CORPUS 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)

Saad M.S.M. (IT14109072)

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

LIST OF TABLES

2.0 INTRODUCTION

This section focuses on the purpose and overview of everything included in this PPD document. Further a list of abbreviations, definitions and acronyms related to this document is provided.

2.1 Purpose

The purpose of this document is to provide the preliminary progress of the “Corpus Preprocessing” 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

Acronyms and Abbreviations

QA	Question Answer
NLP	Natural Language Processing
IE	Information Extraction
IR	Information Retrieval
OCR	Optical Character Recognition
ML	Machine Learning
DNN	Deep Neural Network
NN	Neural Network
POC	Proof Of Concept
i.e	That is

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 corpus preprocessing

Table 2.0 Definitions

2.3 Overview

The remainder of this document contains another eleven chapters. The third chapter provides the details on the background information and overview of previous work based on comprehensive literature survey and the current state of the research problem.

The fourth chapter focuses on the research question of the component on how the problem arises and evolves. The fifth chapter relates to the previous chapter as it provides the relevant objective of the research component.

The sixth chapter is based on the research methodology of the component. It outlines the methods which will be used to achieve the research aim. Further discussion on conceptual framework and identified matrices are taken place.

The seventh chapter provides the details in regards to the data collection procedure to be used and data analysis methods to be used. The eight chapter describes on the benefits for the users and the contribution to the body of knowledge on taking place of this research component.

The ninth chapter focuses on the expected outcome. The tenth chapter is describes the constraints that may limit developers options. The eleventh chapter is about the project plan and schedule which will in return provides the idea of the feasibility of the research.

The twelfth chapter includes all the references made during the writings of this document and the thirteenth chapter includes the appendix of this 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 Language Access and Q&A from Symantec [5].

When it comes to Corpus Preprocessing Information Extraction (IE) [6] is also another technique. In simple IE identifies the keywords and relationship between those. It does this by a process called pattern matching, by looking for predefined sequences in the text. IE infers the relationship between places, people and time to provide the user with meaningful information. This technique is useful when handling with large volume of data. To have the best outcome through this technique traditional data mining process expects the information to be mined already in the form of a relational database. Unfortunately dataset/corpus are available (most of the time) in the form of free natural language documents rather than structured database [7]. This process is depicted in Figure 1.0.

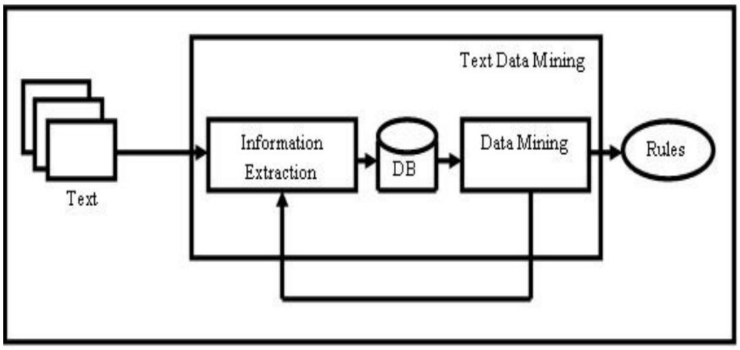


Figure 1.0 Process of Text Extraction

Information Retrieval (IR) is another technique that has been used to address Corpus Preprocessing. 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 return a document or a collection of artifacts that relate to the query [8]. 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.

Categorization is also another technique which involves in identifying the major subjects of a document through inserting the document into a predefined set of topics. It actually does not try to process the actual information. Rather, it only counts words that appears and form that count it identifies the main topics that the document covers. So that categorization technique often relies on a glossary for which topics are predefined and relationships are identified by looking for large terms, narrow terms, synonyms and related terms [9]. The major drawback in this technique of corpus preprocessing is that it misses out the syntactical meanings of the words.

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 have its own flaws where either the accuracy is not high enough or it may take a lot of manual processing and so on. This has meant that while this is a significant problem domain due to the high costs there haven't been any commercially viable solutions yet.

4.0 RESEARCH QUESTION

Corpus Preprocessing is one of primary component in the research project. Initially the corpus will be unstructured text data, which can be understood by humans, not by machines. It simply implies that, there is a necessary for a transformation of such data because many machine learning algorithms including deep neural networks, require inputs to be vectors of continuous values; they won't just work on plain text or strings.

Since the amount of information is rapidly growing and it is becoming difficult to keep manually created knowledge bases and ontologies uptodate. Than being relying on manually created knowledge bases, applying deep learning techniques to unstructured data or the corpus to identify relationship of words and providing the proper transformation of unstructured data to a representation as an input to the neural network training model, will provide an immense help in extracting the needed answer.

To summarize on the research question, the corpus or the dataset will be of unstructured data and cannot be understood by machines or the machine learning algorithms. Therefore the corpus need to be preprocessed into a form where it can be understood by the deep neural networks prevailing the semantic and syntactic relationship of the words.

5.0 RESEARCH OBJECTIVES

When it comes to deep learning QA systems there are two broad categories that can target. There are open datasets such as the Allen AI Science and Quiz Bowl datasets, and closed datasets such as the ones provided by Facebook (bAbI) [10]. Open QA systems require using the information provided in the dataset as well as any additional available knowledge. This requires some Information Retrieving techniques. In real world applications this would be the most likely required solution, however for the purposes of this research, have chosen to focus on a closed QA system, where the answer to a question would depend on the given dataset.

To further narrow down the scope, have chosen to build a question answer system for the **medical emergency domain.** As a case study and proof of concept we plan to implement a QA System that focuses only on answering questions related to medical emergency situations and not addressing open domain questions. The scope has been thus narrowed to, first, increase the accuracy of answers provided and second, to ensure that the project can be completed in the allocated time. 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 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.

Primary objective of corpus preprocessing is that the preprocessed corpus is the input for training model and the answer is extracted through the training model based on the preprocessed corpus. Therefore the preprocessed corpus should contain a proper representation of data as it is the input for the training model. Intention behind the corpus preprocessing is the transformation of raw text into a representation of vectors in a low dimensional vector space along with maintaining the words which are similar in close proximity through understanding the contextual similarity of words and mapping the semantic relationship of words.

6.0 RESEARCH METHODOLOGY

Corpus Preprocessing is one of primary component in the research project. Initially the corpus will be unstructured text data, which can be understood by humans, not by machines. It simply implies that, there is a necessary for a transformation of such data because many machine learning algorithms including deep neural networks, require inputs to be vectors of continuous values; they won't just work on plain text or strings.

Since the amount of information is rapidly growing and it is becoming difficult to keep manually create knowledge bases and ontologies uptodate. Than being relying on manually created knowledge bases, applying deep learning techniques to unstructured data or the corpus to identify relationship of words and providing the proper transformation of unstructured data to a representation as an input to the neural network training model, will provide an immense help in extracting the needed answer.

One of the major application of such transformation of data [11], is Word Embedding which is a natural language technique. It is used to map words or phrases from a vocabulary to a corresponding vector of real numbers. Word embedding aims to create a vector representation [12] with a much lower dimensional space. In contrast Bag of Words [13] approach, which often results in huge and sparse vectors. In Bag of Words approach the dimensionality of the vectors representing each document is equal to the size of the vocabulary.

Word Embedding [14] is also used for semantic parsing, to extract the meaning from text to enable Natural Language Understanding. For a language model to understand the meaning of a word, it need to know the contextual similarity of words. For example if we tend to find diseases in sentences, where diabetes, diarrhea, HIV should be of close proximity. So the vectors created by word embedding preserves these similarities along with the words that regularly occur nearby in text will also be in close proximity.

Word embedding is all about building a low dimensional vector representation from corpus, which preserves the contextual similarity of words. There are word embedding techniques such as,

- 1-of-N vec (one-hot-vec)
- GloVe (Global Vectors)
- Word2vec

In a simple 1-of-N [15] encoding transforms categorical features to a format that works better with classification and regression algorithms. In simple terms every element in the vector is associated with a word in the vocabulary. The encoding of a given word is simply the vector in which corresponding element is set to one and all others are set to zero. Suppose we consider a vocabulary of five words; diabetes, diarrhea, HIV, obesity, paralysis. We could encode the the word obesity as [0, 0, 1, 0, 0]. In such a scenario, the only comparison that can be made between vectors is equality testing or it engages all features and tell which is present and which is absent for a particular set of output.

GloVe [16] is a 'count based' model. Which means GloVe learn their vectors through collecting word co-occurrence statistics in a form of word co-occurrence matrix **X**. Each element $\mathbf{X}_{\mathbf{ij}}$ of such matrix represents how often word ***i*** appears in context of word ***j*** in a large corpus. The number of "contexts" is of course large, since it is essentially combinatorial in size. Then factoring this matrix to yield a lower-dimensional matrix, where each row now yields a vector representation for each word.

In specific, the creators of GloVe illustrate that the "ratio of the co-occurrence probabilities of two words (rather than their co-occurrence probabilities themselves) is what contains information and look to encode this information as vector differences". But when it comes for computation GloVe will be taking more memory [17], because it precomputes the large word into word co-occurrence *(large word x word co-occurrence)* matrix in memory[18]. Also sometimes there is a restriction on vocabulary since GloVe requires memory quadratic in the number of words: it keeps that sparse matrix of all *word x word co-occurrences* in RAM.

Word2vec [19] is a predictive model which is one of the most popular word embedding model. This simply learns relationship between words without any prior knowledge about the domain. The output are vectors, one vector per word with an exceptional linear relationship. Once a vector model is created out of the corpus, word2vec provides two basic tools namely *analogy* and *distance* to use.

- Distance - tool provide a list of words which are closely related to a particular word from the vector model.
- Analogy - tool is provides the ability to query for textual regularities captured in the vector model.

For example, let us assume that we use word2vec to create a vector model of the words appearing in a corpus of medical domain. If the resulting vector space represents diseases and cause of disease is projected in a two dimensional vector space, we can observe a relationship between each disease and the cause of the disease, and also similar diseases are placed closed to each other in vector space.

Further word2vec is based on two architecture: *continous bags of words (CBOW* and *skip-gram (SG)*. Since word2vec as no built-in functionalities for term normalisation, the corpus needs to processed before they could be used for word2vec. Unprocessed corpus contains syntactic variations, stopwords, punctuations which has a negative impact on on how word2vec indexes the term and which will affect the quality of vector space representation.

Before providing the corpus to the word2vec it needs to be processed as follow,

- All punctuations and unnecessary white spaces needs to be removed.
    - Eg: the term *obesity* and *obesity. (with full stop)* can be indexed as two separate terms.
- Transforming all words to lower-case.
    - Eg: the terms *Obesity obesity OBESITY* can indexed as three separate terms.
- Merging muti-word terms.

- ○ Eg: the terms *Human Immunodeficiency Virus* transformed to *human_immunodeficiency_virus* so that word2vec can identify it as a single term.

(Merging multi word terms is not yet finalized, decision will be taken during the implementation phase. To achieve this should create a separate dictionary of multi word terms for the selected domain using knowledge source.)

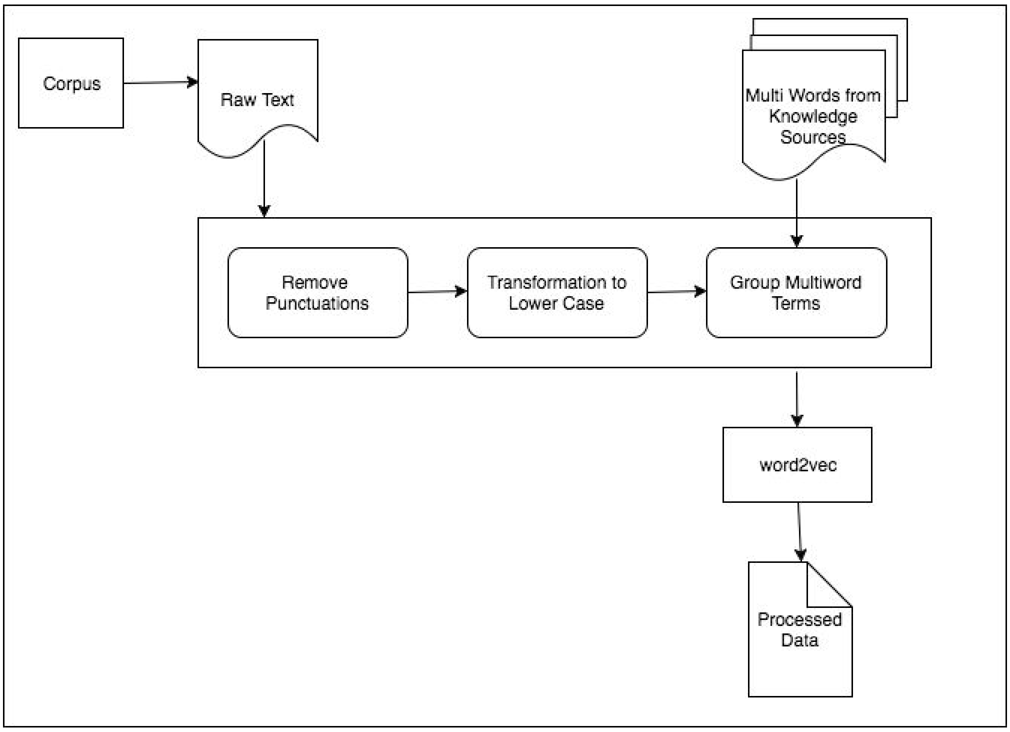


Figure 2.0 Workflow of Corpus Preprocessing

The preferred choice of technology for this component will be Python along with Tensorflow. Tensorflow is a popular machine learning platform with a lot of community support and ease of use, along with Tensorflow keras.io will also be used. Keras is a superset of Tensorflow. Python enables us to carry out string manipulation easily unlike other programming languages since corpuses are text based it will be the preferred choice of language.

7.0 SOURCE FOR TEST DATA AND ANALYSIS

When it comes to deep learning QA systems there are two broad categories that can target. There are open datasets such as the Allen AI Science and Quiz Bowl datasets, and closed datasets such as the ones provided by Facebook (bAbI) [10]. Open QA systems require using the information provided in the dataset as well as any additional available knowledge. This requires some Information Retrieving techniques. In real world applications this would be the most likely required solution, however for the purposes of this research, have chosen to focus on a closed QA system, where the answer to a question would depend on the given dataset.

To further narrow down our scope we have chosen to build a question answer system for the **medical emergency domain.** 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 the college research purpose has been forwarded to the relevant department. If any delays made, dataset available in the UCI Machine Learning Repository [20] will be used which contains more than 300 datasets.

Major data analysis methods which will be used in sequence in the research component of corpus preprocessing is already described in detail under research methodology. In brief data analysis methods of this component to extract out the useful information suggesting conclusions and supporting decision making is explained below.

Stopwords elimination; stopwords are a division of natural language. The motive that stopwords should be removed from a text is that they make the text look heavier and less important for analysts. Removing stop words reduces the dimensionality of term space. Example for stopwords: the, in, a, etc. There are few stopwords removal methods such as; the classic method, methods based on Zipf's Law, the mutual information method and term based random sampling.

All punctuations and unnecessary white spaces needs to be removed. Because having such punctuation and white spaces will make a major difference between the same terms which prevails the same semantic but only because of having punctuation and unnecessary white spaces considered as two different terms in neural network techniques. For example the term *obesity* and *obesity. (with full stop)* can be indexed as two separate terms if punctuations not removed.

Transforming all words to lowercase since having terms in different cases can provide a different impression when provided such data to be trained in neural networks. Though those terms provides the same semantic and syntactic meaning in a graph with such data, the terms will be considered as two different terms in close proximity. To overcome such an issue the dataset needs to transformed in one of the case, and the most widely used case in such scenarios is lowercase.

8.0 ANTICIPATED BENEFITS

Having the research component of Corpus Preprocessing in the Adaptive Artificial Intelligent QA Platform is vital. Since the preprocessed corpus is the input for the training neural network model/component. Further the answer generation component/model depends on the preprocessed corpus to extract the answers. Thus to provide the users with the most accurate and a benefitted answer the corpus needs to be well preprocessed. So that the user will be highly benefitted with the answer than misleading with an incorrect answer.

Also if the corpus is to be trained without preprocessing via training neural network model, the outcome of that model will not be the expected and will never provide an accurate result. Further if the corpus is not to be preprocessed, answer extracted for the question asked can provide an inaccurate result to the user. Hence it proves that the corpus preprocessing is one of the major component in Adaptive Artificial Intelligent QA Platform and it will benefit the end user immensely in providing the accurate result.

By taking place of this research component of Corpus Preprocessing will add up as a contribution to the body of knowledge under deep neural networks. As the research carrying out currently on this component, have understood the major drawbacks of the existing corpus preprocessing techniques as explained in the Literature Review chapter. So this research is focussed on overcoming such drawbacks and coming up with a good corpus preprocessing technique with the use of existing techniques and algorithms along with new enhancements.

So that this research component will provide a benefit for the body of knowledge where as the team has an idea to publish a research paper on Adaptive Artificial Intelligent QA Platform once the research is completed. In the research paper to be published, this specific component will also be described. Therefore anyone who are referring in the context of corpus preprocessing can gain knowledge through that paper to be published hence it will go as a contribution made to the body of knowledge.

9.0 SCOPE AND EXPECTED RESEARCH OUTCOME

The chosen scope to build a question answer system for the **medical emergency domain.** As a case study and proof of concept we plan to implement a QA System that focuses only on answering questions related to medical emergency situations and not addressing open domain questions. The scope has been thus narrowed to, first, increase the accuracy of answers provided and second, to ensure that the project can be completed in the allocated time. 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.

Hence the scope is on providing answers to questions related to medical emergency situations, the corpus preprocessing will be done to the dataset relating to medical emergency domain. The outcome of this research component can be broken down into two parts.

The outcome of the research component once the development is done will be a well preprocessed corpus, which can be used as an input to the training neural network model. Also the outcome will be used in extracting the relevant answer for the questions asked by the end users which then can be converted into logical answers.

The research made on this area and the end result along with the development made will be shared via a research paper or any other media where after this will be available in the internet and any other researchers who are focussing on such a component can use this research already made upon this component as a reference or as a comparison.

10.0 RESEARCH CONSTRAINTS

Major research constraint of this component is the dataset. 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 will depends on the dataset. Further the corpus preprocessing component also can be carried out only if the needed dataset is available.

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 the college research purpose has been forwarded to the relevant department. If any delays made, dataset available in the UCI Machine Learning Repository [20] will be used which contains more than 300 datasets.

11.0 PROJECT PLAN OR SCHEDULE

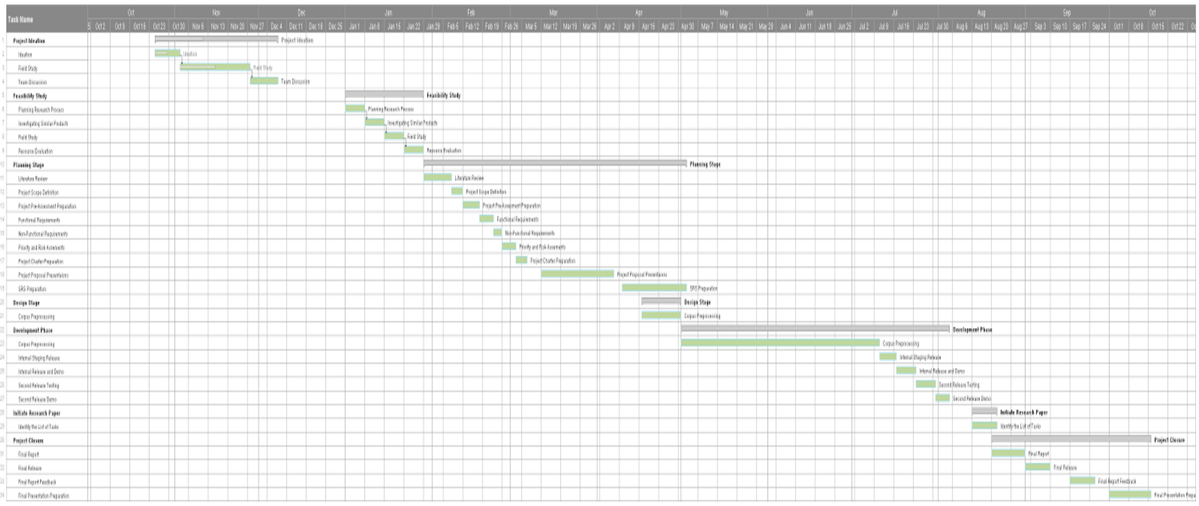


Figure 3.0 Gantt Chart of Corpus Preprocessing

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] "UCI Machine Learning Repository: Data Sets", Archive.ics.uci.edu, 2017. [Online]. Available: http://archive.ics.uci.edu/ml/datasets.html?sort=nameUp&view=list.

[7] Vishal Gupta and Gurpreet S. Lehal, A Survey of Text Mining Techniques and Applications, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009.

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

[9] Saleh Alsaleem, Automated Arabic Text Categorization Using SVM and NB, International Arab Journal of e-Technology, Vol. 2, No. 2, June 2011.

[10] E. Stroh and P. Mathur, “Question Answering Using Deep Learning”, Stanford Reports

[11] "An overview of word embeddings and their connection to distributional semantic models - AYLIEN", AYLIEN. [Online]. Available: http://blog.aylien.com/overview-word-embeddings-history-word2vec-cbow-glove/.

[12] Corrado, Greg, and Jeffrey Dean. "Distributed Representations Of Words And Phrases And Their Compositionality". N.p., 2017.

[13] Y. Zhang, R. Jin and Z. Zhou, "Understanding Bag-of-Words Model: A Statistical Framework". [Online]. Available: https://ai2-s2-pdfs.s3.amazonaws.com/4eb6/00aa4071b9a73da49e5374d6e22ca46eaba6.pdf.

[14] A. Colyer, "The amazing power of word vectors", *the morning paper*, 2016. [Online]. Available: https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/.

[15] J. Collis, "What-is-one-hot-encoding-and-when-is-it-used-in-data-science", *https://www.quora.com*. [Online]. Available: https://www.quora.com/What-is-one-hot-encoding-and-when-is-it-used-in-data-science.

[16] A. Colyer, "GloVe: Global Vectors for Word Representation", *the morning paper*. [Online]. Available: https://blog.acolyer.org/2016/04/22/glove-global-vectors-for-word-representation/.

[17] S. Gouws, "How-is-GloVe-different-from-word2vec", *https://www.quora.com*. [Online]. Available: https://www.quora.com/What-is-word-embedding-in-deep-learning.

[18]"An overview of word embeddings and their connection to distributional semantic models - AYLIEN", *AYLIEN*. [Online]. Available: http://blog.aylien.com/overview-word-embeddings-history-word2vec-cbow-glove/.

[19] "Google Code Archive - Long-term storage for Google Code Project Hosting.", *Code.google.com*. [Online]. Available: https://code.google.com/p/word2vec/.

[20] "UCI Machine Learning Repository: Data Sets", Archive.ics.uci.edu, 2017. [Online]. Available: http://archive.ics.uci.edu/ml/datasets.html?sort=nameUp&view=list.

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

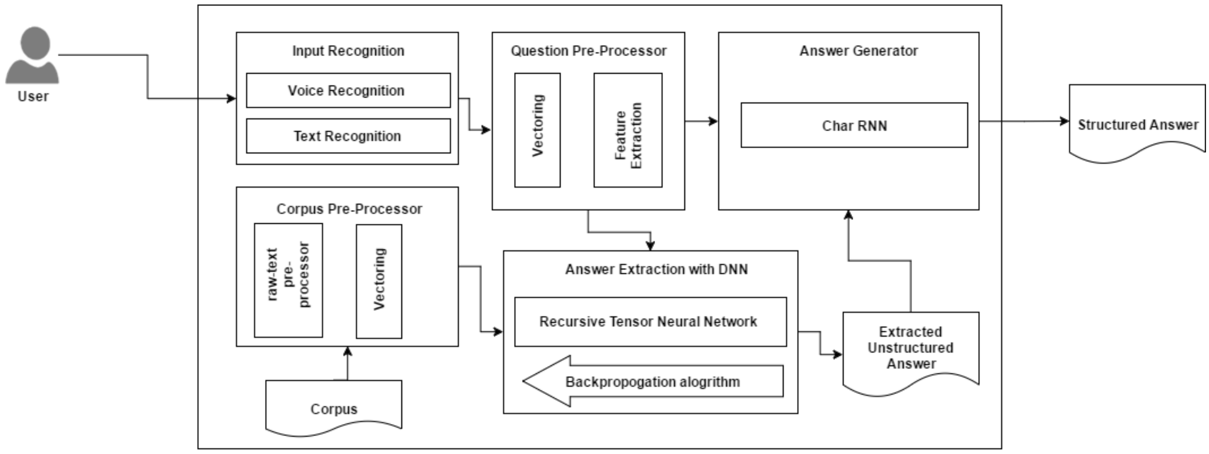
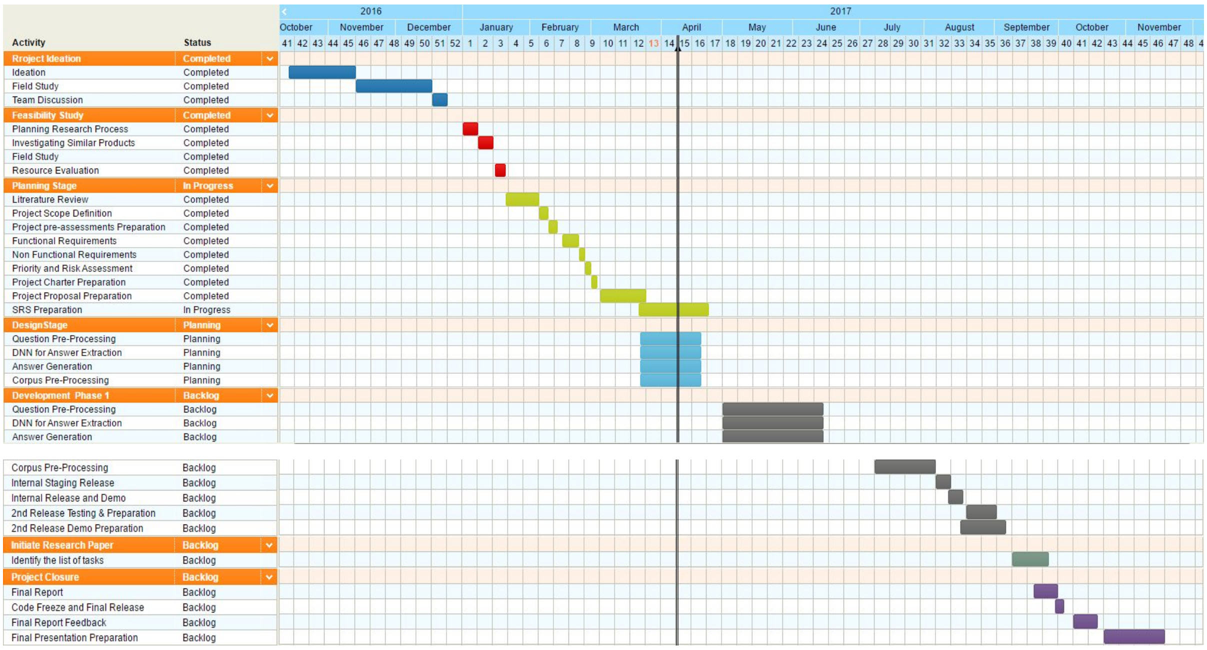


Figure 4.0 System Architecture of Adaptive Artificial Intelligent Question Answer

APPENDIX III: GANTT CHART OF THE OVERALL SYSTEM



Activity	Status
Rroject Ideation	Completed
Ideation	Completed
Field Study	Completed
Team Discussion	Completed
Feasibility Study	Completed
Planning Research Process	Completed
Investigating Similar Products	Completed
Field Study	Completed
Resource Evaluation	Completed
Planning Stage	In Progress
Litrerature Review	Completed
Project Scope Definition	Completed
Project pre-assessments Preparation	Completed
Functional Requirements	Completed
Non Functional Requirements	Completed
Priority and Risk Assessment	Completed
Project Charter Preparation	Completed
Project Proposal Preparation	Completed
SRS Preparation	In Progress
DesignStage	Planning
Question Pre-Processing	Planning
DNN for Answer Extraction	Planning
Answer Generation	Planning
Corpus Pre-Processing	Planning
Development Phase 1	Backlog
Question Pre-Processing	Backlog
DNN for Answer Extraction	Backlog
Answer Generation	Backlog
Corpus Pre-Processing	Backlog
Internal Staging Release	Backlog
Internal Release and Demo	Backlog
2nd Release Testing & Preparation	Backlog
2nd Release Demo Preparation	Backlog
Initiate Research Paper	Backlog
Identify the list of tasks	Backlog
Project Closure	Backlog
Final Report	Backlog
Code Freeze and Final Release	Backlog
Final Report Feedback	Backlog
Final Presentation Preparation	Backlog

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:
Saad M.S.M.