Department of Computer Science and Electronic Engineering, University of Essex

# Project proposal

Submitted as part of the requirements for:

CE888 Data Science and Decision Making

Name: Gao Ziyuan

Student Number: 1605045

Lecturer: Spyros Samothrakis

Date: 23 February 2017

**Abstract:**

This article introduces a project about question answering tasks for Artificial Intelligence. At first, the article shows a global view of the project. Then a research using the similar method is introduced. After that, the article explains the methodology, experiments to be performed and result evaluation in details. Finally, the expected conclusion and planned improvements are shown.

## 1. Introduction:

This project aims at utilizing statistics of libraries to achieve question answering for Artificial Intelligence. The statistics come from bABi tasks, which are tasks of five directories. In the directories is the set of 20 tasks for testing text understanding and reasoning. The tasks in English and Hindi are readable by humans while the third type of tasks are not readable by humans. The forth type of tasks have 10000 training data and the fifth type of tasks are spilt into train and valid portions (90% and 10% spilt). [1] The tasks sequences will be converted to fixed length representations of vectors with references to word2vec method. After that, these vectors are feed to a Random Forest and through this machine learning method, the questions can be read and the answers will display.

## 2. Background:

There are some researches by using Random Forest method. Cancer is a major leading cause of death and conventional methods of diagnosing cancer rely solely on skilled physicians with the help of medical imaging. Eliza Razak realized the existing diagnosis techniques using miRNA suffer from low diagnosis accuracy, sensitivity, and specificity. He circumvents these problems with Random Forest. The results are promising and encouraging. Despite much noise contaminated the preparation process and low miRNA count in body fluids. The proposed system able to identify miRNA markers responsible for classification of cancer. [2]

## 3. Methodology:

In this project, the method which is similar to word2vec method will be used to convert the tasks to representations of vectors. After that, Random Forest is utilized to deal with these vectors. Random Forests grows many classification trees. The input vector is put down each tree in the forest to classify a new object from an input vector. Each tree

gives classification and ‘votes’ for that class. The forest chooses the classification having the most votes.

Back to the point, the bABi tasks are a set of prerequisite toy tasks created by Jason Weston. [1] In tasks library, every single task corresponds to one type of questions and the type of questions determines the outcome of each forest. Assume that the number of stories in every task is N, sample n (n<N) stories at random from the task. This sample is the training set for growing the decision tree. Assume that there are M features in each task, select m (m<< M) variables at random out of the M and the best spilt on these m is utilized to spilt the node of the tree. The value of m is constant during the growth of the forest. [3] [4] Every tree grows in this method and many trees gather into one forest. Therefore, there are 20 forests for one directory. When new stories and questions are given to the forests, the value of the classification chosen by forests will be returned as the answer.

## 4. Experiments:

At the beginning, word2vec method is used to convert the statistics of the library to fixed length representations of vectors.

After that, the type of the questions determines the outcome of the corresponding tree. In view of the same type of questions in each task, Random Forest is created by statistics of every task. The stories are sampled at random and a few variables are selected at random. The best spilt on these variables is used to spilt the node. In this way, one decision tree is created and numbers of trees constitute a whole forest.

Finally, a new story and a question are read in as the input of the project. According to the question, the machine decides to enter into the suitable forest, outputs the values of features chosen by the forest and numbers of sentences indicating the key points of questions.

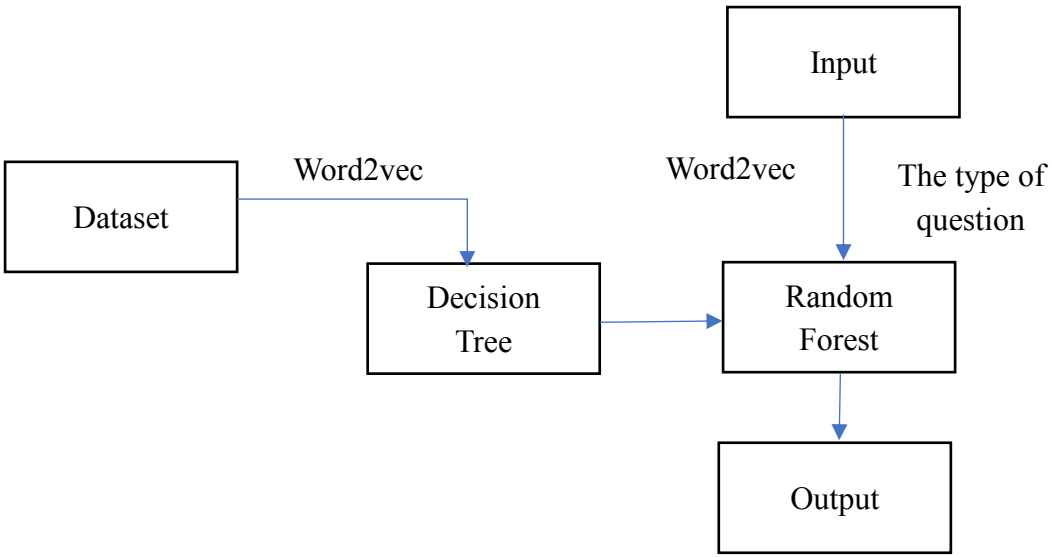



Diagram 1 Structure Diagram of the project

## 5. Discussion:

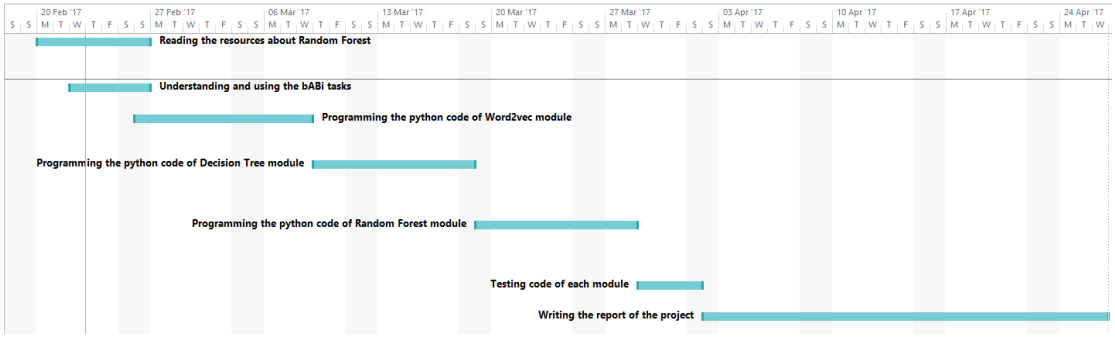
The statistics of the bABi tasks are divided into training set and testing set. The training set is utilized to grow the decision trees and the trees constitute Random Forest. Differently, the testing set is used as the input to test whether every function of the project work normally or not. The output of the testing should be compared with the output of the training. For example, the testing data is set as the input of Random Forest. If output is consistent with output of training set, the codes of Random Forest are no problems; if output is different from output of training set, the examination of each modules is required. From the experiments, the modular mind of programming and the method of comparison test are essential for this project.

## 6. Conclusion:

The project is expected to realize the first goal of converting statistics to fixed length representations of vectors. Then after the experiments of the main codes, the decision trees will be created and Random Forest will grow normally. Hopefully when inputting stories and questions, Random Forest will work, output the correct answers and numbers of sentences indicating the key points of the questions.

Regretfully, the learning of machine is not solved. The additional function of the project is learning new things through inputting new stories and questions continuously. In the future working days, I will make great efforts to try my best to improve the learning function.

## 7. Plan:



## References:

[1] W.Jason, et al. ”*Towards ai-complete question answering: A set of prerequisite toy tasks.*” ArXivpreprint arXiv:1502.05698, Dec.31,2015.

[2] R.Eliza, et al. “*Classification of miRNA Expression Data Using Random Forests for Cancer Diagnosis*”, 2016

[3] B.Leo; C.Adele. “*Random Forests*”. Internet: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#missing1 .

[4] H.Ned, *Introduction to decision trees and random forests*, American Museum of Natural History’s Center for Biodiversity and Conservation