University of Essex – Computer Science

Initial Report

Open Domain Question & Answer System

Osama Rahman – 1304349

Supervisor Name

Secondary Supervisor Name

Table of Contents

Introduction

In modern computing it seems that the next wave of consumer and commercial innovation will be in artificial intelligence. The field, while not new, has recently spiked in popularity. With companies such as Google and Apple competing head to head with their *personal assistants*, and IBM researching and developing Watson, it is an exciting time to be in the field of artificial intelligence.

Open domain question & answer systems are a form of artificial intelligence which, when given a question, will return an answer. Currently, no system has a truly open domain, artificial intelligence isn't quite there yet. However, many systems such as Siri, Google Now, and IBM's Watson, have a substantial domain and will be able to answer a considerable amount of questions. Usually these questions will be trivial such as "What is the weather like today?", or "Who is the current prime minister of the United Kingdom?" Open ended questions such as "Why did Brexit happen?" are not likely to be answered as the answer would have to be meaningful, and would rely on more than textual information to be formulated.

In this project I intend to create an open domain question and answer system, however with more of a niche. Alongside this I intend to implement basic general questions one would find themselves asking Siri or Google now, questions such as "What is the time?", or perhaps simple maths such as "What is six multiplied by 2?". If possible I will be implementing an open source voice API so the user can speak to the system, as opposed to communicating the question via text. This will be discussed in further details below.

Project Goals

Core Objectives:

- Create a user friendly GUI to input and output text and data
    - It is important to create a user interface so the end user feels comfortable using the application. I will be using websites such as Dribbble to gain inspiration while also basing the design on systems such as Siri and Google Now

- Parse natural language user input into a readable query
    - This is needed for the system to retrieve information from the knowledgebase/ database. A user will ask a question, this will then be translated using Natural Language Processing into a computer readable query to retrieve the answer. I will be using an open source natural language processing module.

- Use the query to retrieve data from a database
    - Once translated into a query, from natural language, the system will use this query retrieve the relevant data from a knowledgebase/database.

- Parse the information from the database into a natural language reply
    - Once the data has been retrieved it will be slightly modified to make the systems answer more human like. The answer will be similar to the reply one would expect from an actual human.

Secondary Objectives (To be completed if possible / have time):

- Implement a voice API to allow a user to ask a question through speech
- Implement a large data dump, to make give the system a more open domain
- Implement a way to answer basic maths questions
- Implement a weather API for the system to report back weather
- Implement a world clock API for the system to report global times

Personal Objectives:

- Enhance my programming knowledge and skills
- Further educate myself in the field of artificial intelligence
- Enhance my project management skills and knowledge
- Enhance my professional skills

To summarise the project goals, the core objectives are the ones that will be implemented. Combined, these objectives will allow for a relatively open domain question and answer system. The secondary objectives will be implemented if it is possible, and I am able to do so within the deadline. I have also outlined some personal objectives, seeing as how this is my final year project and I will not only be showing what I have learnt over the years, but I will also be learning throughout the project.

Development

I will be using some open source software, mostly for non-core modules, however I will also be using open source software for core modules, such as the natural language processing. My main development will be with the data retrieval.

In terms of programming languages, I will be predominantly using python.

Natural Language Processing

This module will allow the user to interact with the system in the same way a human would interact with another human, through natural language. I have decided to use an open source project for this. I have not decided which one I will be using yet, as I need to trial each one to see which would suite the needs of the project.

Adapt Intent Parser by Mycroft

https://adapt.mycroft.ai/

This parser is an open source software library used to convert natural language into machine readable data structures.

Natural Language Tool Kit (NLTK 3.0)

http://www.nltk.org/

This toolkit is specifically designed for Python programs. Out of all of the Natural Language tools available, I think this will be the best one for me to use. NTLK features 104 different corpora and lexical resources to translate natural language into machine readable data.

Databases / Data dumps

This is where the information will be kept, where the system will go to find an answer for a question. There are many open source databases / data dumps available, which are great for open domain question and answer systems. Ones I am considering to use are below. The main issue I will run into with this is the sheer size of the database/ data dump. They can range from a few GB to a few TB. What this means is that the larger the size of the data, the longer it will take the system to find the answer.

DBpedia

DBpedia uses Wikipedia as its source of information. DBpedia is more of a knowledgebase than a database. This means that, while structured like a database, it is not a limited concrete technical solution. It will allow for more manipulation and accessibility of data, as a oppose to an immediate retrieval of information.

Voice API's

I plan on adding voice functionality to the system. Creating one from scratch will prove to be too much work. My time would be better spent on the core functionality of the program.

SpeechRecognition 3.4.6 https://pypi.python.org/pypi/SpeechRecognition/

This is a Python library for speech recognition. With support for many SpeechRecognition API's such as CMUSphinx, IBM Speech to Text, and many more this library should prove to be very resourceful.

Information Retrieval

Exactly how this is going to work, I am not sure yet. From my research I have an understanding on the theory and base principles of retrieving the data. In practice, however, I'm sure my development of this module will be slightly different. In essence all that will need to be done is to send a query to the knowledgebase/database, then have the system return the data in the form of a natural language answer. I don't plan on using any open source projects for this section of the program, so I will be writing it entirely from scratch.

Planning and Project Management

Waterfall Methodology

I have decided to use the traditional waterfall project planning methodology. Currently I am on the system design part of the waterfall method. I have chosen this methodology as I appreciate the classic, straight forward approach to software development.

Considering I will be doing this project on my own, it will be easier to keep to the scheduled sequence of events. The requirements of my project are very clear, the system I am building has been built before, and I have read a lot of material on this subject matter so I know exactly how the end project should work. As with any software development project, no matter the methodology, things are likely to change. For example, some secondary goal modules may not end up being developed, or perhaps the way in which a module is required to be developed will change.

Gantt Chart

Using a Gantt chart will allow me to manage the project and keep a timely schedule. This, along with my logbook, will also help me create the final report. What's great about a Gantt chart is that it allows for deadlines to be met in a timely and much easier manner.

Research

Factoid vs Complex Questions

With question and answer systems the types of questions can be split into two main categories, factoid and complex. As the name suggests, factoid questions are answered with short fact based textual answers. Factoid questions are simple; they are questions such as “Where is Apple Computer based?”[1]. Complex questions fi

[1] Klabunde, (2014)

References

Klabunde, R. (2014) ‘Daniel Jurafsky/James H. Martin, speech and language processing’, Zeitschrift für Sprachwissenschaft, 21(1). doi: 10.1515/zfsw.2002.21.1.134.