Deep Neural Network For Answer Extraction

ADAPTIVE ARTIFICIAL INTELLIGENT QUESTION ANSWER SYSTEM

17-107

Preliminary Progress Review

(Preliminary Progress Review Documentation submitted in partial fulfilment of the requirement for the Degree of Bachelor of Science Special (Honours) In Information Technology)

Singhabahu C.P. (IT14126802)

Bachelor of Science (Honours) in Information Technology
(Specialization in Software Engineering)

Department of Information Technology

Sri Lanka Institute of Information Technology

May 2017

1.0 TABLE OF CONTENT

LIST OF FIGURES

2.0 INTRODUCTION

This section focuses on giving an introduction to the Preliminary Progress Review (PPR) document and its content which will include the purpose of this document, acronyms and abbreviations used and an overview.

2.1 PURPOSE

The purpose of this document is to provide the preliminary progress of the "Deep Neural Network For Answer Extraction" component of "Adaptive Artificial Intelligent QA Platform". The document will illustrate the purpose and the main areas to be focussed throughout the component. Also it will explain on research question, objective and methodology, Further indicates the sources for test data, anticipated benefits, expected outcome and constraints. This document is primarily intended to be proposed to the audience who has already referred to the proposal of "Adaptive Artificial Intelligent QA Platform" and also to anyone who has a knowledge on deep learning. Further this document will be used as a reference for the progress of the component.

2.2 ACRONYMS, ABBREVIATIONS AND DEFINITIONS

Abbreviations	**Definitions**
QA	Question Answer
NLP	Natural Language Processing
TN	Tensor Networks
ML	Machine Learning
DNN	Deep Neural Network
NN	Neural Network
PPR	Preliminary Progress Review

Figure 1.0 Acronyms and Abbreviations

2.3 OVERVIEW

This document is organized as follows. In the third chapter a literature review on the related work is provided in the backdrop of our proposed solution. It illustrates different approaches that have been used to realize information extraction.

Fourth and Fifth chapters discusses the research problem and the research objectives. These two chapters basically introduce the research component and the main objectives of the research.

The sixth chapter, which is the most important chapter, describes the methodology used to achieve the research objectives. It includes a detailed explanation of the composition of the proposed solution.

The seventh chapter discusses about the dataset that will be used in the training process of the neural network followed by the chapter eight which depicts the benefits of the proposed solution.

Chapter nine discusses the scope of the project and the expected outcomes and the tenth chapter discusses about the constraints that might hinder the project.

The eleventh chapter provides the timeline and the schedule of the project and the final chapter includes the references which concludes the document.

3.0 LITERATURE REVIEW

Machine learning is a field of Artificial Intelligence that has been gaining a lot of prominence in the current era. It is specifically to do with building systems that are able to learn by themselves without having to be programmed. Deep learning is a subset of techniques of machine learning. Deep learning allows multiple processing layers to breakdown the given data into smaller parts and learn the representations of these data [1]. Deep learning is the state of the art in areas such as speech recognition, natural language understanding, visual object recognition, etc. Convolutional neural networks have brought many breakthroughs in areas such as processing images, pictures and speech, whereas Recurrent networks have been extremely successful in areas such as processing text and speech.

QA is a well researched area from the point of NLP (Natural Language Processing) research. QA has mostly been used to develop intricate dialogue systems such as chat-bots and other systems that mimic human interaction [2]. Traditionally most of these systems use the tried methods of parsing, part-of-speech tagging, etc that come from the domain of NLP research. While there is absolutely nothing wrong with these techniques, they do have their limitations. [3] W.A. Woods et al. shows how we can use NLP as a front end for extracting information from a given query and then translate that into a logical query which can then then be converted into a database query language that can be passed into the underlying database management system. In addition to that there needs to be a lexicon that functions as an admissible vocabulary of the knowledge base so that it is possible to filter out unnecessary terminology. The knowledge base is processed to an ontology that breaks it down into classes, relations and functions [4]. Natural Language Database Interfaces (NLDBIS) are database systems that allow users to access stored data using natural language requests. Some popular commercial systems are IBM's LanguageAccess and Q&A from Symantec [5].

Information retrieval (IR) is another technique that has been used to address the problem of QA. With IR systems pay attention to the organisation, representation and storage of information artifacts such that when a user makes a query the system is able to to return a document or a collection of artifacts that relate to the query [6]. Recent advances in OCR and other text scanning techniques have meant that it is possible to retrieve passages of text rather

than entire documents. However IR is still widely seen as from the document retrieval domain rather than from the QA domain.

Template based question answering is another technique that has been used for QA and is currently being used by the START system which has answered over a million questions since 1993 [7]. START uses natural language annotations to match questions to candidate answers. An annotation will have the structure of 'subject-relationship-object' and when a user asks a question, the question will be matched to all the available annotation entries at the word level (using synonyms, IS-A, etc) and the structure level. When a successful match is found, the annotation will point to an information segment which will be returned as the answer. When new information resources are incorporated into the SMART system, the natural language annotations have to be composed manually [8]. START uses Omnibase as the underlying database system to store information and when the annotation match is found, the database query must be used to retrieve the information. While this system has been relatively successful, it requires a lot of preprocessing which must be done manually.

Our literature survey has found that the QA domain has an active community of researchers and many different approaches have been tried to tackle this problem. While the problem of QA is a very old one, the origins of the problem can be traced back as far as the 1960's, using our access to cheaper and better computational power and newer techniques in data processing we believe we can attempt to solve this issue using a different set of tactics. This will be explained in the next section, the research gap.

4.0 RESEARCH QUESTION

Information Extraction Module can be considered as the backbone of the entire solution. This module is responsible for extracting the relevant information from the pre-processed corpus according to the question fed into the system. It should be able to relate and reason the question with the available corpus and generate an answer in an abstract form which can later be utilized by the answer generation module to generate a meaningful answer.

The research component related to this module mainly deals with the concepts of Deep Neural Networks. The research includes finding the optimal methodology to use DNN techniques to achieve this task. Since there are different sub fields in deep neural networks it is required to narrow it down and use the most appropriate approach to realize the "Information Extraction Module". It is necessary modify and adopt DNN techniques according to our need and evaluate performance of several approaches and select the most suitable approach. Another research component is the method of training the developed DNN model. It is necessary to select appropriate dataset as well as the training methodology to achieve this.

Therefore it is clear that this particular subsystem of the solution, "Information Extraction Module", some challenging research components which would require extensive research on DNN in context of information retrieval and reasoning over relationships in information.

5.0 RESEARCH OBJECTIVES

The primary objective of this module is to generate an answer to the pre-processed question by using the structured data available in the pre-processed corpus. The answer is generated and presented only if it exceeds a satisfactory level of confidence in order to ensure the reliability of the system since the proposed system is expected to operate in cases of medical emergencies and the reliability is of paramount importance. The answer generated will be in an abstract form and it will be used by the answer generation block in order to generate a meaningful answer. There are several research objectives that should be achieved in the cause of the project.

One of the research objective is to get a comprehensive understanding on deep neural networks. It is further required to explore on different types of neural networks and their usage. This knowledge will be beneficial when choosing the most suitable approach to implement the proposed system. It is also required to compare the performance of the system with related work. This can be considered as another important research objective of the project. There are different tools that are available for neural network implementations. This opens doors for another research area to explore on the performance of different tools and their feasibility. Therefore it is clear that development of the Information Extraction Module sets several research objectives that should be fulfilled in order to make the project a success

The end product would be a system that allows the user to ask medical emergency related questions in natural language form and the platform would find the most accurate answer and provide that answer in natural language form as well. The idea is to simulate a situation where the user is interacting with a person in the medical profession as close as possible. The accuracy of the answers will largely depend upon the accuracy of the data in the data set and therefore we cannot guarantee that this will be able to replace an actual medical professional. However the goal in this research is to show that using deep learning techniques we are able to reduce some of the complexities and barriers that are present at the moment and are stopping QA systems from becoming mainstream products. The medical emergency situation was chosen purely out of convenience because of the availability of the dataset. It is only a proof of concept.

6.0 RESEARCH METHODOLOGY

As mentioned before, the main objective of this project is to come up with an Artificial Neural Network (ANN) based solution for Intelligent Information Extraction. In here we are particularly focusing on addressing the Information Retrieval in a form intelligent Question and Answering System. In related work, Ontology based information extraction [9], seems to be a recently emerged subfield of information retrieval. However the inability to reason over discrete and their relationships can be identified as a major drawback in these Ontologies and knowledge base information retrieval/extraction systems [10]. Therefore we present a system which is capable of performing this task with the help of ANN particularly Recursive Neural Tensor Networks.

Artificial Neural Networks are known to perform well in intelligent systems when they are properly tweaked and used. However in order to get the optimal performance of such system it is important to select the most appropriate ANN approach and apply it accordingly. After a comprehensive study in current literature on this matter we decided to use Recursive Neural Tensor Network (RNTN) as our main inspiration for realizing this system. The selection of this methodology was done after analyzing the pros and cons as well as the relevance and the suitability of RNTN compared to other subfields in ANN. Some of the popular choices for ANN are Feedforward neural networks, Recurrent neural networks, Neural Tensor networks etc. However in order to understand the reason behind our selection over these approaches, we will first look at the main objectives expected by this module of the system.

This module, namely “Information Extraction Module” basically acts as the part where the actual decision making happens. The primary objective of this module is to generate an answer to the pre-processed question by using the structured data available in the pre-processed corpus. All these inputs will be in a form of vectors. The answer which is the expected output of this module, is generated and presented only if it exceeds a satisfactory level of confidence in order to ensure the reliability of the system since the proposed system is expected to operate in cases of medical emergencies and the reliability is of paramount importance. The answer generated will be in an abstract form and it will be used by the answer generation block in order to generate a meaningful answer.

Having understood the basic objective of the module now let us take a look at the reason behind our proposed methodology in the context of the problem statement. It should be noted that here we are generating answers from a pre-processed corpus which is already available. The Recurrent Neural Networks (RNN) shows promising performance when implemented in system where it is needed to exhibit dynamic temporal behavior in applications like handwriting recognition, speech recognition. However in our case since we are using data from a predefined corpus this might not be the most appropriate solution to address the requirement. On the other hand Neural Tensor Network specifically Recursive Neural Tensor Networks (RNTN) are more appropriate in scenarios where Natural Language Processing (NLP) and Sentiment Analysis are performed. These RNTNs are able to deal with the hierarchical relation in the words of a sentence which will be beneficial in our case.

In order to implement this ANN system, Python along with Tensorflow library will be used as main tools. This choice of the above tools was done considering their performance and the online support available for these tools compared to others. The neural network will be trained with the expectation of performing information extraction on an arbitrary (generic) rather than a specific corpus in order to realize a versatile system. The training dataset will be provided by National Health Services, England. In addition to that possibility of using datasets which are available online will also be explored.

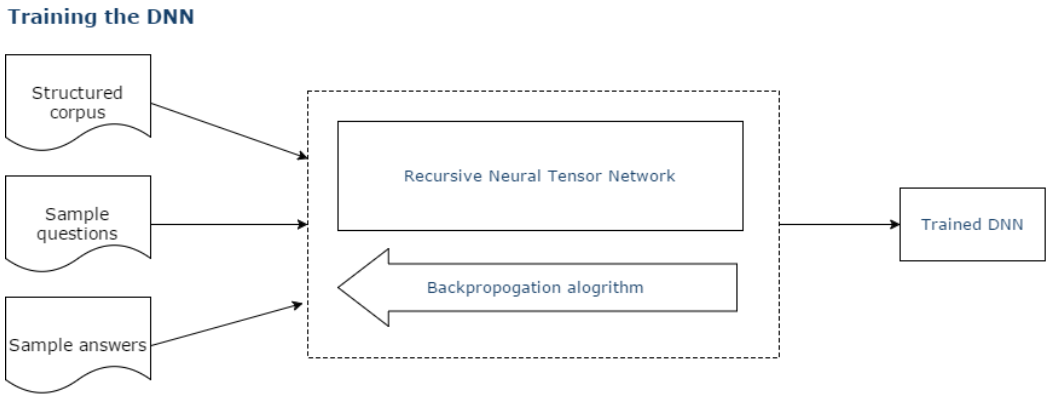

Figure 2.0 Training the RNTN

7.0 SOURCE FOR TEST DATA AND ANALYSIS

Another important factor of this component would be the dataset or the corpus since it can affect the accuracy of the output which will be used by other components like “Answer Generator”. In order to improve the accuracy we need to train our models using reliable datasets or corpuses.

Our main training corpus will be the ECDS dataset provided by NHS, and the team is currently in the process of negotiating with the responsible parties to obtain the dataset. The current state of the dataset negotiation is, that the request made asking for the dataset for the college research purpose has been forwarded to the relevant departments. If any delays occur, dataset available in the UCI Machine Learning Repository [11] will be used which contains more than 300 datasets.

To analyze the results we are using accuracy given by the *Tensorflow* library against the training datasets. There we can modify the system to increase the level of the accuracy. We can also research on optimal number of layers by increasing and decreasing the number of layers used by deep neural networks to improve the results.

8.0 ANTICIPATED BENEFITS

There are lot of benefits of having a system like this. However this section only focus on benefits of having a DNN based implementation of Information Extraction Module. There are several approaches to achieve information extraction such as ontology based information extraction. Since neural networks can show superior performance if they are configured and used properly it would be beneficial to develop such a system utilizing neural network concepts. The system is expected to perform better compared to current implementations of Information Extraction Systems based on ontologies.

Since we are particularly focusing on medical emergencies it is critical to keep the accuracy at a higher level. There can be situations where the user might not convey the message due to 'emergency'. In order to address this the system should be able to reason over the relationships in the context. Since the particular NN, Tensor Neural Networks are capable of achieving this better than other approaches it would be beneficial to have this kind of an implementation.

If the developed system shows superior performance, then it may have the potential to be marketed as a model that can be integrated to different products or as a solution that can be made available online.

Therefore it is clear that there are lot of benefits that can be gained from developing such a system both socially and economically.

9.0 SCOPE AND EXPECTED RESEARCH OUTCOME

DNN has become an emerging field in intelligent systems. It has rapidly become matured over the past decade. However there are lot of unexplored areas in DNN and our problem of interest, Information Extraction is one of them. Therefore there is a significant research gap which needs to be filled and the objective of this research is to contribute as much as possible in this context.

The expected research outcome is to explore and adopt Tensor Neural Network concepts to realize intelligent information extraction. The overall system should be able to create a platform that is able to learn from given data set and then produce direct answers that users can rely on. For the platform to be effective it is important that users can interact with it as naturally as possible. So the user must be able to type in a generic question like they would be talking to a person and the platform should produce an answer that is both factually and grammatically correct. This model should be able to facilitate this process by utilizing DNN techniques to perform Information Extraction.

10.0 RESEARCH CONSTRAINTS

There are several research constraints in realizing the proposed system. Since there are several sub fields in NN it is required to select the most suitable type of NN to be used. For neural networks to perform well it is necessary train the model properly. The training process require a suitable dataset that can be used in the training process. Since we are particularly focussing on medical domain it is a bit difficult to find a suitable dataset.

Another major constraint is that the training process is time consuming and in order to speed up this process it is required to use high end servers with GPU processing capabilities. Using these resources are costly and infeasible at the development stage. Therefore the development iterations might be time consuming.

The lack of expertise and online help in the context of Information Extraction using DNN is another constraint in this research component. Since we are focusing more on an implementation rather than a highly mathematical study, online help would be highly beneficial. Therefore this is another research constraint that we have to face.

It is clear that like every research there are several research constraints that should be tackled and dealt with appropriately in order to make this research a success.

11.0 PROJECT PLAN OR SCHEDULE

This figure below shows the project plan we follow as a team. By using this Gantt chart it will be easier to schedule tasks and workload within the team. This increases the efficiency of the team as well.

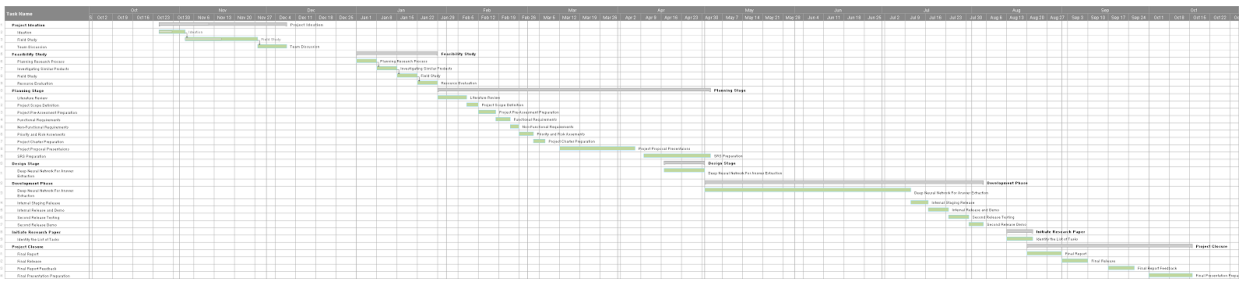

Figure 3.0 Gantt Chart of the Component

12.0 REFERENCES

[1] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning", Nature, vol. 521, pp. 436 – 444, 2015

[2] Silvia Quarteroni, “A Chatbot-based Interactive Question Answering System”, 11th Workshop on the Semantics and Pragmatics of Dialogue, 2007

[3] W.A Woods, R.M. Kaplan and B. Nash-Webber, “The lunar sciences natural language information system”, BBN Rep. 2378, Bolt Beranek and Newman, Cambridge, Mass., USA, 1977

[4] T.R. Gruber, “A translation approach to portable ontology specifications”, Knowledge Acquisition, 5 (2), 1993.

[5] R. Dale, H. Moisl and H. Sommers, Handbook of Natural Language Processing, 1st ed. New York: Marcel Dekker AG, 2006, pp. 215 - 250.

[6] L. Hirschman and R. Gaizauskas, “Natural language question answering: the view from here”, Natural Language Engineering, 7 (4), 2001, pp. 275-300.

[7] "The START Natural Language Question Answering System", Start.csail.mit.edu, 2017. [Online]. Available: http://start.csail.mit.edu/index.php. [Accessed: 26- Mar- 2017]

[8] B. Katz, G. Borchardt and S. Felshin, “Natural Language Annotations for Question Answering”, Proceedings of the 19th International FLAIRS Conference (FLAIRS 2006), 2006

[9] Daya C. Wimalasuriya Dejing Dou, “Ontology-based information extraction: An introduction and a survey of current approaches “, Journal of Information Science Vol 36, Issue 3, pp. 306 - 323

[10] Socher, Richard, Chen, Danqi, Manning, Christopher D. and Ng, Andrew Y. Reasoning with neural tensor networks for knowledge base completion. In Advances in Neural Information Processing Systems, 2013a.

[11] J. Martens, “Deep learning via Hessian-free optimization”, In Proceedings of the 27th International Conference on Machine Learning (ICML). ICML 2010, 2010.

APPENDIX I: OVERALL ARCHITECTURE OF THE SYSTEM

The end product would be a system that allows the user to ask medical emergency related questions in natural language form and the platform would find the most accurate answer and provide that answer in natural language form as well. The idea is to simulate a situation where the user is interacting with a person in the medical profession as close as possible. The accuracy of the answers will largely depend upon the accuracy of the data in the data set and therefore we cannot guarantee that this will be able to replace an actual medical professional. However the goal in this research is to show that using deep learning techniques we are able to reduce some of the complexities and barriers that are present at the moment and are stopping QA systems from becoming mainstream products. The medical emergency situation was chosen purely out of convenience because of the availability of the dataset. It is only a proof of concept.

In order to achieve this, system is broken it down into four different components. Each of these components form a critical part of the system and carry out a critical function. They also have deeply integrated deep learning techniques in each component, which we have described in great detail in the methodology section. In the next section we have a brief overview of what each of the sub objectives are supposed to accomplish.

The four component of the system are,

- Corpus Preprocessing
- Question Preprocessing
- Deep Neural Network For Answer Extraction
- Answer Generation

This document contains the in-depth details of one of the research component. I.e Corpus Preprocessing.

APPENDIX II: SYSTEM ARCHITECTURE DIAGRAM

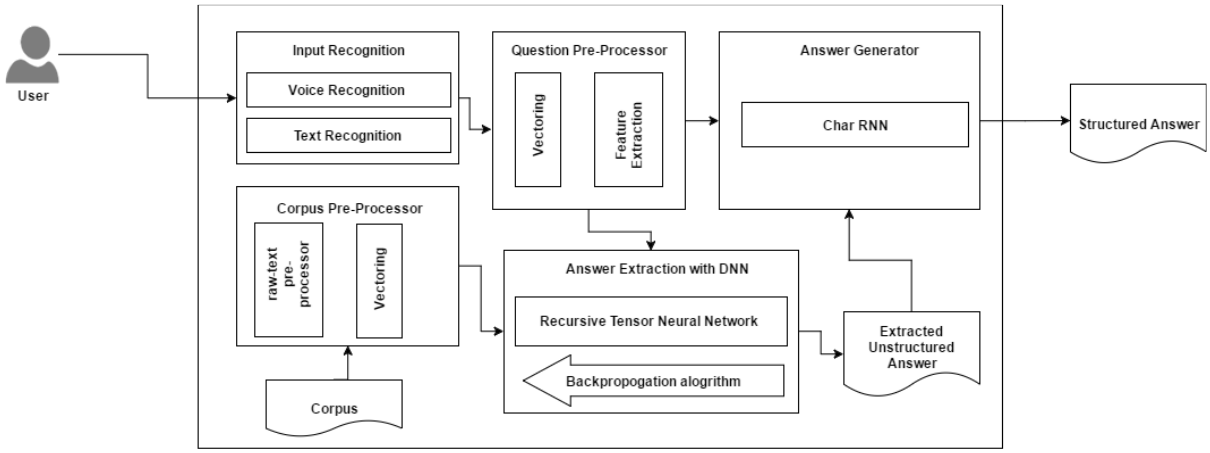

Figure 4.0 System Architecture of Adaptive Artificial Intelligent Question Answer

APPENDIX III: GANTT CHART OF THE OVERALL SYSTEM

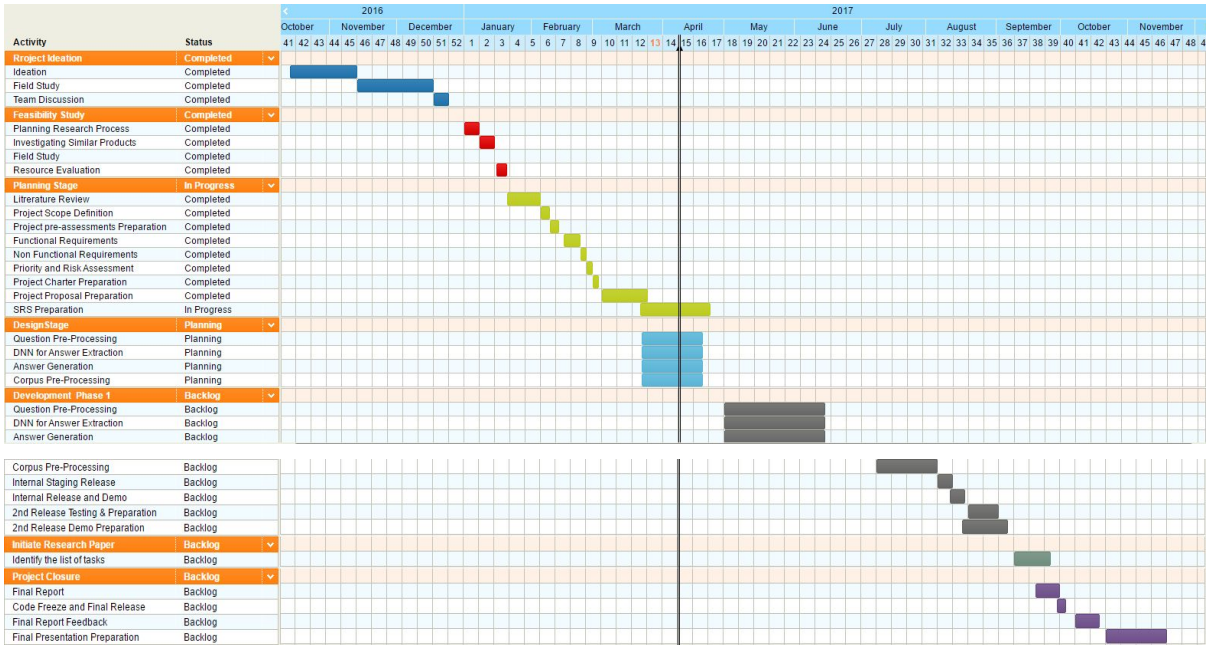

Figure 5.0 Gantt Chart of the Overall System

APPENDIX IV: SPECIFIC TASKS TO BE FOCUSSED

1. Requirements Gathering
2. System Design
3. Research
4. Implementation
5. Web Interface
6. Mobile Application (Android only)
7. System Testing
8. Continuous Integration and Deployment

APPENDIX V: DESCRIPTION OF PERSONAL AND FACILITIES

The description of the personnel involved in this project is as follows:

Supervisor: Mr. Yashas Mallawarachi

External supervisor: Mr. Anupiya Nugaliyada

Implementation team of Adaptive Artificial Intelligent Question Answer System:

Akram M.R. (Leader)

Deleepa Perera

Singhabahu C.P.

Saad M.S.M.

The owner of this document and Corpus Preprocessing component:

Singhabahu C.P.