Deep Neural Network For Answer Extraction

ADAPTIVE ARTIFICIAL INTELLIGENT QUESTION ANSWER SYSTEM

17-107

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

October 2017

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. Also I hereby grant to Sri Lanka Institute of Information Technology the non-exclusive right to reproduce and distribute my dissertation in whole or part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as article or books).

Signature: Date:

Signature of the Supervisor: Date:

ABSTRACTION

This final report covers the “Deep Neural Network For Answer Extraction” component of “Adaptive Artificial Intelligent QA Platform”. The goal of this module is to build a neural network which can adapt to any type of textual data given in a predefined format. Furthermore it should be able to create a model that can successfully receive and predict the output in a vector format. This will be fed to answer generation component to produce an answer that can be understood by humans.

The goal of this project is to build a Question Answer system that is capable of processing the information in a large dataset and allows the user to gain knowledge from this dataset by asking questions in natural language form. The system is capable of understanding this question responds to the user’s query in natural language form as well. The goal is to make the user feel as if they were interacting with a person.

ACKNOWLEDGEMENT

I would like to thank my supervisor Mr. Yasas Mallawarachi, head of Software Engineering Services at Sri Lanka Institute of Information Technology for his guidance and for helping me putting this project together.

Furthermore I would like to thank my external supervisor PhD student Anupiya Nugaliyadde for his guidance, help and for always being available through the project process.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF ABBREVIATIONS

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

1.0 INTRODUCTION

The information growth rate in the world is increasing at a rapid rate. It has become impossible to keep up with the amount of information that is being added to any given domain. A regular human cannot keep up with all new that gets generated. Due to the wide availability of digital input and output devices and the ease of use of these devices people are creating more and more raw data. If we are able to process this data into meaningful information fast, it would be possible for people to become more productive and to get the information they want faster.

A key idea in social sciences is that a rational human being makes the best decision when he/she is able to access and use all the available relevant information. However this is not very practical. The best way to access all the available information for a given domain would be the internet, but a human is not able to go through all this information and understand and process it in a timely manner to make a good decision.

To a certain extent search engines have been able to tackle this problem. Search engines have become the center of the internet because that is what we use as the gateway to the internet. Rarely do we actually type in a specific web address. Instead we would type in a query into a search engine and use the information that is provided by the search engine to access the data that we want. Over the years search engines have become better and are using much more powerful techniques than simple keyword matching and page link ranking. Still search engines only provide us resources through which we have to sort through and find the answers that we need to find. We cannot use them to give us a direct answer.

We thought that this situation can be improved much more and by using deep learning techniques we created 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 is 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 means that we would not need to have a specific query language or we would not need to structure the corpus (dataset) manually.

1.1 BACKGROUND LITERATURE

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.

1.2 RESEARCH GAP

As explained in the introduction, 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 three different techniques that have been used to tackle this problem, NLP, Information Retrieval and Template based question answering. All three methods have 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 an important problem domain due to the high costs there haven't been any commercially viable solutions yet.

The solution that we proposed for this problem domain is one that is based on deep learning. Deep learning can be defined a subset of machine learning techniques that uses non-linear information processing to identify and extract features and patterns in data, classification and transformations. There are three key reasons that deep learning has become so popular in the recent past. First, the hugely increased processing abilities and availability of general purpose GPU's, the vast amount of training data that has become available and the many advances made in the recent past in the field of deep learning that has made the task of training artificial neural networks more efficient [23].

Deep learning makes extensive use of Artificial Neural Networks (ANN) to complete a given task. ANN's are inspired from the neural networks found in the human brain. The brain consists of an intricate network of neuron cells. Researchers have found success in trying to replicate this structure on silicon chip. Each neuron would consist of what is known as an activation function and the neurons would be connected to each other via connections called tensors [24]. The entire ANN would consist of an input layer of neurons, multiple hidden layers and an output layer. Biases are assigned to neurons and weights are assigned to the

tensors. These values influence what eventually becomes the output of the ANN. When training an ANN we will be adjusting these weights and biases to get the desired output for a known input data set and known scenario [25]. When working with a large ANN with several tens of neurons it can become a tedious task to adjust the weights and biases manually. There are algorithms we can use such as backpropagation to help us adjust these values automatically until we get to the desired output.

Since deep neural networks are exemplary at recognising patterns and processing data extremely fast, we believe that by applying deep learning techniques to this problem domain we will be able to overcome many of the drawbacks of the other approaches. We will be able to have a higher accuracy because of the state of the art neural network training paradigms, reduce manual tasks by allowing an ANN to process and structure the corpus and automatically extract the required features and we will not need to use an underlying database engine so therefore we will not need to adopt or develop a different query language.

1.3 RESEARCH PROBLEM

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.

1.4 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.

2.0 METHODOLOGY

2.1 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.

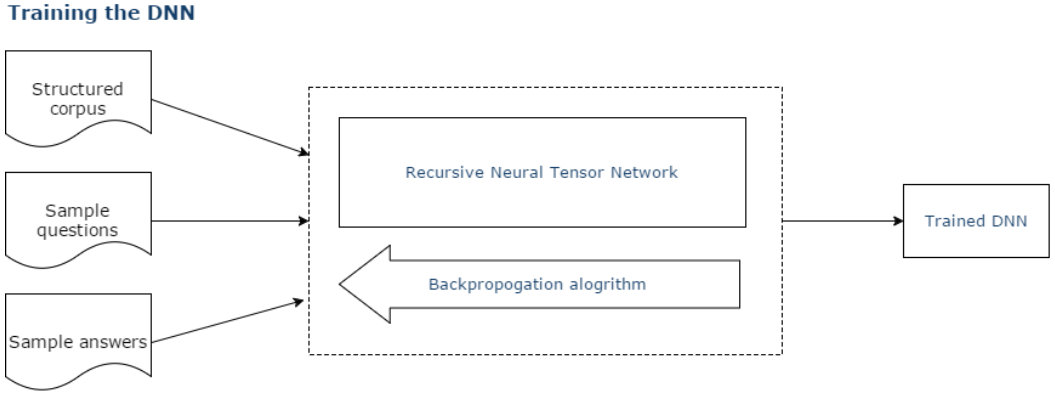

Figure 1.0 Training the RNTN

2.1 TESTING AND IMPLEMENTATION

//TODO

3.0 RESULTS & DISCUSSION

3.1 RESULTS

//TODO

3.2 RESEARCH FINDINGS

This component plays a critical role in our project since it handles the training process of the model and also it acts as a middle tier which is be used by other components to maintain communication between each other.

Neural networks is the main technology behind this module. Here I discuss what we have learned by implementing NN based solution.

3.2.1 THE PROBLEM OF LONG-TERM DEPENDENCIES

One of the appeals of RNNs is the idea that they might be able to connect previous information to the present task, such as using previous video frames might inform the understanding of the present frame. If RNNs could do this, they'd be extremely useful. But can they? It depends.

Sometimes, we only need to look at recent information to perform the present task. For example, consider a language model trying to predict the next word based on the previous ones. If we are trying to predict the last word in "the clouds are in the sky," we don't need any further context – it's pretty obvious the next word is going to be sky. In such cases, where the gap between the relevant information and the place that it's needed is small, RNNs can learn to use the past information.

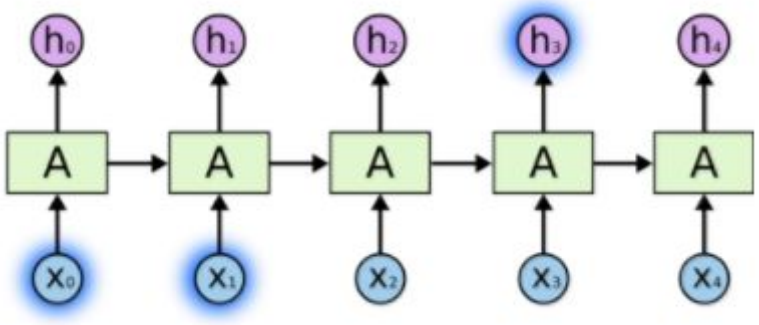

Figure 2.0 RNN

But there are also cases where we need more context. Consider trying to predict the last word in the text "I grew up in France… I speak fluent French." Recent information suggests that the next word is probably the name of a language, but if we want to narrow down which language, we need the context of France, from further back. It's entirely possible for the gap between the relevant information and the point where it is needed to become very large.

Unfortunately, as that gap grows, RNNs become unable to learn to connect the information.

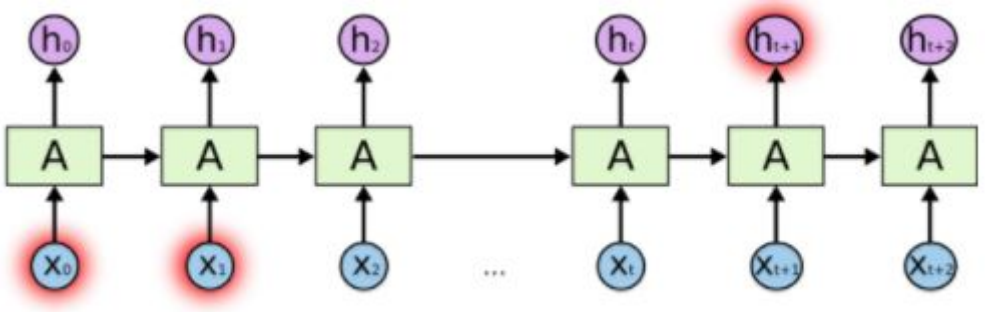

Figure 3.0 Issues with RNN

In theory, RNNs are absolutely capable of handling such "long-term dependencies." A human could carefully pick parameters for them to solve toy problems of this form. Sadly, in practice, RNNs don't seem to be able to learn them. The problem was explored in depth by Hochreiter (1991) [German] and Bengio, et al. (1994), who found some pretty fundamental reasons why it might be difficult.

3.2.2 THE CORE IDEA BEHIND LSTMS

The key to LSTMs is the cell state, the horizontal line running through the top of the diagram. The cell state is kind of like a conveyor belt. It runs straight down the entire chain, with only some minor linear interactions. It's very easy for information to just flow along it unchanged.

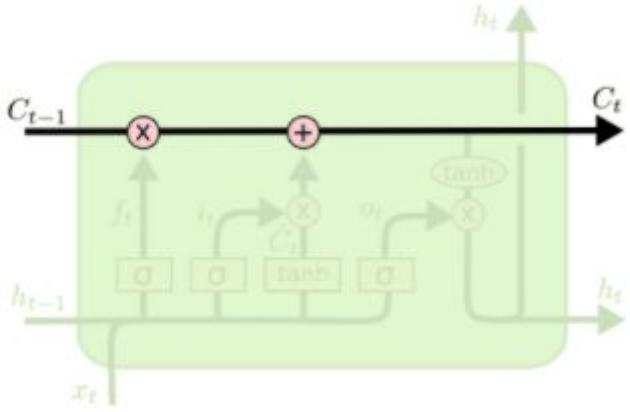

Figure 4.0 LSTM cell

The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates. Gates are a way to optionally let information through. They are composed out of a sigmoid neural net layer and a pointwise multiplication operation.

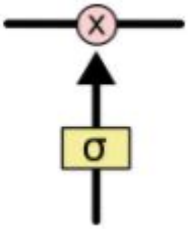

Figure 5.0 Gate Structure

The sigmoid layer outputs numbers between zero and one, describing how much of each component should be let through. A value of zero means "let nothing through," while a value of one means "let everything through!". An LSTM has three of these gates, to protect and control the cell state.

3.3 DISCUSSION

//TODO

4.0 CONCLUSION

//TODO

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.

[21] "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]

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

[23] L. Deng and D. Yu, "Deep Learning: Methods and Applications", Foundations and Trends in Signal Processing, vol. 7, no. 34, pp. 197–387, 2013.

[24] S. Wang, Interdisciplinary computing in Java programming language, 1st ed. Boston, Mass.: Kluwer Academic, 2003.

[25] N. Gupta, "Artificial Neural Network", Network and Complex Systems, vol. 3, no. 1, pp. 24-28, 2013.

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

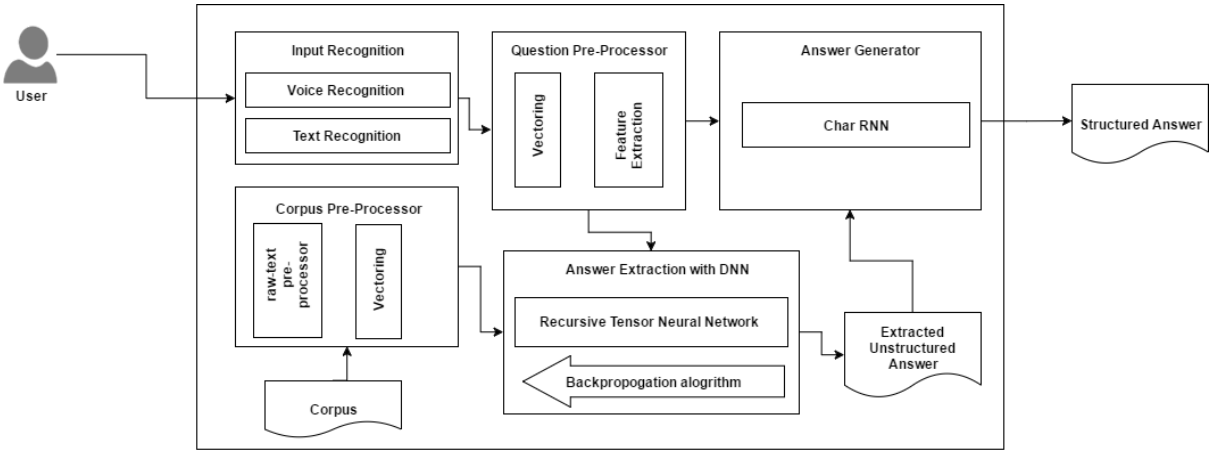

Figure 3.0 System Architecture of Adaptive Artificial Intelligent Question Answer

APPENDIX III: 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

APPENDIX IV: FOCUSSED TASKS

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 PERSONNEL 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 “Deep Neural Network For Answer Extraction” component:

Singhabahu C.P