with the chatterbot.

D2 Please describe how you will ensure that individuals cannot be identified indirectly (e.g. via other information that is collected)?

> No personal details will be collected before, during or after the study

D3 Please describe how you will protect participants confidentiality?

> Conversation logs will be cleared of any identifying details before being written up in the report.

Debriefing

E1 Will participants have the opportunity to obtain feedback or the results after the project has ended?

- ○ Yes
- ◉ No

Data

F1 How and where will data and documentation be stored?

> Data will be collected in the form of conversation logs and will be kept on university servers which are protected by security that the university employs

F2 Will you be collecting personal data or sensitive personal data as part of this project?

- ○ Yes
- ◉ No

Additional Information

G1 Do you have any additional information or comments which have not been covered in this form?

- ○ Yes
- ◉ No

G2 Do you have any additional documentation which you want to upload?

- ◉ Yes
- ○ No

G2.1 Please attach a copy of any other materials relevant to this application

Type	Name	File Name	Date	Version	Size
Additional Documentation	Risk Assessment	RA_Project_SoftwareDevelopment_270918.pdf	18/10/2018 12:00:00 AM	1	369.0 KB

Signatures

H1 I confirm that all information in this application is accurate and true. I will not start this project until I have received Ethical Approval.

- ◉ I confirm
- ○ I do not confirm

H2 Please notify your supervisor that this application is complete and ready to be submitted by clicking "Request" below. Do not begin your project until you have received confirmation from your supervisor - it is your responsibility to ensure that they do this.

Signed: This form was signed by Jim O'Shea (J.D.Oshea@mmu.ac.uk) on 19/10/2018 11:04

H3 By signing this application you are confirming that all details included in the form have been completed accurately and truthfully.

Signed: This form was signed by Pritam Sangani (pritam.sangani@stu.mmu.ac.uk) on 18/10/2018 15:02

Research Insurance Checklist

ADMINISTRATIVE DETAILS

Lead Investigator Name (Title/Forename/Surname)	**Pritam Sangani**
Contact Email Address	**pritam.sangani@stu.mmu.ac.uk**
Full Title of the Research	**Eliza – Long Term Memory**

SECTION 1 – TECHNIQUES, TESTING AND INTERVENTIONS

Does your research study involve:

☐ **Physically invasive techniques?**

This refers to any test in which the skin of the participant is broken or an implement is inserted into any opening of the human body (e.g. eyes, ears, nose, mouth, lungs, stomach, rectum, vagina and urethra) or involves the taking of body samples such as saliva, hair, urine, faeces, sputum, skin, nails, or taking biopsies of any form for any purpose, or any form of scanning such as DEXA scans, Ultrasound scans, MRI, fMRI, CT, or PET scanning.

☐ **Ingestion of food stuffs or drugs?**

This refers to the consumption of any substance which may impact on psychological or physical state. Substances may include but are not limited to food, beverages or drugs.

☐ **Physical testing?**

This refers to any test in which a participant must perform an action resulting in the use of any muscle of the body and/or involves the use of scanning procedures, eye-trackers, mounted body cameras, sensors or electrodes, or the taking of swabs from any cavity of the body, respiratory challenge testing or recording of peak flows, EEG, ECG, Exercise ECG, Treadmill work.

☐ **Psychological intervention?**

This refers to any test which purposely alters the mood of the participant or involves administering personality inventories, or any other form of psychological test.

OR

✓ **I confirm that my research does not fall into any of the above categories *(please go straight to Section 3)***

Manchester Metropolitan University

SECTION 2 – CLINICAL TRIALS INSURANCE

Please complete this section only if you ticked one of the boxes in Section 1.

Does your research study involve:

☐ **Pregnant persons as participants with procedures other than blood samples being taken from them?**

☐ **Children aged five or under with procedures other than blood samples being taken from them?**

☐ **Activities being undertaken by the lead investigator or any other member of the study team in a country outside of the UK?** ***If 'Yes', please refer to the 'Travel Insurance' guidance on Page 1 of this form.***

OR

☐ **I confirm that my research does not fall into any of the above categories**

SECTION 3 – OTHER HAZARDS

Does your research study involve:

☐ **Working with Hepatitis, Human T-Cell Lymphotropic Virus Type iii (HTLV iii), or Lymphadenopathy Associated Virus (LAV) or the mutants, derivatives or variations thereof or Acquired Immune Deficiency Syndrome (AIDS) or any syndrome or condition of a similar kind?**

☐ **Working with Transmissible Spongiform Encephalopathy (TSE), Creutzfeldt-Jakob Disease (CJD), variant Creutzfeldt-Jakob Disease (vCJD) or new variant Creutzfeldt-Jakob Disease (nvCJD)?**

☐ **Working in hazardous areas or high risk countries?** ***Please refer to the 'High Risk Countries' guidance on Page 1 of this form.***

☐ **Working with hazardous substances outside of a controlled environment?**

☐ **Working with persons with a history of violence, substance abuse or a criminal record?**

OR

✓ **I confirm that my research does not fall into any of the above categories**

Appendix B
Showcase Materials

The slide deck used to present the project at the showcase event, held on 4th of April 2019, can be found below:

ELIZA Plus

Implementing a long-term memory mechanism in a chatterbot

by Pritam Sangani

Contents

1. Problem
2. Project Task
3. Implementation
4. Evaluation
5. Conclusions
6. Demo

Problem

- Turing Test / Imitation Game
- ELIZA chatterbot

Project Task

“

Implement and integrate a long-term memory mechanism in the ELIZA chatterbot that can refer to past conversations in the current context.

”

Implementation

- Web Interface
- Chatterbot
- Long-Term Memory

Evaluation

- Scenario-based
- Usability Evaluation
- Functionality Evaluation

PERSON

You are a young adult talking about a special person that you are trying to impress. You talk to ELIZA about that person, mentioning them by name.

PRODUCT

You have just bought an expensive product from Amazon. You talk to ELIZA about how excited you are for the product to arrive and start using it, mentioning the product by name during your conversation.

GPE/PLACE

You are going on holiday with your friends this Summer after being recommended the destination by some other friends. Talk to ELIZA about how excited you are to be travelling on holiday and the great things you have already heard, mentioning the place you are going to by name.

Conclusions

- Named Entity Recognition can extract key information
    - Limited by accuracy of model
- Pattern-matching and substitution is too primitive for the complex nature of conversation flow

Appendix C

Product Submission

The product supporting this project was submitted on the 22nd of February 2019. The link to the source code is:
`https://stummuac-my.sharepoint.com/:u:/g/personal/16039231_stu_mmu_ac_uk/ES_ubr59LnVIizGlB17pY9oBnrsORInKanzs_mhgUhRQ5g`

Appendix D

Evaluation Scenarios

The following three figures show the three scenarios written as part of the evaluation study of the product implemented to test the hypotheses set for this project.

PERSON

You are a young adult talking about a special person that you are trying to impress. You talk to ELIZA about that person, mentioning them by name.

Figure D.1: Scenario 1 - tasked to talk about a special person.

PRODUCT

You have just bought an expensive product from Amazon. You talk to ELIZA about how excited you are for the product to arrive and start using it, mentioning the product by name during your conversation.

Figure D.2: Scenario 2 - tasked to talk about an expensive new product that the participant has just bought.

GPE/PLACE

You are going on holiday with your friends this Summer after being recommended the destination by some other friends. Talk to ELIZA about how excited you are to be travelling on holiday and the great things you have already heard, mentioning the place you are going to by name.

Figure D.3: Scenario 3 - tasked to talk about a holiday the participant is going on with friends in the Summer.

Appendix E

Questionnaire

E.1 Questions in Questionnaire

The table below lists the questions asked as part of the evaluation questionnaire and the mode of answer for each question.

Question Number	Question Title	Mode of Answer
1	I found the user interface easy to navigate	rating: 1-5 (1=Strongly Disagree, 5=Strongly Agree)
2	The web application did not break unexpectedly	True/False
3	The chatterbot always responded when I sent it a message	True/False
4	I found the chatterbot to be welcoming	rating: 1-5 (1=Strongly Disagree, 5=Strongly Agree)
5	I found the chatterbot to be entertaining and engaging	rating: 1-5 (1=Strongly Disagree, 5=Strongly Agree)
6	I found that the chatterbot was able to pick up a conversation that I talked about in the first session, during the second session	rating: 1-5 (1=Strongly Disagree, 5=Strongly Agree)
7	The chatterbot could stay on theme for a reasonable amount of time	rating: 1-5 (1=Strongly Disagree, 5=Strongly Agree)
8	How likely would you be to talk to ELIZA Plus again?	rating: 1-5 (1=Very Unlikely, 5=Very Likely)
9	Enter any other feedback that you have regarding the user interface or the chatterbot below	Free Text (optional)

Table E.1: Table of questions asked in Evaluation Questionnaire

E.2 Full Responses

Participant 1		
Question Number	**Question Title**	**Answer**
1	I found the user interface easy to navigate	5.0/5.0
2	The web application did not break unexpectedly	True
3	The chatterbot always responded when I sent it a message	True
4	I found the chatterbot to be welcoming	4.0/5.0
5	I found the chatterbot to be entertaining and engaging	3.0/5.0
6	I found that the chatterbot was able to pick up a conversation that I talked about in the first session, during the second session	4.0/5.0
7	The chatterbot could stay on theme for a reasonable amount of time	2.0/5.0
8	How likely would you be to talk to ELIZA Plus again?	3.0/5.0
9	Enter any other feedback that you have regarding the user interface or the chatterbot below	The chatterbot picked up conversation but didn't stick to the same topic for as long as I would like. Also the chatterbot didn't ask about topics that I had talked about in the first session from the beginning of the second session - it took quite long for this to happen.

Table E.2: Table of Full Responses for Participant 1

Participant 2		
Question Number	**Question Title**	**Answer**
1	I found the user interface easy to navigate	5.0/5.0
2	The web application did not break unexpectedly	True
3	The chatterbot always responded when I sent it a message	True
4	I found the chatterbot to be welcoming	4.0/5.0
5	I found the chatterbot to be entertaining and engaging	4.0/5.0
6	I found that the chatterbot was able to pick up a conversation that I talked about in the first session, during the second session	4.0/5.0
7	The chatterbot could stay on theme for a reasonable amount of time	1.0/5.0
8	How likely would you be to talk to ELIZA Plus again?	4.0/5.0
9	Enter any other feedback that you have regarding the user interface or the chatterbot below	It was quite amusing to see the chatterbot getting stuck with some responses, but I felt that it was picking up conversation that I had talked about in the first session. The only bad thing is that it sometimes tried to get me to talk about something else quite quickly after moving on to a new topic.

Table E.3: Table of Full Responses for Participant 2

Participant 3		
Question Number	**Question Title**	**Answer**
1	I found the user interface easy to navigate	5.0/5.0
2	The web application did not break unexpectedly	True
3	The chatterbot always responded when I sent it a message	True
4	I found the chatterbot to be welcoming	2.0/5.0
5	I found the chatterbot to be entertaining and engaging	5.0/5.0
6	I found that the chatterbot was able to pick up a conversation that I talked about in the first session, during the second session	3.0/5.0
7	The chatterbot could stay on theme for a reasonable amount of time	1.0/5.0
8	How likely would you be to talk to ELIZA Plus again?	5.0/5.0
9	Enter any other feedback that you have regarding the user interface or the chatterbot below	The chatterbot moved on to a new topic way too quickly

Table E.4: Table of Full Responses for Participant 3

Participant 4		
Question Number	**Question Title**	**Answer**
1	I found the user interface easy to navigate	5.0/5.0
2	The web application did not break unexpectedly	True
3	The chatterbot always responded when I sent it a message	True
4	I found the chatterbot to be welcoming	5.0/5.0
5	I found the chatterbot to be entertaining and engaging	5.0/5.0
6	I found that the chatterbot was able to pick up a conversation that I talked about in the first session, during the second session	3.0/5.0
7	The chatterbot could stay on theme for a reasonable amount of time	3.0/5.0
8	How likely would you be to talk to ELIZA Plus again?	5.0/5.0
9	Enter any other feedback that you have regarding the user interface or the chatterbot below	It's really funny to see the chatterbot get stuck with some responses.

Table E.5: Table of Full Responses for Participant 4

Participant 5		
Question Number	**Question Title**	**Answer**
1	I found the user interface easy to navigate	5.0/5.0
2	The web application did not break unexpectedly	True
3	The chatterbot always responded when I sent it a message	True
4	I found the chatterbot to be welcoming	4.0/5.0
5	I found the chatterbot to be entertaining and engaging	5.0/5.0
6	I found that the chatterbot was able to pick up a conversation that I talked about in the first session, during the second session	4.0/5.0
7	The chatterbot could stay on theme for a reasonable amount of time	2.0/5.0
8	How likely would you be to talk to ELIZA Plus again?	2.0/5.0
9	Enter any other feedback that you have regarding the user interface or the chatterbot below	It's a cool chatbot - only bad thing is that the chatterbot was moving on to a new topic quite quickly and I kept on getting same responses frequently.

Table E.6: Table of Full Responses for Participant 5