IMPROVING ATTENTION
NEURAL REASONING MODEL FOR QUESTION ANSWERING

18-090

Preliminary Progress Review

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

S.Sudheesan (IT15109668)

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

Department of Information Technology

Sri Lanka Institute of Information Technology

May 2018

TABLE OF CONTENT

1. INTRODUCTION

Question Answering (QA) systems has been widely popular for the past few decades. Although it was initially implemented using pure Natural Language Processing techniques, a major challenge was to allow the system to think similar to human perception. Deep Neural Networks has been proven to overcome this challenge mainly due to its ability to reason. Reasoning can be defined as the ability to reason between entities and relationships. Improving reasoning would make the system think closer to how a human would and it would further open infinite possibilities for Artificial Intelligence.

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

1.1 Purpose

The purpose of this document is to provide the initial progress of the individual component of "NEURAL REASONING MODEL FOR QUESTION ANSWERING". The document will demonstrate the purpose and the main areas to be focused throughout the component. This document will describe the how a selected component is significant to achieve anticipated outcome and what kind to impact that attention mechanism would make This document is primarily envisioned to be projected to the audience who has already referred to the proposal of "NEURAL REASONING MODEL FOR QUESTION ANSWERING" and also to anyone who has knowledge on deep neural networks. The document will exemplify identity and significant, technical objectives, technical approaches train data analysis, anticipated benefits and constraints of the component.

1.2 Scope

This document tends to cover aspects of “Improving Attention” component of “NEURAL REASONING MODEL FOR QUESTION ANSWERING”. Our research area contains for four major components Data Preprocessing, Improvement of Attention, Improvement of Reasoning and Improvement of Memory, among those individual component this document conceals the main characteristics of Improvement of Attention in a QAS to achieve a human level of intelligence in answer prediction task.

This paper will describe the Attention Mechanism’s contribution to use to uplift the overall performance of an intelligent QAS by overcoming current state-of-art techniques. And also the document will emphasize how attention mechanisms are different from other state of art techniques which are used in existing intelligent question and answering systems.

1.3 Overview

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 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. Throughout our research we are proposing a “NEURAL REASONING MODEL FOR QUESTION ANSWERING” which is Deep learning based model. The research contains four major components which is context preprocessing, improvement in reasoning; improvement in attention and improvement in memory out of these four components this paper tends to cover the aspects of improvements of attention.

Even though Deep Neural Networks is the state-of-art technique used to tackle difficult sequence prediction models it suffers from a limitation that all in the input sequences are forced to be encoded in to a fixed length vector. Attention is the mechanism which is used in Encoder Decoder model that identifies which parts of the input sequence are relevant to predict output, whereas prediction is the process of using the relevant information to select the appropriate output. Utilizing attention mechanisms in a QAS system will make a significant impact on the intelligence of the answers for the questions asked by the user.

Using proper attention mechanism in question and answering will train the model to select only the relevant parts of the question which is important to generate the answer and ignore the rest.

The main goal of this research component is to improve attention in our proposed model in order predict answers more intelligently like humans.

2. Statement of the work

2.1 Background information and overview of previous work based on literature survey

Automated QA systems have been gaining a lot of prominence since the early 60’s. QA has mostly been used to develop intricate dialogue systems such as chat-bots and other systems that mimic human interaction [1]. The methodologies used in these systems vary from Information retrieval based statistical approaches to machine learning approaches and then to deep learning based evolving systems.

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. [1] 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 [2]. 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 [3].

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 [4]. 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 [5]. 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 [6]. 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.

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.

Recurrent Neural Networks are special because of its ability to take the previously perceived information into consideration. While its counterpart 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.

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.

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. RNN contains an interplay of Reasoning, Attention and Memory, commonly referred to as the 'RAM model' in Deep Neural Networks.

Humans are capable of understanding what is the information that they have to consider in order to answer correctly for a question asked by another person machines are lack of these capability. Injecting attention into a question and answering system will make the model answer in such a way that human does.

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. [7]

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.[8] 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.[8] 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 up to 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[9], Handwriting synthesis and Image Caption Generation as well.[10]

2.2. Identification and significance of the problem

There is a wealth of information on the internet. Search engines present a ranked list of relevant documents in response to users' formulated keywords based on various aspects such as popularity measures, keyword matching, frequencies of accessing documents, etc. However, they do not truly accomplish the task of information retrieval as users have to examine each document one by one for getting the desired information; it makes information retrieval a time consuming process. A system which can respond to natural languages question may tackle this problem involved in information retrieval mediums such Search engines. The latest improvements in deep learning methodologies made significant impact in natural language QAs for producing human like answers for a given question.

Humans are capable of understanding what is the information that they have to consider in order to answer correctly for a question asked by another person machines are lack of these capability. Lack of this capability causes singled fixed length problem in Recurrent Neural Networks (Encoder Decoder Model). This means all the input sequences are forced to be encoded into a single fixed length vector. The problem causes frequently when the length input sequence inputted to the model is larger than the length of the training sequence. It seems somewhat unreasonable to assume that we can encode all information about a potentially very long sentence into a single vector and then have the decoder produce a good translation based on only that. Let's say your question is 50 words long. The first word of the English translation is probably highly correlated with the first word of the source sentence. But that means decoder has to consider information from 50 steps ago, and that information needs to be somehow encoded in the vector. Recurrent Neural Networks are known to have problems dealing with such long-range dependencies. In theory, architectures like LSTMs should be able to deal with this, but in practice long-range dependencies are still problematic. With an attention mechanism we no longer try encode the full source sentence into a fixed-length vector. Rather, we allow the decoder to "attend" to different parts of the source sentence at each step of the output generation.

2.3. Technical objectives (specify s/w and h/w requirements)

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:

- Ø Pyhton
- Ø Tensorflow
- Ø Keras.
- Ø Theano.

2.4. Detail design (Technical approach)

Answer selection is a difficult task, as typically there is a large number of possible answers which need to be examined. Furthermore, although in many cases the correct answer is lexically similar to the question, in other cases semantic similarities between words must be learned in order to find the correct answer [11, 12]. Additionally, many of the words in the answer may not be relevant to the question.

Consider, for example, the following question answer pair:
How do I freeze my account? Hello, hope you are having a great day. You can freeze your account by logging into our site and pressing the freeze account button. Let me know if you have any further questions regarding the management of your account with us.

Intuitively, the key section which identifies the above answer as correct is “[...] you can freeze your account by [...]”, which represents a small fraction of the entire answer.

Earlier work on answer selection used various techniques, ranging from information retrieval methods [13] and machine learning methods relying on hand-crafted features [14, 15]. Deep learning methods, which have recently shown great success in many domains. We propose a new architecture for question answering and we augment this design with a more sophisticated attention mechanism.

Early RNN designs were based on applying a deep feed forward network at every time step, but struggled to cope with longer sequences due to exploding and diminishing gradients [16]. Other recurrent cells such as the LSTM and GRU cells have been proposed as they alleviate this issue; however, even with such cells, tackling large sequences remains hard. Consider using an LSTM to digest a sequence, and taking the final LSTM state to represent the entire sequence; such a design forces the system to represent the entire sequence using a single LSTM state, which is a very narrow channel, making it difficult for the network to represent all the intricacies of a long sequence [17]. Attention mechanisms allow placing varying amounts of emphasis across the entire sequence [17], making it easier to process long sequences; in QA, we can give different weights to different parts of the answer while aggregating the LSTM outputs along the different time steps:

We extended the basic encoder–decoder by letting a model search for a set of input words, or their annotations computed by an encoder, when generating each target word. This frees the model from having to encode a whole source sentence into a fixed-length vector, and also lets the model focus only on information relevant to the generation of the next target word.

Above image demonstrate a simple attention model. An attention model is a method that takes n arguments y_1 … y_n and a context c. It returns a vector z which is the summary of the y_i focusing on the information linked to context c. If we consider test data consists of the question, answer and the context paragraph the arguments y_1 … y_n will be the question sequence and the context will be the proposed answer. The goal of an attention mechanism is to construct an overall representation of the candidate answer a, which is later compared to the question representation to determine how well the candidate answers the question;

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 a particular domain question answering system.

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 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.

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.

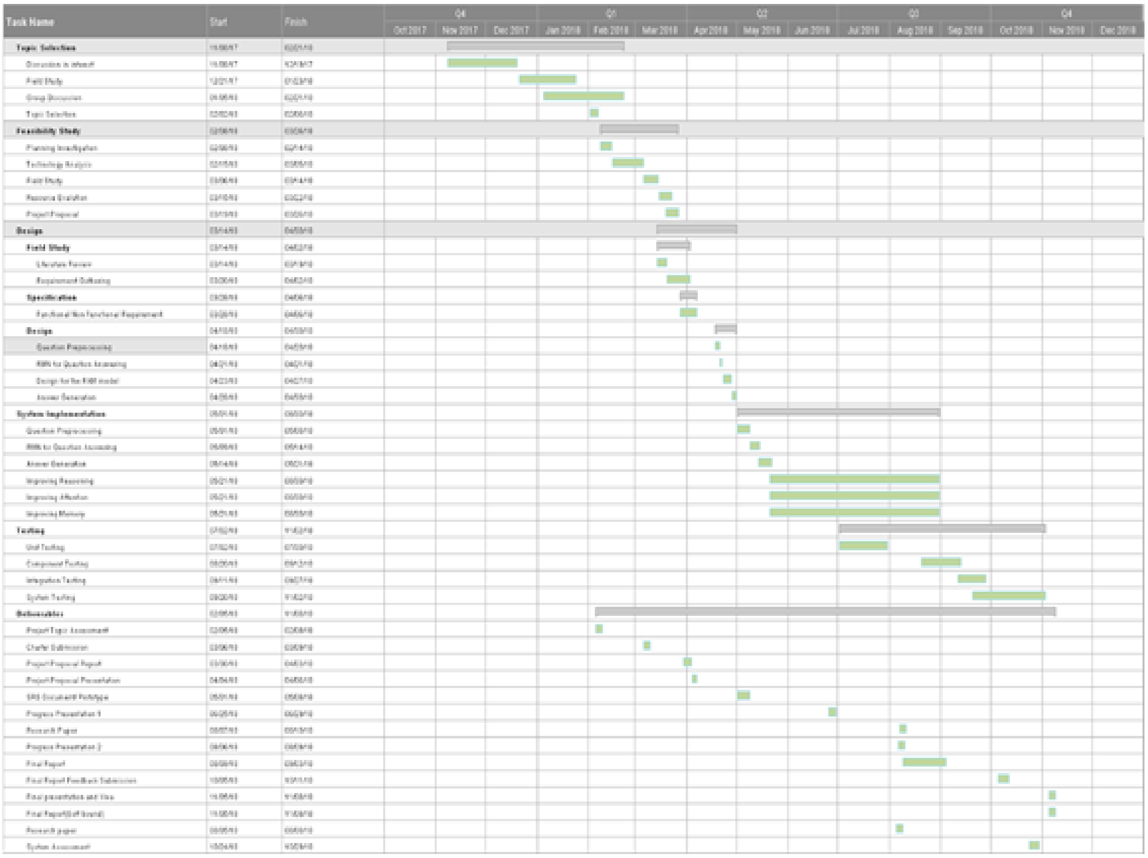

Figure 3.0 Gantt
Chart of the Component

4. RESEARCH CONSTRAINTS

One of the main constraint 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]

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

[2]

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

[3]

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

[4]

"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]

[5]

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

[6]

Y, Bengio, P. Simard, and P. Frasconi, “ Learning long-term dependencies with gradient descent is difficult.” *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[6]

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

[7]

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

[8]
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].

[9]
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].

[10]
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].

[11]
Wtlab.um.ac.ir, 2011. [Online]. Available: https://wtlab.um.ac.ir/images/e-library/Question_Answering/A%20survey%20on%20question%20answering%20technology%20from%20an%20information%20retrieval%20perspective.pdf

[12]
“The question answering systems: A survey", *Aast.edu*, 2012. [Online]. Available: http://www.aast.edu/papers/staffpdf/19955_401_18_QA%20Survey%20Paper%20(IJRRIS).pdf.

[13]
Egr.msu.edu,2001.[Online].Available:https://www.egr.msu.edu/~jchai/QAPapers/Redundancy-Clarke01.pdf.

[14]

Nlp.stanford.edu, 2010. [Online]. Available: https://nlp.stanford.edu/pubs/wang-manning-coling10.pdf.

[15]

Aclweb.org, 2007. [Online]. Available: http://www.aclweb.org/anthology/D07-1003.

[16]

Bioinf.jku.at,2018.[Online]. Available: http://www.bioinf.jku.at/publications/older/2604.pdf.

[17]

D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", Arxiv.org, 2014. [Online]. Available: https://arxiv.org/abs/1409.0473. [Accessed: 14- 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