QUESTION PREPROCESSING

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)

AKRAM M.R (IT14109386)

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

Department of Information Technology

Sri Lanka Institute of Information Technology

May 2017

Table Of Contents

LIST OF FIGURES

LIST OF TABLES

1.0 Introduction

1.1 Purpose

The purpose of this report is to present “Preliminary Progress” of the question pre-processing component in the **Adaptive Artificial Intelligent QA System.** This report will give the reader a comprehensive understanding of the project scope, objectives, methodologies, constraints and the schedule of the project. The intended audience for this document is any party who has referred to the project proposal document of the platform and any other party interested in the projects project progress which falls under the CDAP module.

1.2 Acronyms, Abbreviations and Definitions

QA	Question Answer
NLP	Natural Language Processing
IE	Information Extraction
IR	Information Retrieval
QP	Question Preprocessing
ML	Machine Learning
DNN	Deep Neural Network
NN	Neural Network
POC	Proof Of Concept
i.e	That is
AAIQA	Adaptive Artificial Intelligent Question Answer
HoG	Heart of Gold

Table 1.0 Acronyms and Abbreviations

Definitions

Corpus	Dataset or a collection of data
Supervised Learning	Analyzes the training data and produces an inferred function, which can be used for mapping new examples
Unsupervised Learning	Is a cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data
End User	The person who actually uses a particular product
Preprocessing	Extract meaningful sets of data in the context of question preprocessing

Table 2.0 Definitions

1.3 Overview

This report contains a further eleven sections. The third section in this report illustrates the literature review comparing existing solutions and methodologies in the QP domain. and the next three sections describes the **research question, research objective** and **research methodology**. Research question addresses three vital questions the Question Preprocessing component is trying answer, the research methodology addresses the methods and techniques used to achieve the objectives mentioned in section four. Section six describes the sources and the training dataset that will be used to train and build the QP model, furthermore the section describes the data collection and analysis methods used.

Section seven and eight discusses anticipated benefits for the users of the overall system and the benefits to future work in the area of question preprocessing. Section eight specifically discusses the expected research outcome and the scope of the QP component. Section nine bring forwards the constraints of the entire system which the corpus and how we plan to mitigate the risk of the two constraints mentioned. Section ten consists of the gantt chart and further sections consist of references and appendix.

2.0 Literature Review

Artificial Intelligence 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 such that it ends up to be a giant truth statement. Deep learning is a subset of AI. Deep learning allows multiple processing layers to breakdown the given data into smaller parts and learn the representations of these data. Deep learning has been a breakthrough research area achieving high success rate in the areas of speech recognition, image processing etc.

QA is a well researched domain in the point of NLP and NLU. One of the significant tasks in a QA system is to represent the Natural Language question such as "What is the weather today ?" etc, in a machine readable/understandable format. Traditionally when QA systems use an Informational Retrieval (IR) based approach with an ontology built with significant human effort, the task of QP would be to generate a query out of the given question and the query can be executed in the underlying DBMS system [2]. The knowledge base in such system is always domain specific and further drilled down to categories within the domain. The primary task of QP in such an instance would be identify in what direction or category the question is presented by the user.

Another approach for QA system is closed domain QA systems. The question analysis in the closed domain system by A. Frank and co[1] uses a popular architecture known as Heart of Gold (HoG), they present an hybrid approach utilizing NLP techniques. A question is linguistically analysed by the Heart of Gold (HoG) NLP architecture, which flexibly integrates deep and shallow NLP components, In this architecture the initial stages of QP uses syntactic and semantic analysis.

Another approach for QA systems is the rule based approach. Rule based QA systems are an extended form of IR based systems. Rule Based QA doesn't use deep language understanding or specific sophisticated approaches [1]. A broad coverage of NLP techniques are used in order to achieve accuracy of the answers retrieved. Some popular rule based QA systems such as Quarc and Noisy channel generates heuristic rules with the help of lexical and semantic features in the questions. For each type of questions it generates rules for the semantic classes like who, when, what, where and Why type questions.

The current trends in QA systems is to use a neural network based approach. Where the neural nets uses a preprocessed dataset to learn patterns and extract information accordingly. Pre-processing of the question and the dataset uses a popular NLP technique known as word embedding. Word embedding has the ability to map words with a semantic relationship in close proximity in a low dimensional vector space [3], this allows the neural network to understand the context of the question and the context of the dataset which is achieved in the preprocessing layer of the dataset. However this embedding layer in QP has several variations and includes different other techniques to enhance the representation of the vector.

There is a wealth of information on the internet. For any given domain we are able to find a huge amount of information. However to use this information effectively, there needs to be a system to process the data and extract out the meaningful information. Further it is important to provide a simple and seamless way of interacting with this data. This has given rise to the field to natural language question answering where a user must be able to ask a question in everyday language and receive a factually correct answer quickly. In the literature survey we discussed few different techniques that have been used to tackle this problem, NLP, Information Retrieval and Information Extraction, Categorization. All the methods has its own flaws where either the accuracy is not high enough or it may take a lot of manual processing and so on.

3.0 Research Question

The key focus area in the question preprocessing component is to preprocess the question presented by the user in natural language form such that it is represented in a machine readable format. Since the **Adaptive Artificial Intelligent QA System** uses a neural network based deep learning approach, the task in QP is to generate a vector that the neural networks can understand. One of the main challenges in generating a vector is to maintain the syntax and semantics of the question asked and further map it in the most efficient and effective way such that the processing of the vector by the neural networks to extract the answer consumes as less time as possible and achieves the highest accuracy.

QP component should be able to identify the most efficient representation of the vector space and it needs to follow strict process in order quantify the most effective and efficient representation.

In summary the research question put forward for question preprocessing can be identified as the following

1. What is the most efficient machine readable representation of the natural language question ?
2. How can we preserve the syntax and semantics of the question within the representation ?
3. How can we ensure the representation is the most effective and efficient for the chosen neural network model for the system ?

Throughout the research of this component we plan to address the above mentioned questions utilizing various deep learning and natural language processing techniques.

4.0 Research Objective

QA systems can take many different avenues in terms of its approach as illustrated in the literature review. Since we will be utilizing a neural network based approach the question presented by the user in natural language form will need to be represented in a vector space such that the vector space preserves the syntax and semantics of the question. In trying to achieve the aforementioned tasks we will need to identify and map these requirements to stable technical approaches. These approaches will also need to be tweaked in order to comply with the variation of the neural network model that will be deployed in the final system. Since the **Adaptive Artificial Intelligent QA System** targets the **Medical Emergency** domain the semantics of that domain needs to be studied and incorporated in the preprocessing stage.

Thus, we can conclude that the main objective of the QP component is to construct a vector space representing the question asked by the user. The vector should be constructed in a way it preserves and maps the semantic relationships of the words in close proximity in a low dimensional vector space. The importance of preserving the semantics is that it enables the neural network to identify the context of the question through the pattern it's represented in. Due to this a uniform technique needs to be identified that can also preserve the context and enables the neural network to perform efficient and effectively.

5.0 Research Methodology

In a Q&A based system understanding and processing a question is vital to provide the user with the most appropriate answer. Machines do not understand text as humans do thus questions inputted as texts need to be transformed into a vector that preserves the context of the question that was presented by the user. The objective of this area is to provide the neural network with the most effective vector format that would preserve and best represent the context of a question in a vector so that the DNN can utilize the vector to process and find the most appropriate answer.

For this purpose we will be using few natural language modelling techniques such as word embedding,syntactic analysis, Identification of question type such as a wh-question analysis.
Word Embedding is used to map words or phrases from a vocabulary to a corresponding vector. This type of representation has two important and advantageous properties:

1. It is a more efficient representation
2. It is a more expressive representation

Word embedding maps each word to a vector space.The Embedding layer will map each token from the question to its corresponding vector space, which preserves the contextual similarity of words in the vector space.

The embedding layer in the question processing component can be done through a popular pre-trained word embedding model known as word2vec or glove(exact model will be chosen based on further trial and error)[5]. Word2vec is a small two layer neural network. It contains two distinct models (CBOW and skip-gram), each with two different training methods (with/without negative sampling) and other variations [5]. To top that, it also contains a sharp pre-processing pipeline, whose effects on the overall performance is yet to be evaluated.

The other sub-component is Syntactic analysis of a question. Syntactic analysis is the process of identifying the structure of a sentence,The interplay of syntax and semantics of natural language questions is of interest for question representation. Researchers in the area of question understanding reccomend the use of a TreeLSTM neural network for this purpose[6], since it is capable to capture long distance interaction on a tree. The other option would be to go with a chain structured LSTM but the critical downfall of it for this specific task is that it fails to capture long distance interaction on a tree[6].

To obtain the parse tree information some of the available open source parsers such as the Noah's Ark parser [7], or the Standford Core Parser can be used.

Questions by nature are composed to fulfill different types of information. A what question and a how question requires different types of information [6]. In Order to incorporate this a Wh-Analysis will be required, thus an additional layer for question adaption will be required. Wh-word is basically the question word which is one of who, why, where, which, when, how, what and rest. “rest” are the questions that don't have any question word. Example :- Name of a disease that cause bowel bleeding?. This process segregates questions into different types and considers the type of question for answer generation, a recommended approach for this would be to encode question type information into a one hot vector which is a trainable embedding vector [6], and it is incorporated in the training process.

For the purpose of implementation the popular deep neural network library tensor flow along with the python programming language will be used. For the purpose of syntactic analysis the Standford core NLP parser will be used.Python enables us to carry out string manipulation easily unlike other programming languages since user inputs are text based it will be the preferred choice of language.The choice of the above mentioned technology is due to the widely available community support and free distribution of the software.

Work Flow diagram of the question preprocessing component is given below.

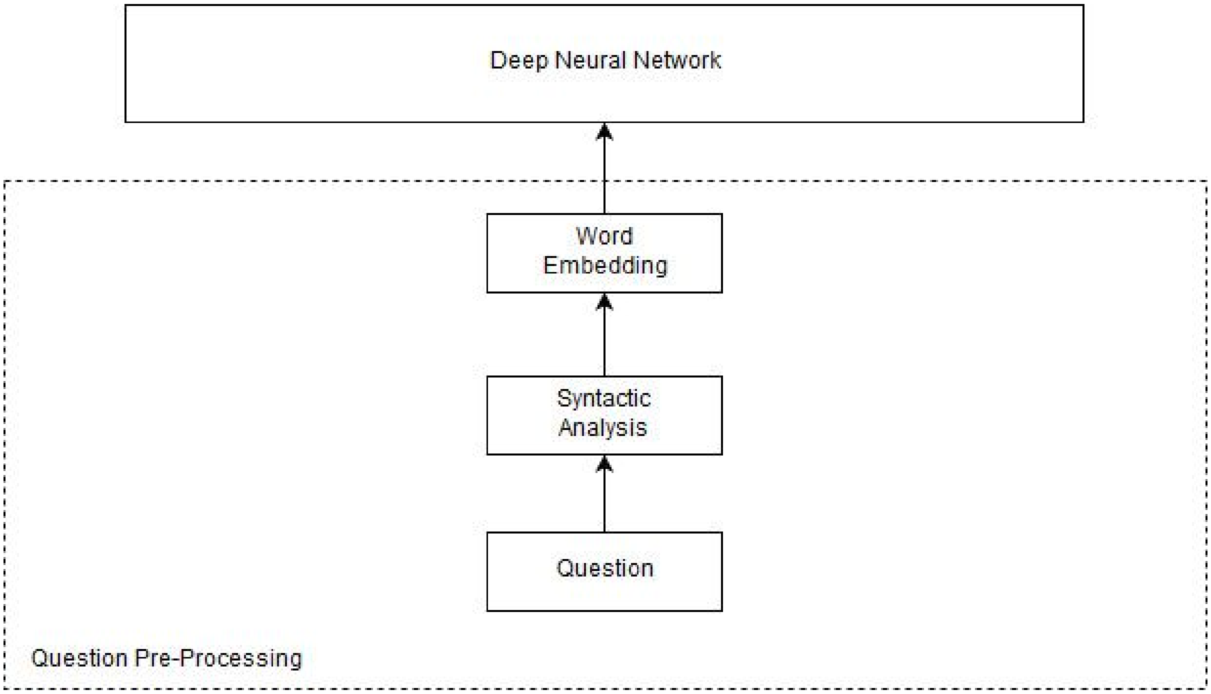

Figure 1.0 Question pre-processing module

6.0 Source For Test Data and Analysis

6.1 Corpus

The training corpus used to support the domain the QA system is built on which is the **Medical Emergency** domain will be the ECDS dataset [8]. Our team is currently working on acquiring this dataset. Until then we will be using a dataset from UCI Machine Learning Repository [9] .
For the purpose of the QA component the CNN daily QA dataset will be used to evaluate the pre-trained word embedding model and tweak and measure its performance although the Questions in the dataset are not related to the **Medical Emergency** domain we could still use them to figure out the efficiency of the representation of the question. Furthermore, a small test data set with domain specific questions will be compiled to perform integration tests with when the entire system is integrated and ready to go live.

6.2 Evaluation and Analysis of results

Since word embedding is one the vital parts in QP it is relevant to focus on techniques that try to evaluate word embedding models. Thus, a paper by Tobias Schnabel, Igor Labutov. David Mimno, Thorsten Joachims presents a solid evaluation approach for word embedding layers [10]. They categorize the evaluation approach to four main types,

1. Relatedness: These datasets contain relatedness scores for pairs of words;
2. Analogy: This task was popularized by Mikolov et al. (2013a). The goal is to find a term x for a given term y so that x : y best resembles a sample relationship a : b.
3. Categorization: Here, the goal is to recover a clustering of words into different categories
4. Selectional preference: The goal is to determine how typical a noun is for a verb either as a subject or as an object (e.g., people eat, but we rarely eat people)

We can use the similar approach to analyze and evaluate our QP model.

7.0 Anticipated Benefits

The AAIQA System for **Medical Emergency** domain presents the users with the ability to look up to help in a medical emergency situation. This can drastically help patients who are victims of such situations and also assist people involved in such situations. Apart from the practical benefits the QP component enables the NN layer to process a vector that has the near exact representation of the question in a machine understandable format.

Apart from user benefits identifying an approach that deviates from current approaches and if the evaluation of the approach proves to have out performed current approaches this enables future work in the QP domain to adapt and evolve the approach we have used.

8.0 Scope and Expected Research Outcome

The outcome of the research in question preprocessing is to construct a low dimensional vector utilizing a text preprocessing layer and a word embedding layer. The overall outcome would be to generate a representation of a natural language question which is more expressive and efficient such that it preserves original semantic relationship of the question. The outcome of that would the efficient functioning of the neural network such that it requires minimum time to process the vector to identify and extract an answer. The work carried out in this component will be documented in the research paper that is planned to be written at the closure of the project. Thus, it will be available as a reference for future work.

The scope of this component would be to expect a natural language question in the form of a text as input and transform it to a machine understandable format as an output such that the syntax and semantics of the question is preserved. The output of this component will an input for the Neural Network layer that extracts the answer.

9.0 Research Constraints

The constraints in this research area can be mainly categorized into two parts,

1. Corpus
2. Lack GPU for training

The corpus is one of the main constraints. As previously mentioned in some other chapters of this document that the end system will be based upon the medical emergency domain. Where the system will answer questions related to medical emergency domain. As the chosen system is a closed QA system, it depends on the dataset.

The corpus will be the ECDS dataset provided by NHS, and the team is currently in the process of negotiating with the responsible people to obtain the dataset. The current state of the dataset negotiation is that the request made asking for the dataset, for this research purpose has been forwarded to the relevant department and they have got back to us mentioning the request is being processed. If any delays made, datasets available in the UCI Machine Learning Repository [9] will be utilized.

Since Deep Neural Networks require high powered GPUs to be able to train fast these resources are scarce. Thus, the ideal solution for this problem is to utilize such resources given by public cloud vendors as services such as Google Cloud, AWS, and Azure. One of the mentioned cloud providers services will be utilized to train the combined systems neural network models.

10.0 Project Plan

The following gantt chart illustrates the project plan for Question Preprocessing component.

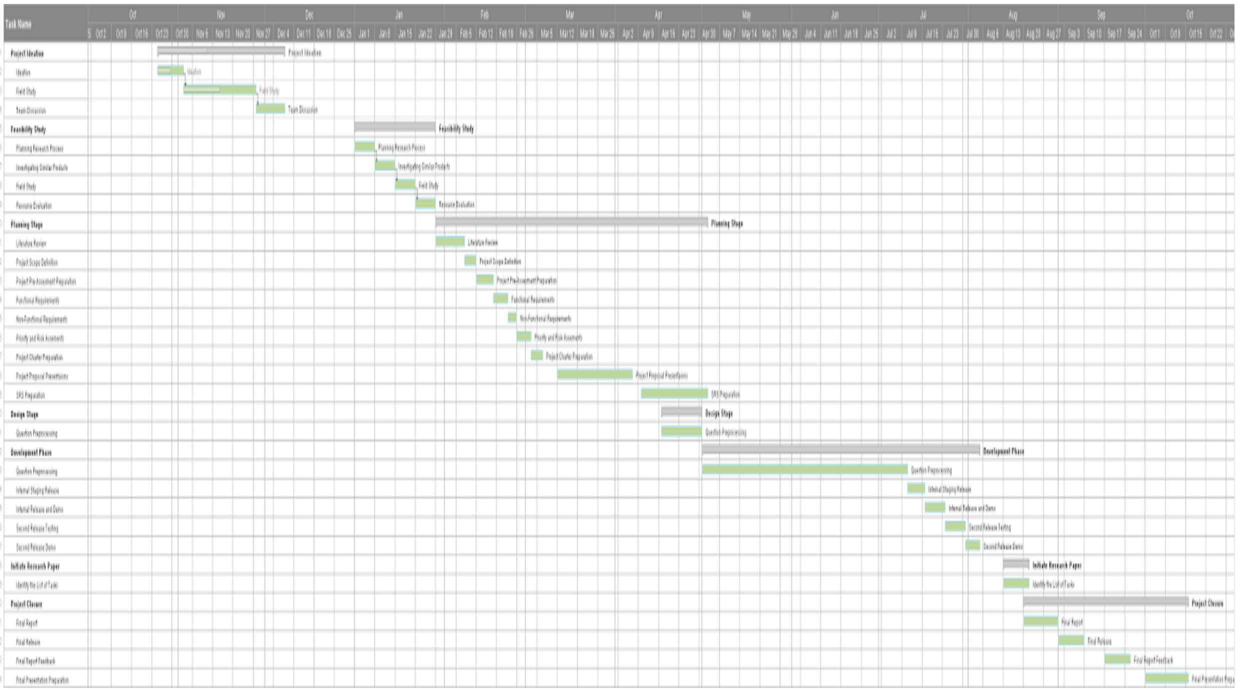

Figure 2.0 Gantt Chart

11.0 Reference

[1] A. Frank, H. Krieger, F. Xu, H. Uszkoreit, B. Crysmann, B. Jörg and U. Schäfer, "Question answering from structured knowledge sources", *Journal of Applied Logic*, vol. 5, no. 1, pp. 20-48, 2007.

[2] P. Gupta and V. Gupta,*p* "A Survey of Text Question Answering Techniques", *International Journal of Computer Aplications*, vol. 53, no. 4, pp. 1-8, 2012.

[3]"On word embeddings - Part 1", *Sebastian Ruder*, 2017. [Online]. Available: http://sebastianruder.com/word-embeddings-1/. [Accessed: 01- May- 2017].

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

[5] Ruder, Sebastian. "On Word Embeddings - Part 3: The Secret Ingredients Of Word2vec". *Sebastian Ruder*. N.p., 2017. Web. 26 Apr. 2017.

[6]J. Zhang, X. Zhu, Q. Chen, L. Dai and H. Jiang, "Exploring Quesiton Understanding and Adaptation in Neural-Network-Based Question Answering", 2017.

[7] "Noahs-ARK/semafor", *GitHub*, 2017. [Online]. Available: https://github.com/Noahs-ARK/semafor. [Accessed: 27- Apr - 2017].

[8] N. England, "NHS England » Emergency Care Data Set (ECDS)", *England.nhs.uk*, 2017. [Online]. Available: https://www.england.nhs.uk/ourwork/tsd/ec-data-set/. [Accessed: 01- May- 2017].

[9] "UCI Machine Learning Repository: Data Sets", *Archive.ics.uci.edu*, 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets.html. [Accessed: 01- May- 2017].

[10] T. Schnabel, I. Labutov, D. Mimno and T. Joachims, "Evaluation methods for unsupervised word embeddings", 2017.

APPENDIX I: SYSTEM ARCHITECTURE DIAGRAM

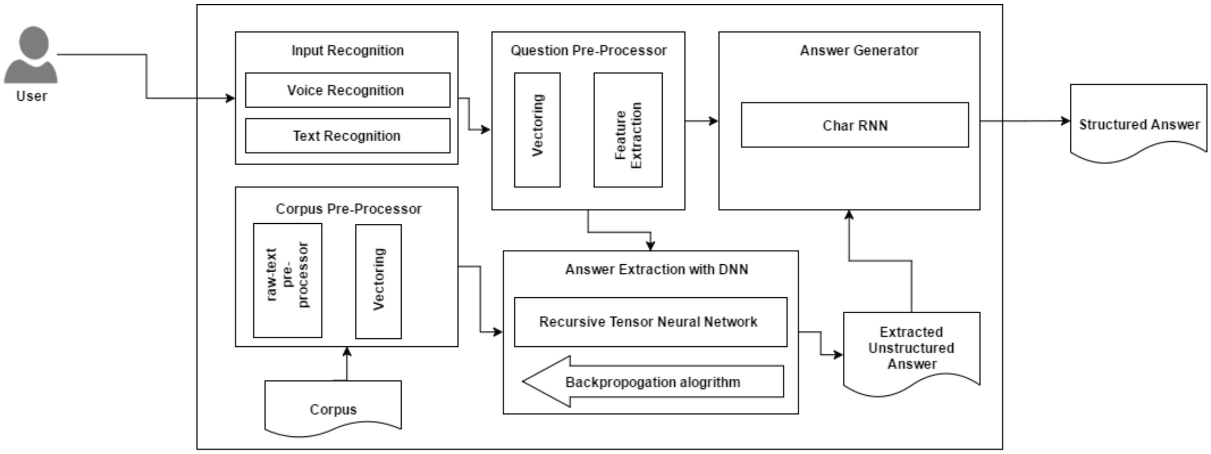

Figure 3.0 System Architecture of Adaptive Artificial Intelligent Question Answer

APPENDIX II: 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 Question Preprocessing component:
Akram M.R