NEURAL REASONING MODEL FOR QUESTION ANSWERING

18-090

Project Proposal Report

K.S.D.Ishwari (IT15067098)
A.K.R.R.Aneeze (IT15060372)
S.Sudheesan (IT15109668)
H.J.D.A. Karunaratne (IT15047748)

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

Department of Information Technology

Sri Lanka Institute of Information Technology

May 2018

A NEURAL REASONING MODEL FOR QUESTION ANSWERING

18-090

Project Proposal Report

(Proposal documentation submitted in partial fulfilment of the requirement for the Degree of Bachelor of Science Special (Honors) in Information Technology)

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

Department of Information Technology

Sri Lanka Institute of Information Technology

May 2018

TABLE OF CONTENTS

1. INTRODUCTION

Question Answering(QA) is one of the research areas that have gained a lot of interest over time.

Its ability to simulate human interaction through responding to user responses increase the appeal of a product or a system holds for its user base. Out of the various technologies used in QA recently, deep learning stands to be the most promising. Both models of neural networks, feedforward and recurrent have been used in several such researches. Feed-forward neural networks were used extensively for some time but now the tide is turning towards Recurrent neural networks. It's ability to ingest its own inputs through the feedback loop brings past decisions to the equation, increasing the accuracy of the answers provided over time. Since the recursive strategy used by traditional RNN results in eventual generalization, variants such as Long Short Term Memory(LSTM) and Gated Recurrent Units(GRU) are used to maintain distinctions, learning even through large number of steps. Other than these, attention mechanisms and memory networks are used in tandem to improve on memory usage and reasoning capabilities.

Although systems based on the above mentioned deep learning techniques have provided accurate state of the art results, it is still in need of vast improvements. This research is addressing one of those shortcomings by improving the performance of QA using a memory based Neural Network that significantly improves complex reasoning.

This document contains the approaches currently followed in QA systems, the proposed methodology to improve the the RAM model, and also the benefits that can be anticipated.

1.1. Purpose

The purpose of this document is to provide the preliminary progress of the work done towards the “Neural Reasoning Model for Question Answering” research.

This document is primarily intended to be proposed to the audience who has already referred to the proposal of “Neural Reasoning Model for Question Answering” 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 model.

1.2. Scope

The document will illustrate the purpose and the main areas to be focused throughout the component. The literature review section will demonstrate the research gap that this project is attempting to fulfill by analyzing the existing solutions in the QA domain, revealing the advantages and disadvantages of those systems, and how the proposed system can improve on those failings. Furthermore, the research objectives and the proposed methodology will be discussed. Latter part of the document will contain the data sources, anticipated benefits and the anticipated deliverables for this project.

1.3. Overview

The proposed system aims to achieve a remarkable improvement in the Reasoning, Attention and Memory(RAM) Model in Deep Neural Networks, thus achieving effective and accurate answers in the Question Answering system. The main goal of the proposed model is to enhance the performance of the QA systems, thus improving the user experience with Chatbots, Search Engines etc.

2. STATEMENT OF THE WORK

2.1. Background Information and Overview of Previous work based on Literature Survey

Question Answering is a complex and challenging Natural Language Understanding task.

Deep learning fundamentally differs from machine learning because of the ability it has to learn underlying features in data using neural networks. A Neural Network is a computational model which is inspired by the way neurons process information in the human brain[47] .

The basic unit of computation in a neural network is a neuron also referred to as a node or a unit. These neurons receive inputs from an external source or from other nodes and then compute an output. Each input has a weight associated with it that depicts the importance of it compared to other inputs.

Neurons can be classified into layers according to the task they perform.

1. Input layer
2. Hidden layer/s
3. Output layer

Input Layer does not do any computation. It just passes the information to the next layer which is most often a hidden layer.
Hidden layers perform intermediate computations and processing. They also transfer the weights from the input layer to the following layer which is most often another hidden layer.
One of the challenges in creating a neural network is determining the number of hidden layers needed.

Output layer maps the computed inputs from the hidden layer/layers to the desired output. They use an activation function to achieve this.

An activation function is responsible for the definition of an output of a node given its input/inputs. It is an abstract representation of the frequency of the firing rate of a neuron. At its most basic state this function is a binary determining whether the neuron fires or not. This example is a linear state of the activation function. It is the non-linearity of these function that enables the the network to compute complex problems from a relatively small number of nodes.

The accuracy of a neural network relies heavily on how successful the training process is. To achieve this it usually requires a large data-set. In the training process the inputs from the data-set are fed to the neural network and then its outputs are compared with the outputs of the data-set. Once this is done a function can be created that shows how far the network's outputs are deviated from the expected outputs. This function is called a cost function. Ideally the value returned by the cost function should be zero, which means that there is no deviation between the outputs from the network and the expected outputs.
If there is a deviation, in order to reduce it the weights between the neurons should be changed. Instead of randomly altering these weights a technique called Gradient Descent is used for this purpose. This technique is used to find the minimum of the cost function. This is achieved by changing the weights by small increment after each data-set iteration. By computing the gradient of the cost function for a given data-set iteration, the direction of the minimum can be determined. Through numerous iteration the weights are automatically optimized by the neural network which in turn increases the accuracy of it.[48]

There are several classes of neural networks. The one we are focusing on in this research are the Recurrent Neural Networks(RNN).

Recurrent Neural Networks are special because of its ability to take the previously perceived information into consideration. Feed-forward neural networks are only concerned with the state of the inputs at a given time, RNN takes the past decisions into consideration. The decision a recurrent net reached at time step t-1 affects the decision it will reach one moment later at time step t. This is because of the feedback loop that

enables RNNs to ingest their own outputs moment after moment as input retaining memory in the sequence itself.

The sequential information in an RNN can be preserved in the hidden state of the network and it can be carried forward several steps so that it would affect processing new inputs. This enables RNN to find co-relations between inputs that are separated by many moments. As long as the memory of a recurrent net can persist the past decisions of would influence the current decision making.

Backpropagation is essential for a RNN to achieve its purpose. In a feed-forward net backpropagation moves backward from the final error through the outputs, weights and inputs of each hidden layer, assigning those weights responsibility for a portion of the error by calculating their partial derivatives or the relationship between their rates of change. These derivatives are then used by the gradient descent to adjust the weights up or down on whichever direction to reduce the error. RNNs extend this to form a new concept called Backpropagation through time(BPTT). Time here is defined in the form of a well-defined, ordered calculations linking one time step to the next.

One of the major fallbacks of the traditional RNNs is the vanishing gradient problem. These RNNs were not able to form connections between the final outputs and the events that happened several steps before. Because of this the gradient for that particular period cannot be calculated which resulted in the “vanishing gradient” for that period in the graph. One reason for the vanishing gradient is that the information passing through the network passes through several stages of multiplication. If we were to explain this problem using basic mathematics, if a value is to be multiplied by another value more than 1(even slightly) continuously it can become immeasurably large. In the same way if the multiple is less than 1, the value tends to do the inverse. The first example causes Exploding Gradients but this can be solved easily by truncating. But the vanishing gradient caused by the second example is harder to solve because it can become too small for computers to work with.

To solve this problem a variation of RNN came up with Long Short-Term Memory units, or LSTMs. LSTMs preserve a constant error that can be that can be back-propagated through many time and layers, sometimes as many as 1000. This opens up channels to link cause and effect remotely. LSTM achieves this by storing the information outside the normal flow in a gated cell. The cell makes decisions about what to store, and when to allow reads, writes and erasures, via gates that open and close. These are analog gates implemented with element-wise multiplication of sigmoids. Being analog makes the gates differentiable which is an advantage in backpropagation.

These gates act on the signals they recieve and similar to the neurons these block or pass the information based on its importance which is calculated using their own sets of weights. These weights are adjusted during the recurrent net learning process. That is, the cells learn when to allow data to enter, leave or be deleted through the iterative process of making guesses, backpropagating error, and adjusting weights via gradient descent.[49]

RNNs have been able to make a significant improvement in contrast to other machine learning techniques due to their ability to learn and carry out complicated transformations of data its ability maintaining long term as well as short term dependencies. They are said to be 'Turing-Complete'[38], therefore having the capacity to simulate arbitrary procedures. RNN contains an interplay of Reasoning, Attention and Memory, commonly referred to as the 'RAM model' in Deep Neural Networks.

Researches have been done throughout regarding building models of computation with various forms of explicit storage. Google's DeepMind project introduced Neural Turing Machines that extend the capabilities of neural networks by coupling them to external memory resources, which they can interact with by attentional processes. [39] It is shown that NTMs can infer simple algorithms such as copying, sorting, and associative recall from input and output examples. NTMs have demonstrated how to learn a model to sort a small set of numbers as well as a host of other symbolic manipulation tasks. This is done through a large, addressable memory, by differentiating Turing's enrichment of finite-state machines by an infinite memory tape. Short Term memory in NTMs are handled by

resembling a working memory system, designed to solve tasks that require the application of approximate rules to data that are quickly bound to memory slots known as “rapidly-created variables”. Additionally, NTMs use an attentional process to read from and write to memory selectively.

End-to-end Memory Networks introduces a recurrent attention model over a possibly large external memory. [42] Since they are trained end-to-end it requires significantly less supervision during training. This technique is applicable to synthetic question answering and to language modeling. A RNN architecture is presented where recurrence reads from a long term memory multiple times before outputting a symbol. This work evaluates how this is crucial towards maintaining good performance of the model.

Another research that has been done in this area is by Facebook AI research where they have attempted to show that some basic algorithms can be learned from sequential data using a recurrent network associated with a trainable memory. [40] Although, machine learning has progressed over the years, with the scaling up of learning algorithms, alternative hardware such as GPUs or large clusters have been compulsatory. It is not practical with real world applications. This approach has increased the learning capabilities of recurrent nets by allowing them to learn how to control an infinite structured memory.

Neural Training Machines were more expensive than previously considered due to the utilization of an external memory. Thus, Reinforcement Learning Neural Turing Machines (RLNTM) were introduced. [41] RLNTMs use a Reinforcement Learning Algorithm to train Neural Network that interacts with interfaces such as memory tapes, input tapes and output tapes, to solve simple algorithmic tasks. RLNTMs use Reinforce algorithm to learn where to access the discrete interfaces and to use the backpropagation algorithm to determine what to write to the memory and to the output. RLNTMs have succeeded at problems such as copying an input several times to the output tape, reversing a sequence, and a few more tasks of comparable difficulty.

Overtime, different attention models have shown promising results as well. Researches have been done on how complex sequences with long-range structure can be generated with LSTM RNNs.[43]

In addition to that, Neural Machine Translation is another approach to machine translation, which attempts to build and train a single large neural network that reads and outputs a correct translation instead of having small sub-components that are tuned separately.[44] Most machine translators are encoder-decoder models where the encoder neural network reads and encodes a sentence to a fixed length vector and the decoder output the translation from the encoded vector. The major issue with this system is that the neural network should be able to compress all the important information in the sentence to a fixed-length vector. Thus, dealing with longer sentences becomes difficult. The distinguishing feature of this approach is that it does not encode the entire input sentence into a fixed length vector. Instead, it encodes the input into a sequence of vectors and choses a subset of these vectors adaptively while decoding.[44] This removes the potential challenge in dealing with long sentences and assists in retaining the necessary information.

A method has been proposed, utilizing a local attention-based model for Abstractive Sentence Summarization which upto date remains a challenge for Natural Language Processing. This model focuses on sentence-level summarization rather than using extracted potions of the sentences to prepare a condensed version. It uses a neural language model with an input encoder that learns a latent soft alignment over the input text to help inform the summary.

There are attention-based models introduced for Speech Recognition[45], Handwriting synthesis and Image Caption Generation as well.[46]

Although machine learning have achieved significant accomplishments, reasoning tasks remains elusive.

2.2. Identification and Significance of the Problem

Question Answering(QA) is one of the research areas that have gained a lot of interest over time. Out of the various technologies used in QA recently, deep learning stands to be the most prominent technology among the prevailing technologies. This research is addressing the shortcomings of performance and accuracy of QA, by improving and supporting the RAM model. The idea is to build a Question Answering system using Deep learning technique by Improving on the existing state-of-the-art leveraging optimized Deep Learning techniques by combining and improving Reasoning, Attention and Memory mechanisms to assess the system in a real context.

The significance of the research is that we will be improving question preprocessing, Reasoning, Attention and Memory of our deep neural network which are vital and technically challenging tasks to carry out.

Improvement of question preprocessing: Provide the neural network with the most effective vector representation.

Improvement in Reasoning: Improving reasoning by making it capable to infer over multiple facts in a way insensitive to the number of supporting facts and data.

Improvement in Attention: Improving focus on relevant parts of the input more than the irrelevant parts and improvement in finding correspondence between source and target words.

Improvement in Memory: Improvement in the ability of the neural network to retain the most important data inputted and identify the threshold level to retain memory.

2.3. Technical objectives

According to the explanations given in previous parts “NEURAL REASONING MODEL FOR QUESTION ANSWERING” project is based on dnn (Deep Neural Network) where the preprocessing, attention, reasoning and memory layers of dnn are enriched to produce better results than existing state-of-art techniques. This part of the report will demonstrate the Software and Hardware requirements of the research which will be used for performing technical tasks in order to improve the above mentioned for components.

2.3.1 Specific software requirements

The most renowned and accepted programming language for the machine learning problems is Python. Python is widely used across the world for solving machine learning problems for its readability and wide range of machine learning libraries. Our research project will be using Python as the main programming language.

TensorFlow is an open source artificial intelligence library, using data flow graphs to build models. It allows developers to create large-scale neural networks with many layers and TensorFlow allows distribution of computation across different computers, as well as multiple CPUs and GPUs within a single machine. TensorFlow provides a Python API so we will be using TensorFlow for our deep learning research.

Keras is a high-level neural network API, helping lead the way to the commoditization of deep learning and artificial intelligence. It runs on top of a number of lower-level libraries, used as backend, including TensorFlow and Theano. Keras code is portable, meaning that you can implement a neural network in Keras using Theano as a backend and then specify the backend to subsequently run on TensorFlow, and no further changes would be required to your code. Theano is a Python library for fast numerical computation that can be run on the CPU or GPU.It is a key foundational library for Deep Learning in Python that you can use directly to create Deep Learning models or wrapper libraries that greatly simplify the process. And this research will be using Keras and Theano as key software resources.

Overall Software Stack :

- Python
- Tensorflow
- Keras.
- Theano.

2.4. Detail Design

Context Preprocessing

Preprocessing is the initial step in QA which enables question and context understood. The outputs of this components will be responsible for outputs of all other components. Machines do not understand natural language similar to humans. Thus, it is necessary to interpret the text inputted in an accurate machine-understandable form yet preserving the context. Vector Representations have been a popular [1][2].

The main objective of this component would be provide the neural network with the most effective vector representation which the rest of the components can utilize in order to provide the user with the most accurate answer. In order to do so, the Question type, the expected answer type and the question focus should be identified[3]. There multiple several techniques such as Word Embedding, Syntactic Analysis[4] and WH-question analysis to achieve this.[5]

In the proposed approach we will be focusing on Word Embedding. Word Embedding is used to map words or phrases to the corresponding vector. It allows words with similar meaning to have similar representation. It is done through algorithms such as Word2Vec, GloVe. The two learning models introduced to learn the word embedding in Word2Vec are Continuous Bag of Words model which predicts the current word based on its context and Continuous and Continuous Skip-Gram model which predicts the surrounding words given a current word.[7] But in this approach it is proposed to use GloVe algorithm which is extension to Word2Vec. It combines both the global statistics of matrix factorization techniques like Latent Semantic Analysis with the local context-based learning in Word2Vec.[7] The approach followed may change based on trial and error.

Deep Neural Network for Answering

Next component creates a Deep Neural Network which will be used for Information Extraction. In previous related work, this has been done using Knowledge Bases and Ontologies but they lack the ability to reason. Therefore, we present a system which is capable of reasoning with the help of a Deep Neural Network.

The proposed framework will use Reinforced Memory Networks (R-MN) technique which is the current state of the art. [9] This approach combines Memory Networks and Reinforced learning to achieve superior performance than the conventional methods such as Long Short Term Memory (LSTM) and Dynamic Memory Networks (DMN). In this approach, the text sequence and the question is passed to the Memory Network and Reinforced Learning component. The output of the Memory Network will be fed again to the Reinforced Learning component. The mentioned technique is ideal since Memory Networks cannot achieve reasoning on its own and Reinforcement learning has proven to be performing up to super human extents.

Improvement of Reasoning Attention and Memory (RAM) in order to improve performance

Deep learning techniques have shown to outperform humans in computer vision, pattern recognition. [36][37] However, QA has not shown similar achievements or human level performance through deep learning. This shows that deep learning lacks reasoning, attention and memory[11]. The improvements in these areas have led to very promising results [12][13][14] yet they are required to make progress on complex tasks.

The proposed system aims to address each of these areas in order to improve them, thus improving the performance of the Memory based Network.

Reasoning is an area in Machine Learning which is possibly what led to the drastic improvement in performance in contrast to the Ontology-based or Knowledge based Question Answering Systems. Humans are capable of reasoning about entities and relationships intuitively. For example, a human would know when to cross the road

safely when there are multiple vehicles approaching from both sides. However, machines possess this capability up to some extent only. Machines would have to utilize the previously acquired knowledge to do so. This would allow machines to reason about their entities and relations from unstructured data. Solving this would open infinite possibilities for Artificial Intelligence.

In order to improve reasoning, the proposed framework would focus on Memory and Attention which are crucial for reasoning.

Memory is another important aspect in DNNs as the cutting edge Deep Learning models continue to push the limits of GPU RAM. This system aims to develop a desirable model to train more data consuming less memory. The dependencies between the operations in the deep network is represented using Computation Graphs. There are mainly two ways: performing back-propagation on the same graph or explicitly representing a backwards path to calculate the required gradients. In addition, mxnet[25] have introduced explicit backward path for gradient calculation. Also, methods such as In-place memory sharing [25], Standard memory sharing[25] are also proven to optimize memory. Furthermore, the proposed system will consider Static and Dynamic Memory Allocation methods and Parallelization to optimize memory.

Next aspect that affects reasoning would be Attention. Before Attention was introduced, translation relied on reading a complete sentence and compressing all information into a fixed-length vector. Thus, a sentence with hundreds of words represented by several words will surely lead to information loss, inadequate translation. Attention allows machine translator to look over all the information the original sentence holds, then generate the proper word according to current word it works on and the context. MAC network developed by Stanford approaches problems by decomposing them into a series of attention-based reasoning steps. [18]

2.5. Sources for test data & analysis

The amount of data that we create is growing at a staggering pace and keeping track of it becomes more difficult. Accordingly, it is important that we use techniques to make the information that we need accessible. A question answer system is concerned with automatically answering questions posed by humans naturally, which would return the most specific or direct answer rather than returning a set of documents relating like in search engines. There are two broad types of question answering systems that could be built upon. One is Closed-domain question answering system, which basically deals with questions under a specific domain (bAbi – a closed domain dataset provided by Facebook). The other type is Open domain question answering system, which is concerned with question answering about almost any kind of question posed, therefore, it is required to use the information provided in the dataset as well as any additional available knowledge(WIKIQA - an open domain dataset). However, for the purpose of the research, we have settled our center of interest towards an open domain question answering system. This does not necessary mean that it will answer every open domain question. We would focus on answering general questions with an enhanced Deep Neural Network than state-of-the-art systems.

We plan to build a QA system using deep learning to help interpret a question posed and provide a suitable answer from a given context. We will facilitate the QA system to be built under any domain, complying with certain conditions to increase performance. The precision and the weight of reliability of the provided answers will majorly depend upon the accuracy of the context. However the goal in this research is to emphasize that by using deep learning techniques by combining Reasoning, Attention and Memory mechanisms, we are able to reduce some of the complexities and barriers that are present at the moment and provide an innovative product that utilizes the platform.

2.6. Anticipated Benefits

As we demonstrated above the latest improvements in deep learning methodologies made significant impact in natural language QAS for producing human like answers for a given question. This lead to creation of many QAS in open domain and closed domain with their significant changes and specialties. Through “NEURAL REASONING MODEL FOR QUESTION ANSWERING” research we tend to improve four major components of a dnn which are preprocessing, attention, reasoning and memory in order to produce a human like answers for a given question captivating state-of-art techniques. So that the user will be highly benefited with the answer than misleading with an incorrect answer. This will make user more enthusiastic to us our QAS system more and more.

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 focused 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.

By using attention mechanism on the facts, we refine the internal representation of facts based on the information (relevance) from the question. By aggregating relevant information into the final representation, we provide necessary information to the answer selection layer, to predict the answer to the question. This form of multi-hop “search” (on facts) allows the model to learn to perform some sophisticated reasoning for solving certain challenging tasks. This will ignore the irrelevant parts from the answer automatically so he answers will be more related to the question.

Humans are capable of reasoning about entities and relationships intuitively. However, machines possess this capability up to some extent only. Machines would have to utilize the previously acquired knowledge to do so. This would allow machines to reason about

their entities and relations from unstructured data. Solving this would open infinite possibilities for Artificial Intelligence. Through the reasoning the model will gain artificial intelligent which makes answers more human interactive.

Apart from technical benefits identifying an approach that deviates from current approaches and if the evaluation of the approach proves to have outperformed current state-of-art approaches this enables future work in the Intelligent Question and Answering domain to adapt and evolve the improvement of RAM model we have used.

3. 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.

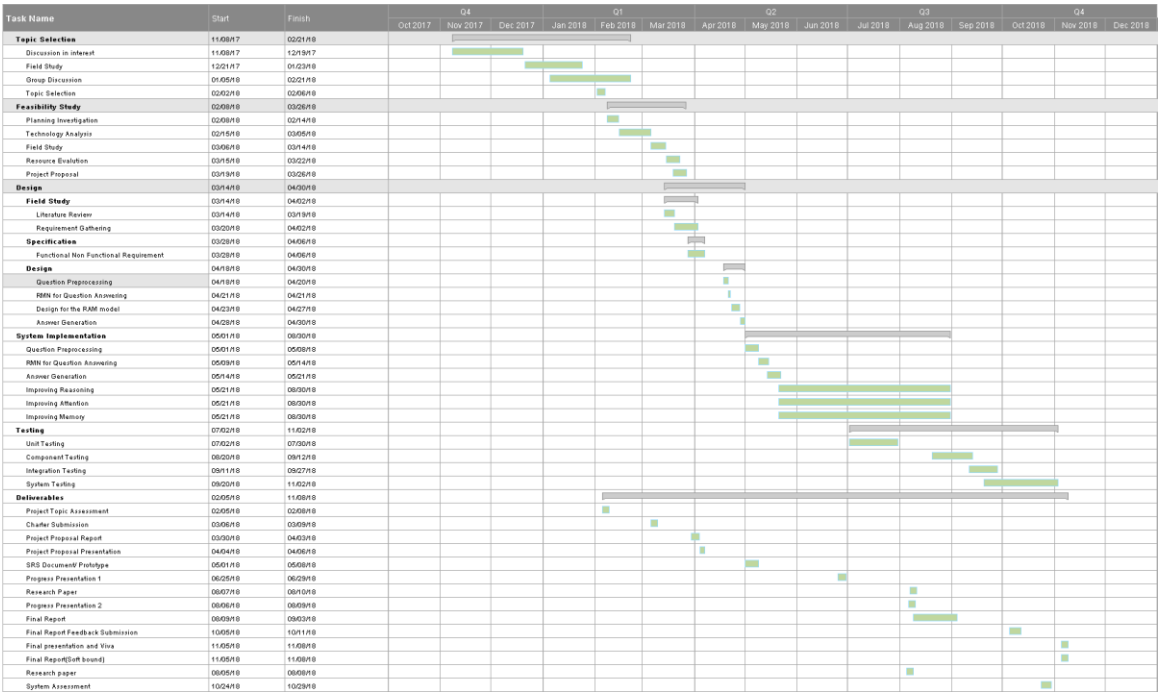

Figure 3.0 Gantt Chart of the Component

4. RESEARCH CONSTRAINTS

One of the main constraints that we have to face is the suitability of the dataset for our model. Although, we are planning to use the TriviaQA[29] dataset initially, we may have to change the dataset based on Trial and Error.

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.

Another challenge is that there must be continuous monitoring of the training process to avoid overfitting the model to the given dataset. The challenge is to balance overfitting and underfitting when training the model. This can be challenging sometimes with a neural net that is as complicated as RNN's.

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.

5. SPECIFIED DELIVERABLES

As the main outcome of the project we aim to deliver a learning domain specific QA system with better performance through enhanced reasoning, memory and attention.

6. REFERENCES

[1] Aclweb.org, 2018. [Online]. Available: http://www.aclweb.org/anthology/P08-1028. [Accessed: 01- Apr- 2018].

[2] Homepages.inf.ed.ac.uk, 2018. [Online]. Available: http://homepages.inf.ed.ac.uk/s0453356/composition.pdf. [Accessed: 01- Apr- 2018].

[3] A. Ben Abacha and P. Zweigenbaum, "MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies", 2018. .

[4] Linguistics.ucla.edu, 2018. [Online]. Available: http://linguistics.ucla.edu/people/stabler/isat.pdf. [Accessed: 01- Apr- 2018].

[5] Aclweb.org, 2018. [Online]. Available: http://www.aclweb.org/anthology/D15-1237. [Accessed: 01- Apr- 2018].

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

[7] J. Brownlee, "What Are Word Embeddings for Text? - Machine Learning Mastery", Machine Learning Mastery, 2018. [Online]. Available: https://machinelearningmastery.com/what-are-word-embeddings/. [Accessed: 01- Apr- 2018].

[8] Arxiv.org, 2018. [Online]. Available: https://arxiv.org/pdf/1703.04617.pdf. [Accessed: 01- Apr- 2018].

[9] 2018. [Online]. Available: https://www.researchgate.net/publication/320658588_Reinforced_Memory_Network_for_Question_Answering. [Accessed: 01- Apr- 2018].

[10] J. Hanlon and J. Hanlon, "How To Solve The Memory Challenges Of Deep Neural Networks - TOPBOTS", TOPBOTS, 2018. [Online]. Available: https://www.topbots.com/how-solve-memory-challenges-deep-learning-neural-networks-graphcore/. [Accessed: 01- Apr- 2018].

[11] "RAM Workshop", Thespermwhale.com, 2018. [Online]. Available: http://www.thespermwhale.com/jaseweston/ram/. [Accessed: 01- Apr- 2018].

[12] Memory Networks. Jason Weston, Sumit Chopra, Antoine Bordes. International Conference on Representation Learning, 2015

[13] Teaching Machines to Read and Comprehend. Karl Moritz Hermann et. al. arXiv Pre-Print, 2015.

[14] Large-scale Simple Question Answering with Memory Networks. Antoine Bordes, Nicolas Usunier, Sumit Chopra, Jason Weston. arXiv Pre-Print, 2015.

[15] L. Bottou, "From machine learning to machine reasoning", Machine Learning, vol. 94, no. 2, pp. 133-149, 2013.

[16] Arxiv.org, 2018. [Online]. Available: https://arxiv.org/pdf/1410.3916.pdf. [Accessed: 01- Apr- 2018].

[17] Proceedings.mlr.press, 2018. [Online]. Available: http://proceedings.mlr.press/v48/kumar16.pdf. [Accessed: 01- Apr- 2018].

[18] Arxiv.org, 2018. [Online]. Available: https://arxiv.org/pdf/1803.03067.pdf. [Accessed: 01- Apr- 2018].

[19] "How does an attention mechanism work in deep learning? - Quora", Quora.com, 2018. [Online]. Available: https://www.quora.com/How-does-an-attention-mechanism-work-in-deep-learning. [Accessed: 01- Apr- 2018].

[20] "How does attention model work using LSTM? - Quora", Quora.com, 2018. [Online]. Available: https://www.quora.com/How-does-attention-model-work-using-LSTM. [Accessed: 01- Apr- 2018].

[21] Aclweb.org, 2018. [Online]. Available: http://www.aclweb.org/anthology/C16-1290. [Accessed: 01- Apr- 2018].

[22] S. Dwivedi and V. Singh, "Research and Reviews in Question Answering System", 2018. .

[23] "Machine Learning for Question Answering | Machine Learning Group", Mlg.ulb.ac.be, 2018. [Online]. Available: http://mlg.ulb.ac.be/questionanswering. [Accessed: 01- Apr- 2018].

[24] "Intro to Deep Learning for Question Answering", Slideshare.net, 2018. [Online]. Available: https://www.slideshare.net/TraianRebedea/intro-to-deep-learning-for-auestion-answering. [Accessed: 01- Apr- 2018].

[25] "Optimizing Memory Consumption in Deep Learning — mxnet documentation", Mxnet.incubator.apache.org, 2018. [Online]. Available: https://mxnet.incubator.apache.org/architecture/note_memory.html. [Accessed: 01- Apr- 2018].

[26] Aclweb.org, 2018. [Online]. Available: http://www.aclweb.org/anthology/P15-2116. [Accessed: 01- Apr- 2018].

[27] Thespermwhale.com, 2018. [Online]. Available: http://www.thespermwhale.com/jaseweston/ram/papers/paper_18.pdf. [Accessed: 01- Apr- 2018].

[28] "A Brief Overview of Attention Mechanism – Synced – Medium", Medium, 2018. [Online]. Available: https://medium.com/@Synced/a-brief-overview-of-attention-mechanism-13c578ba9129. [Accessed: 01- Apr- 2018].

[29] "TriviaQA", Nlp.cs.washington.edu, 2018. [Online]. Available: http://nlp.cs.washington.edu/triviaqa/. [Accessed: 01- Apr- 2018].

[30] J. Bian, B. Gao and T. Liu, "Knowledge-Powered Deep Learning for Word Embedding", 2018. .

[31] S. Dwivedi and V. Singh, "Research and Reviews in Question Answering System", 2018.

[32] "Understanding LSTM Networks -- colah's blog", Colah.github.io, 2018. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed: 01- Apr- 2018].

[33] Cs224d.stanford.edu, 2018. [Online]. Available: https://cs224d.stanford.edu/reports/StrohMathur.pdf. [Accessed: 01- Apr- 2018].

[34] J. Hanlon and J. Hanlon, "How To Solve The Memory Challenges Of Deep Neural Networks - TOPBOTS", TOPBOTS, 2018. [Online]. Available: https://www.topbots.com/how-solve-memory-challenges-deep-learning-neural-networks-graphcore/. [Accessed: 01- Apr- 2018].

[35] Krizhevsky, A., Sutskever, I. & Hinton, “G. ImageNet classification with deep convolutional neural networks. In Proc. Advances in Neural Information Processing Systems”, 2012. [Accessed: 03- Apr- 2018]

[36] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Hamburg, 2015, pp. 681-687. [Accessed: 03- Apr- 2018]

[37] WildML. (2018). *Attention and Memory in Deep Learning and NLP*. [online]Available at: http://www.wildml.com/2016/01/attention-and-memory-in-deep-learning-and-nlp/ [Accessed 4 Apr. 2018].

[38] Siegelmann, H. T. and Sontag, E. D. (1995). “On the computational power of neural nets”. Journal of computer and system sciences.

[39] Alex Graves, Greg Wayne, Ivo Danihelka. (2014). ‘Neural Turing Machines”. [Online]. Available: https://arxiv.org/pdf/1410.5401v2.pdf. [Accessed 4 Apr. 2018].

[40] Armand Joulin Facebook, Tomas Mikolov (2015). “Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets”. [Online]. Available: https://arxiv.org/pdf/1503.01007.pdf. [Accessed 4 Apr. 2018].

[41] Wojciech Zaremba, Ilya Sutskever (2016). “REINFORCEMENT LEARNING NEURAL TURING MACHINES”. [Online]. Available: https://arxiv.org/pdf/1505.00521.pdf. [Accessed 4 Apr. 2018].

[42] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus (2015). “End-To-End Memory Networks”. [Online]. Available: https://arxiv.org/pdf/1503.08895.pdf. [Accessed 4 Apr. 2018].

[43] Alex Graves (2014). “Generating Sequences With Recurrent Neural Networks”. [Online]. Available: https://arxiv.org/pdf/1308.0850.pdf. [Accessed 4 Apr. 2018].

[44] Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio. (2016) “NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND

TRANSLATE". [Online]. Available: https://arxiv.org/pdf/1308.0850.pdf. [Accessed 4 Apr. 2018].

[45] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Yoshua Bengio. (2015) "Attention-Based Models for Speech Recognition". [Online]. Available: https://arxiv.org/pdf/1308.0850.pdf. [Accessed 4 Apr. 2018].

[46] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho , Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio. (2016) "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". [Online]. Available: https://arxiv.org/pdf/1502.03044.pdf . [Accessed 4 Apr. 2018].

[47] Towards Data Science. (2018). *A Gentle Introduction To Neural Networks Series — Part 1*. [online] Available at: https://towardsdatascience.com/a-gentle-introduction-to-neural-networks-series-part-1-2b90b87795bc [Accessed 13 May 2018].

[48] freeCodeCamp. (2018). *Want to know how Deep Learning works? Here's a quick guide for everyone.*. [online] Available at: https://medium.freecodecamp.org/want-to-know-how-deep-learning-works-heres-a-quick-guide-for-everyone-1aedeca88076 [Accessed 13 May 2018].

[49] Chris V. Nicholson, S. (2018). *A Beginner's Guide to Recurrent Networks and LSTMs - Deeplearning4j: Open-source, Distributed Deep Learning for the JVM*. [online] Deeplearning4j.org. Available at: https://deeplearning4j.org/lstm.html#recurrent [Accessed 13 May 2018].

APPENDIX I: DESCRIPTION OF PERSONAL AND FACILITIES

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

Supervisor: Mr. Yashas Mallawarachchi

Co-supervisor: Mr. Anupiya Nugaliyadde

Implementation team:

K.S.D.Ishwari

A.K.R.R.Aneeze

S.Sudheesan

H.J.D.A. Karunaratne