

QUIZ GENERATION

A Project Report

Submitted by

Vivek R. Bhave	111508015
Shreyansh R. Gopawar	111508073
Aditya D. Neralkar	111508076

in partial fulfillment for the award of the degree

of

Information Technology

Under the guidance of

Dr. Y. V. Haribhakta

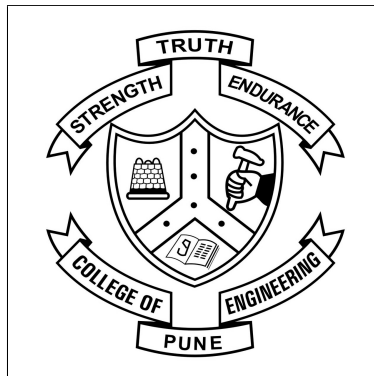
College of Engineering, Pune

DEPARTMENT OF COMPUTER ENGINEERING

AND

INFORMATION TECHNOLOGY,

COLLEGE OF ENGINEERING, PUNE-5



April, 2019

**DEPARTMENT OF COMPUTER ENGINEERING
AND
INFORMATION TECHNOLOGY,
COLLEGE OF ENGINEERING, PUNE**

CERTIFICATE

Certified that this project, titled “QUIZ GENERATION” has been successfully completed by

Vivek R. Bhawe	111508015
Shreyansh R. Gopawar	111508073
Aditya D. Neralkar	111508076

and is approved for the partial fulfillment of the requirements for the degree of “B.Tech. Information Technology”.

Dr. Y. V. Haribhakta

Project Guide

**Department of Computer Engineering
and Information Technology,
College of Engineering Pune,
Shivajinagar, Pune - 5.**

Dr. V. Z. Attar

Head

**Department of Computer Engineering
and Information Technology,
College of Engineering Pune,
Shivajinagar, Pune - 5.**

Abstract

Asking questions to students has always been regarded as the best method to gauge the students learning capabilities. However, with the amount of content available, generating questions from domain experts, is a time consuming and expensive process.

The task of Automatic Question Generation(AQG) aims to generate questions from a given piece of text which shall appropriately test the students mastery over the material. Plenty of research in the area of AQG has been done over the years. The current state of the art models use a bidirectional Long Short Term Memory encoder decoder system to solve the problem. The work presented here carries out an important task in one of the stages of the entire AQG system.

The model we have designed takes the sentence to generate questions on, as input, and tags a few words of the sentence as answer words on which questions can be created. The model has been trained using the SQUAD and Google datasets. The model has been created using LSTM networks and other linear layer neural networks.

As a part of the system, we also built a custom web crawler. On input of a specific topic to the crawler, it can find a list of relevant web pages and is also able to filter in the important content from those webpages. Thus, this is an end to end system, which can be used in many areas like education, quizzing, entertainment, etc.

Contents

List of Tables	ii
List of Figures	iii
1 Introduction	1
1.1 Importance of Automatic Question Generation (AGS) Systems	1
1.2 Neural Networks	2
1.3 Recurrent Neural Networks (RNN)	2
1.3.1 Problems with RNN	4
1.4 Long Short Term Memory	4
2 Literature Survey	6
3 Problem Statement	13
3.1 Motivation	13
3.2 Objectives	13
4 System Requirement Specification	15
5 System Design	17
5.1 Topic Input & Data gathering	17
5.2 Pre processing block	19
5.3 Question Generation - Algorithm	21
5.4 Output	22

5.5	Results	23
5.5.1	Query: College of Engineering, Pune	23
5.5.2	Query: Madhuri Dixit	24
5.5.3	Query: Narendra Modi	25
5.5.4	Vocabulary Based	26
6	Machine Learning Approaches	27
6.1	Dataset generation:	27
6.2	Models:	28
6.2.1	Model using LSTM	28
6.2.2	Simple Dense Layer Model	29
6.2.3	Results of Simple Dense Layer Model	30
7	References	31
8	Publication Details	33

List of Tables

2.1	Comparison Table	10
-----	----------------------------	----

List of Figures

1.1	Basic Neural Network Diagram	2
1.2	Recurrent vs Feed-forward Neural Network	3
1.3	Structure of a Neuron	3
1.4	Structure of a Basic LSTM node	5
5.1	Architecture of Model	18
5.2	Query: College of Engineering, Pune	23
5.3	Query: Madhuri Dixit	24
5.4	Query: Narendra Modi	25
6.1	Softmax Activation Formula	29
6.2	Sigmoid Function	30
6.3	Binary Cross-Entropy Formula	30
6.4	Results of Simple Dense Model	30

Chapter 1

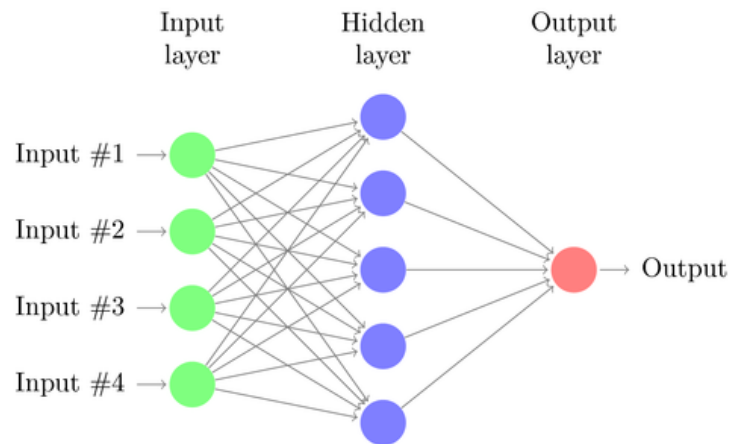
Introduction

1.1 Importance of Automatic Question Generation (AGS) Systems

Gauging a students knowledge over a certain topic is one of the fundamental problems in education systems. Quizzing students has long been accepted as one of the best techniques to solve this problem. Studies conducted over the last several decades have found that providing students with frequent and good number of quiz questions leads to better understanding of the topics, than spending an equal amount of time studying notes or textbooks. Quizzes can be used at several places in the education domain. They can be used in Massive Open Online Courses (MOOC) to test if a student has grasped the concept properly, they can be used as standalone tests in universities, or they can also be used as practice by students to discern their command over the subject.

However, with the volume of information available, generating questions from domain experts is an expensive and time consuming process. The problem of Automatic Question Generation (AQG), in the field of Natural Language Processing, aims to generate questions from a given text, such that students are appropriately tested on their mastery over the subject.

Figure 1.1: Basic Neural Network Diagram



1.2 Neural Networks

A neural network is an interconnected collections of neurons. The connections between the neurons are modelled as weights and the final value that the neuron stores is the weighted summation of all the inputs. A neural network consists of an input layer, an output layer and multiple hidden layers.

1.3 Recurrent Neural Networks (RNN)

Recurrent neural network is a type of artificial neural network. Neural networks had a shortcoming that that they could not remember previous inputs given to the network. This shortcoming was solved by recurrent neural networks with the help of hidden layers. Because of their internal memory, RNNs are able to remember important things about the input, which allows them to make precise predictions of the input that is yet to be processed. Sequential data is an ordered data where related things follow each other. RNN's have proved extremely useful while working with sequential data.

Unlike the normal feedforward neural networks, recurrent neural networks feed the output of previous step to the current step. This helps the network to memorise the previous inputs. Thus RNN cycles through the information

Figure 1.2: Recurrent vs Feed-forward Neural Network

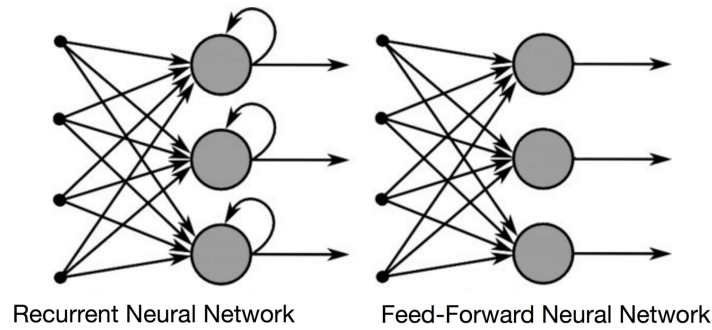
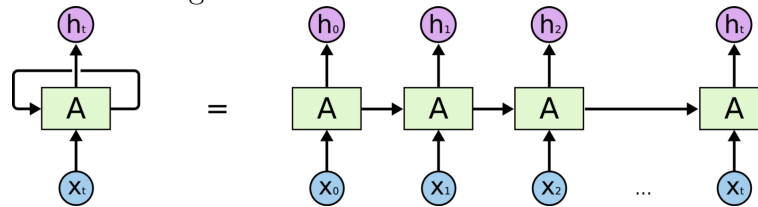


Figure 1.3: Structure of a Neuron



and when taking a decision it takes into consideration the current input as well as whatever it learned from the previous output. As an example, in the word 'Teacher', by the time a feed forward neural network reaches 'c', it has forgotten about the previous inputs 't', 'e' and 'a'. A RNN has the power to remember that. Recurrent Neural Network adds immediate past to the present. This is important since past data contains information about what is coming next.

RNNs apply weight to current and previous input and tweaks their weights using either the Gradient Descent or Backpropagation Through Time algorithms. Gradient Descent is an algorithm that is used to iteratively minimize a given function. Backpropagation Through Time (BPTT) is performing backpropagation on an unrolled RNN. Unrolling is a visualization and conceptual tool, which helps you to understand whats going on within the network. RNN can be viewed as a sequence of Neural Networks that is trained one after another with backpropagation.

The diagram above shows the RNN being unrolled after the equal sign.

The different timesteps are visualized and information gets passed from one timestep to the next. Within BPTT the error is back-propagated from the last to the first timestep, while unrolling all the timesteps. This allows calculating the error for each timestep, which allows updating the weights.

1.3.1 Problems with RNN

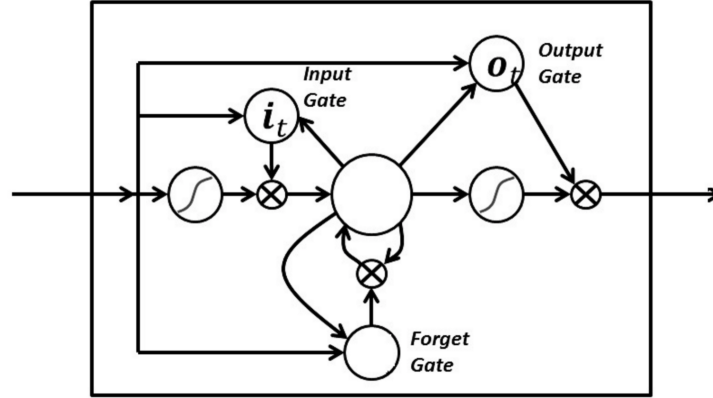
Although RNN's are quite powerful because of their inherent ability to remember previous inputs while processing each new one, they have to be used with care as a few of the following problems might occur.

1. Training RNNs is a difficult task.
2. **Exploding gradient problem** : Exploding gradient is when the algorithm assigns an extremely high importance to the weights, without much reason. But fortunately, this problem can be easily solved by truncating or squashing the gradients.
3. **Vanishing gradient problem** : Vanishing gradient is when the values of a gradient are too small and the model stops learning or takes way too long. This problem is solved by the concept of LSTM.

1.4 Long Short Term Memory

Long Short Term Memory (LSTM) is an artificial recurrent neural network. LSTM can be considered to be an extension to the recurrent neural network, as LSTM has extended memory as compared to RNN. Thus an LSTM is capable of learning things that have very large time lags between them. The memory used by an LSTM can be attributed to a computer memory as it has functions like read, write or delete information. This is quite different from that of a RNN which uses a vector.

Figure 1.4: Structure of a Basic LSTM node



The memory in LSTM are gated cells. Gated cell means that the cell decides whether or not to store or delete information based on the importance it assigns to the information. The assigning of importance takes place through weights which are learned through the algorithm.

A basic LSTM node consists of input gate, output gate and forget gate. The input gate decides whether or not to let new input in. The forget gate deletes the information because it is not important. The output gate impacts the output at the current time step. The gates in a LSTM are analog, in the form of sigmoids, meaning that they range from 0 to 1. The fact that they are analog, enables them to do backpropagation with it.

The problem of vanishing gradients which is very common in recurrent neural networks is solved by LSTM as it keeps the gradient steep enough and therefore the training short and the accuracy high.

Chapter 2

Literature Survey

1. **Computational Intelligence Framework for Automatic Quiz Question Generation**

The model described in this paper, uses a rule based approach and is able to generate three types of questions, namely, Fill in the Blank, True/False, and Wh Questions. The training phase of this algorithm involves, detecting topically important sentences, NER and POS Tagging, sanitizing the input sentence and then constructing a tree for further processing of each type of question. For Fill in the Blank, it gives weight to each word involved and marks important words for setting up blanks. The true or false sentence involves identifying modal verbs, and also factual type of sentences and further rule based processing. For Wh questions, patterns identified during the training phase are matched, and used to generate questions.

2. **Automatic Question Generation System**

This is a traditional supervised learning approach and the model can be broken down into a sequence of processes. The text input is stemmed to get the meaning of each word and then passed to the Phrase Mapper. Then the key phrases are extracted using pre trained documents and the

document is summarized. In the end, Nouns are filtered in, on which questions can be generated.

3. **Automatic Question Generation for Intelligent Tutoring Systems**

This system generates only Multiple Choice Questions, and has been trained on Wikipedia articles. The wikipedia articles are processed and unigrams, bigrams are extracted. The keywords from them are then extracted and their weights using the TF IDF weighting scheme are calculated. Using the above distractors can be generated. Whenever a query is fired to generate questions on, the wikipedia articles of the keyword is taken up, and rule based methods are applied to the sentences to create questions and presented to the users including the distractors as other viable options.

4. **Automatic Question Generation from Childrens Stories for Companion Chatbot**

This paper was specifically targeted for generating questions from childrens stories, and was trained and tested on much lesser data than others. The model is said to work in two parts, that is one for question generation and the other for ranking of the questions. The model uses the part of speech tags and dependency parsing algorithms along with pre trained language rules for the question generation phase. For ranking of the questions, the model uses logistic regression to determine the acceptability probability of the question.

5. **Deep Guessing**

Generating Meaningful Personalized Quizzes on Historical Topics by Introducing Wikicategories in Doc2Vec: The aim of this model is to load

all of the wikipedia articles, classify them into categories, and use the knowledge from the above, for generation of distractors in multiple choice questions. The model works on a novel idea of creation of paragraph vectors, whose advantage is that they are trained from unlabeled data and thus can work well for tasks which do not have enough labelled datasets. After the score of wiki categories is obtained, thresholds are decided to classify the records into categories and thus, are eventually used for the creation of meaningful options in the questions generated.

6. **Automatic question generation on the basis of the discourse connectives**

This problem of question generation has been divided into two modules in this paper. The first part is that of Content selection and the next that of Question formation. The Content selection phase consists of recognizing the part in text that is relevant and important to generate questions on. The second phase of Question formation includes several subtasks like Word Sense disambiguation of the discourse connectives then the Identification of the type of question to be created and eventually applying syntactic transformations on the context. The paper mainly takes into consideration the seven main discourse connectives - although, as a result, because, for example for instance and since. Using the output generated so far, the type of question gets decided and then the question can be formed.

7. **Semantic Based Automatic Question Generation**

The paper explains a system that applies two fundamental Natural Language Processing concepts namely Semantic Role Labeling and Named Entity Recognizer technique. These tasks are used to convert the in-

puted sentence to semantic pattern. The model described in the paper has developed a system which has identified patterns for each type of question. The question types under consideration here are all of the Wh-questions - who, when, why, where, what. The system for classification uses learning, storage memory, feature extraction, and associative Retrieval.

The sentence given as input will be first parsed using Named Entity Recognition and SRL technique. The output of these two algorithms has a direct correlation with the exact question type to be created. Thus after the question type identification, the question pattern is known using the pretrained rules.

8. **Automatic Multiple Choice Question Generation System for Semantic Attributes Using String Similarity Measures**

The paper has described a model which first selects a factual sentence and an important word to generate questions on from the text given as input. This selection is done on the basis of the semantic labels and named entities in the sentence. Then for actually generating a question the SRL and NER tag is used, which directly helps in finding out the type of question.

Then the model focuses on finding distractors i.e. the incorrect options of a question. For this task different similarity measure between sentences of the data set is taken into account. Eventually, when a question is generated, the measures of similarity between the actual question and the sentences of the input text is considered and sorted. The top three sentences obtained from the above procedure, are considered for finding a relevant important word to the question generated which are eventually

used as distractors.

Table 2.1: Comparison Table

Sr No.	Algorithm	Methodology	Type of Question	Evaluation of Result
1.	Computational Intelligent Techniques	Rule Based Approach - Training for patterns and then generate	Fill in the Blanks, True/False, Wh questions	Manual checking for finding difference between Automatic Questions and Domain Experts - users found difference
2.	Automatic Question Generation System	Series of pre processes like stemming, summarize, noun extract, and then rule based question generation	All Wh questions along with type like - Person, Definition, Procedure, Consequence, etc	Compression and Omission Ratio considered - acceptable results

3.	AQG for Intelligent Tutoring Systems	Keywords from Wikipedia and then TF-IDF, then rule based	MCQs along-with distractors	Manual evaluation with domain expert
4.	AQG from Childrens Stories for Companion Chatbot	POS Tagger, Dependency Parsing and Logistic Regression	Only Wh questions	Manual evaluation with over 50% acceptance
5.	Deep Guessing: Quizzes on Historical Topics by Wikicategories in Doc2Vec:	Generating paragraph vectors from wikipedia articles, categorization for distractors	MCQ Questions with distractors	Manual evaluation of 45 MCQs
6.	AQG on the basis of the discourse connectives	Content selection and Question formation	Question generation like Why, when, where, in which	Manually evaluated for semantic and syntactic soundness of question by two evaluator

7.	Semantic based AQG	Feature Extraction using NER, SRL, Sentence pattern finding and get question type pattern	Only Wh questions	Manual evaluation with small dataset, About 87% accuracy
8.	AQG System for Semantic Attributes Using String Similarity Measures	Extracting sentence, check viability by similarity of sentences and select top three sentences	Only MCQ type questions	Manual evaluation with 75% success.

Chapter 3

Problem Statement

The problem of question generation, is currently solved by domain experts, who create questions over text or topics by their intuition or knowledge. Automatic Question Generation can solve this laborious, time consuming task quite efficiently and with much higher accuracy. Such an automatic question generation system can have applications in multiple domains.

3.1 Motivation

The main motivation of designing an Automatic Question Generation System, was to contribute in the education domain. Quiz questions can be of utmost importance for teachers to ensure that students are able to understand the material taught in class. Similarly quizzes can be used by students to find out their preparedness of the subject at hand.

3.2 Objectives

The objective of the project is to devise a system to generate a set of questions based on a topic provided by the user. The questions generated by this system, should be similar to questions generated by a domain expert. The difficulty level of the questions generated, should be such that they appropri-

ately test the knowledge of the user.

Chapter 4

System Requirement Specification

The system uses the following technologies and libraries relating to them -

Technology:

- Python 3.6

Libraries:

- Numpy
- Scikit
- Keras
- TensorFlow
- Torch
- Pytorch

The hardware specifications of the system used for running experiments are -

- Nvidia GEFORCE RTX 2080Ti
- Ram 16 GB
- 512 SSD

- 2 TB HDD

The dataset used in training, validation and testing of the model was the SQUAD 2.0 (Stanford Question Answering Dataset). This is the most recent and state of the art dataset used as a benchmark for testing Automatic Question Generation Systems.

Chapter 5

System Design

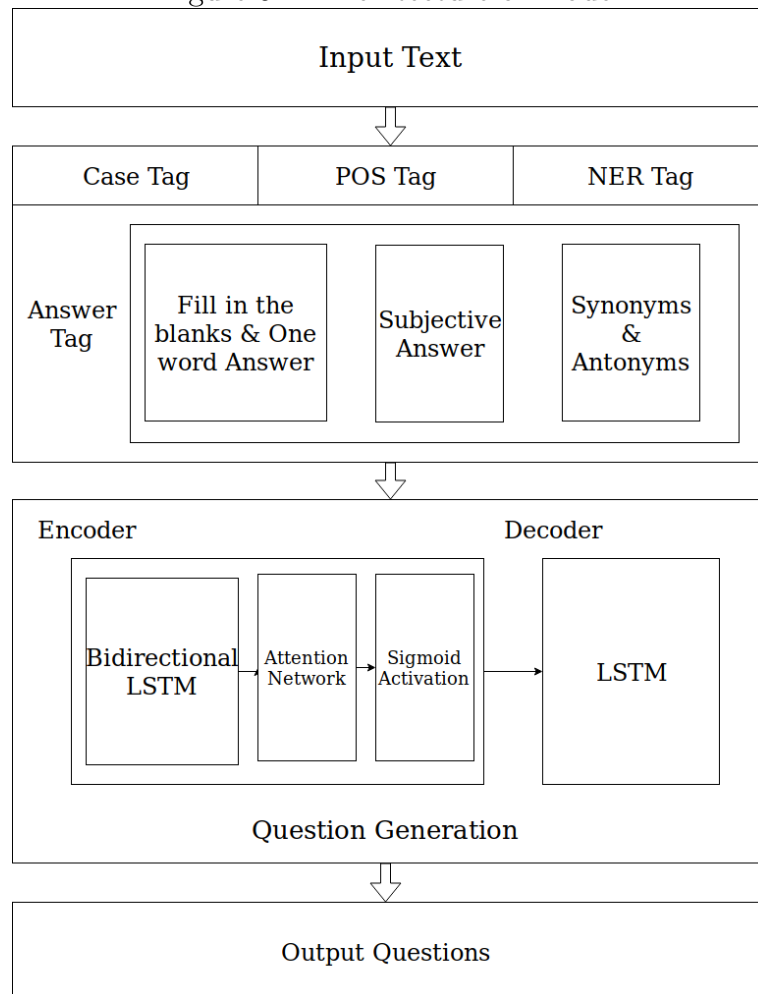
The complete work presented here can be divided into several parts right from taking in a topic as input, finding out relevant documents, tagging words as probable answer words, and eventually generating questions from it.

5.1 Topic Input & Data gathering

Initially, the user shall enter his/her topic of interest into the software. The software shall then crawl the web and gather links to various web pages relevant to the topic given by the user. On getting the links, further the software shall scrape those links and parse the HTML pages to get the information content in those web pages.

However, all the content received can be classified into three parts. The first part consists of markup language tags and syntax, the second part shall contain data unrelated to the topic of interest, and the third shall be the data which is useful and relevant both to the topic of interest. The data classified into the first part, is removed by searching for markup language tags in the obtained data. On finding such tags, the lines containing such tags shall be eliminated from the useful data text. For identifying data in the second part, the semantic meaning of the words and the web page from where the data

Figure 5.1: Architecture of Model



has been input is taken into consideration. Thus eventually the system has data which can be used for generating questions on the users topic of interest.

5.2 Pre processing block

The processing block consists of four independent tasks. The output of each of these four tasks works as features for the question generation algorithm.

The four tasks and their work can be summarized as follows -

1. Case Tagging - This feature determines whether the first letter of the word is in lower or upper case. The output of this case thus determines whether the word is the starting word of the context. In quite a few cases, the starting word of the context dominates in meaning, and hence has been considered as a feature in the problem.
2. NER Tagging - The 7 set Stanford NER Tagging is used in the system. The NER Tagging helps in finding out the type of a noun to which the word belongs, i.e if it is a place, a date or a person. Depending on the tag different kinds of 'Wh' questions are created and hence the feature is of great significance in the problem.
3. POS Tagging - POS tags are important in generating patterns based on their frequency in the contexts. Frequently occurring series of consecutive POS tags that form answer words in the training sentences are most probable to be the best sources for generating questions in test contexts as well.
4. Answer Words - This feature is the most important of all the four features. This feature talks of whether the word in consideration is an answer word or not, in the question to be formed. Only those words

that answer words are given paramount importance in the question generating process. So to achieve this goal, different rule based algorithms are used depending on the type of questions to be formed

- (a) Fill in the blanks & One Word Answers: In this question type, a factual important sentence is chosen from the input text. NER Tags of the words in this sentence are found out. Through our experimentation it was found out that consecutive words with same NER tags are a good fit as probable answer words. It is also ensured that the tags of such probable answer words do not belong to the OTHER category.
- (b) Subjective Questions: To generate Subjective Questions, possible patterns which lead to subjective questions were researched in the dataset. On research, words like because, means, states that, defines, etc were identified as words whose sentences had content for generating subjective questions. The words in a sentence following these special words were tagged as possible answer words.
- (c) Vocabulary based Questions: The software offers a special feature for improving students language skills. Once the content collection phase is completed, the system identifies difficult and rarely used words in the passage. On identification of such words, a user is presented with the meanings of such words and asked to identify the special word in the passage. Conversely, a user may also be provided with a word opposite in meaning to the special word, and would then be asked to identify the special word selected by the system. The tasks of identifying special words, the word meanings and words opposite in meaning to the given word are achieved using

the word net library.

Some machine learning models using RNNs have been explored for tagging purpose the answer words. These approaches are aimed at tagging answer words for forming all types of question mentioned above. They are discussed in Section : Machine Learning Approach.

In order to achieve this goal, a model containing a LSTM and dense layers is developed. The output of the model is binary string which is 1 for the Answer Tag and 0 for non answer words. The output of this preprocessing subsystem is fed to the Question Generation Algorithm subsystem as input.

5.3 Question Generation - Algorithm

The Question Generation Algorithm performs three major tasks -

1. Fill in the Blanks: To create Fill In The Blanks questions, all the words that are marked as Answer tag are made blank.
2. Synonym and Antonym words are asked directly.
3. Basic tense filling grammar is generated by entering blank in the verb part of the sentence and the base word is provided in brackets.
4. Subjective and One word WH Questions is generated by passing the preprocessed sentence to QG-Net model. The QG-Net model contains the following parts and each is explained below:

- **Context Reader and Encoder**

Context reader is a bi-directional long short-term memory (bi-LSTM) network with 600 hidden layers. Bi-LSTM processes the input words

in both the forward and backward direction. Encoder creates context word vectors from the preprocessed input provided.

- **Decoder and Softmax activation**

The question-generator generates a question word one-by-one by following two internal calculations. The unidirectional LSTM in the decoder maps the current question word with fixed-sized vector, which is the hidden state of the network. Then the second calculation, i.e. the softmax function calculates the probability distribution over all words from a fixed question vocabulary.

- **Relevance of Question**

To select the relevant question, pointer network is used. Pointer network calculates the output word probabilities as a mixture of two probabilities, one over the question vocabulary and the other over the input context vocabulary

5.4 Output

In the end, the system shall generate questions on either the topic of interest specified to the user or the text given as input to the system. The system can generate various types of questions. The first type is Fill in The Blank questions, where a factual or important word is left blank in a sentence from the input text. The second type is a one word answer, where an important sentence from the input text, is converted to a question and presented to the user. It also generates basic grammar questions and synonym & antonym words. Finally, the system can also generate reading comprehension type of questions, wherein the input text on which questions are generated is made available to the user while he/she is answering the questions.

Figure 5.2: Query: College of Engineering, Pune

No.	Question	Your Answer	Correct Answer	Score
1	what is the name of pune 's engineering institute ?		College of Engineer0	
2	what is the name of the autonomous engineering institute affiliated with savitribai ?		Pune	0
3	what is the name of the oldest engineering colleges in asia ?		College of Engineer0	
4	where is the oldest engineering colleges in 1854 located ?		Asia	0
5	when was the university of kansas founded ?		1854	0
6	what school of engineering does pune belong to ?		College of Engineer0	
7	what is another name for college of engineering ?		Pune	0
8	when was the " poona model " referred to as ?		1950s	0
9	_____ , pune (coep) is an autonomous engineering institute affiliated to savitri		College of Engineer0	
10	college of engineering , _____ (coep) is an autonomous engineering institute affiliated to savitribai		Pune	0
11	established in 1854 , it is one of the oldest engineering colleges in asia , after _____		College of Engineer0	
12	established in 1854 , it is one of the oldest engineering colleges in _____ , after college of engineer		Asia	0
13	established in _____ , it is one of the oldest engineering colleges in asia , after college of engineer		1854	0
14	the students and alumni of _____ , pune are colloquially referred to as coepians		College of Engineer0	
15	the students and alumni of college of engineering , _____ are colloquially referred to as coepians		Pune	0
16	the college 's study model was referred to , in the early _____ , as the " poona model "		1950s	0
17	Word that means: an association organized to promote art or science or education		institute	0
18	Word that means: institute, enact, or establish		Established	0
19	Word that means: be relevant to		referred	0
20	Word that means: be relevant to		referred	0
21	Word that means: a complex of buildings in which an institution of higher education is housed		college	0
22	Antonym of: abolish		Established	0
23	Antonym of: middle		early	0
				Total: 0
				Percentage: 0.0

OK

5.5 Results

Questions generated after querying terms using the developed system are as follows:

5.5.1 Query: College of Engineering, Pune

Many of the questions generated on the given topic are similar to the Fill in the Blanks questions, being quite factual in nature. Most of the questions created, not only would be similar to the ones created by domain experts, but as a group of questions also appropriately test the knowledge obtained the user after reading the material. The overall results obtained are quite encouraging to be used for any of the proposed applications.

The blank words are completely factual, which can be answered on the basis of ones general knowledge. Hence the quality of the questions is quite good. However, in a few cases the blanks where the sentence under consideration uses pronouns in reference to the answer. In such cases the question cannot be answered by anyone without reference to the context of the paragraph.

Figure 5.3: Query: Madhuri Dixit

No.	Question	Your Answer	Correct Answer	Score
1	when was ' madhuri ' born ?		May 1967	0
2	what is the name of beyonce 's married name ?		Madhuri Dixit	0
3	how long has she been one of the most popular and highest-paid film actresses actresses ?		1980s	0
4	who won the 2008 padma shri ?		India	0
5	in what year did the government of india receive her with the padma shri ?		2008	0
6	who was the fourth highest civilian honour of the country ?		Padma Shri	0
7	madhuri dixit (born 15 _____) , also known by her married name madhuri dixit nene , is		May 1967	0
8	_____ (born 15 may 1967) , also known by her married name madhuri dixit nene , is		Madhuri Dixit	0
9	one of the most popular and highest-paid hindi film actresses in the late _____ and 1990s , she ha		1980s	0
10	in 2008 , the government of _____ awarded her with the padma shri , the fourth highest civilian ho		India	0
11	in _____ , the government of india awarded her with the padma shri , the fourth highest civilian ho		2008	0
12	in 2008 , the government of india awarded her with the _____ , the fourth highest civilian		Padma Shri	0
13	Word that means: broadcasting visual images of stationary or moving objects		television	0
14	Word that means: a female actor		actress	0
15	Word that means: be cognizant or aware of a fact or a specific piece of information; possess knowledge or inform		known	0
16	Word that means: a person of considerable prominence		personality	0
17	Word that means: a well-known or notable person		name	0
18	Word that means: someone who finds financing for and supervises the making and presentation of a show (play c		producer	0
19	Word that means: support or hold in a certain manner		born	0
20	Word that means: have the quality of being; (copula, used with an adjective or a predicate noun)		been	0
21	Word that means: express approval of		praised	0
22	Word that means: have as a part, be made up of		including	0
23	Word that means: the people who live in a nation or country		country	0
24	Word that means: a tangible symbol signifying approval or distinction		honour	0
25	Antonym of: ignore		known	0
26	Antonym of: unmarried		married	0
27	Antonym of: unknown		known	0

OK

5.5.2 Query: Madhuri Dixit

Although a reasonable number of quality questions have been generated, some of the problems seen in other test results are evident here. The QG-Net model, on some occasions substitutes the word beyonce in place of a possible noun whose pronoun has been used in the original sentence. Unable to get the reference to the noun from the overall context, this substitution is used by the model. In a few rare cases, the model does not accurately identify the question type and hence uses the wrong Wh word, in creating the appropriate question.

In this case as well, the blank words are factual and hence, the quality of questions has been maintained. The blanks include all aspects of the persons life including but not limited to, important dates in the persons life, names of popular plays and films and her various roles in the films. However, more than one question can be created on a few of the sentences and hence care must be taken to provide only one of such questions to the user.

Figure 5.4: Query: Narendra Modi

No.	Question	Your Answer	Correct Answer	Score
1	what country was damodardas in 14th and current prime minister ?		India	0
2	in what month and year was narendra born ?		September 1950	0
3	who is the 14th and current prime minister of india ?		Narendra Damodardas	0
4	who was the chief minister of gujarat ?		Member of Parliament	0
5	what was the chief minister of varanasi in 2001 ?		Gujarat	0
6	when was the chief minister of gujarat ?		2001	0
7	what organization does modi belong to ?		Bharatiya Janata Party	0
8	narendra damodardas modi (pronounced ['nɛrɛndrə dɑmɔdər'dɑs 'mɔdi:] (listen) ; born 17 septem		India	0
9	narendra damodardas modi (pronounced ['nɛrɛndrə dɑmɔdər'dɑs 'mɔdi:] (listen) ; born 17 sept		September 1950	0
10	_____ (pronounced ['nɛrɛndrə dɑmɔdər'dɑs 'mɔdi:] (listen) ; born 17 sept		Narendra Damodardas	0
11	he was the chief minister of gujarat from 2001 to 2014 , and is the _____ for		Member of Parliament	0
12	he was the chief minister of _____ from 2001 to 2014 , and is the member of parliament for varan		Gujarat	0
13	he was the chief minister of gujarat from _____ to 2014 , and is the member of parliament for var		2001	0
14	modi is a member of the _____) , and of the rashtriya		Bharatiya Janata Party	0
15	Word that means: hear with intention		listen	0
16	Word that means: speak, pronounce, or utter in a certain way		pronounced	0
17	Word that means: a leader engaged in civil administration		politician	0
18	Word that means: do duty or hold offices; serve in a specific function		serving	0
19	Word that means: support or hold in a certain manner		born	0
20	Word that means: one of the persons who compose a social group (especially individuals who have joined and pe		member	0
21	Word that means: a person who performs voluntary work		volunteer	0
22	Word that means: one who loves and defends his or her country		nationalist	0
23	Word that means: a group of people who work together		organisation	0
24	Antonym of: noncurrent		current	0
25	Antonym of: unborn		born	0
26	Antonym of: nonmember		member	0
27	Antonym of: defense		attack	0

5.5.3 Query: Narendra Modi

The quality of questions in this case is similar to the one observed in the other cases. Alongwith the problem of incorrect identification of the question type, a couple of other problems exist in rare cases. Sometimes, the reference to a particular domain gets shifted to another one during computation in the QG Net model, which generated a question on Eisenhower, although no reference has been made to this person in the article. In a few unstructured sentences, the model does not perform well enough and mixes up possible part of phrases with one another, eventually generating an incorrect question.

With the text collected being quite detailed, questions have been generated around all the activities of the person. Most of the questions created are about dates of events, or about names of other people he dealt with throughout his life. Such features are largely dependent on the input text and can greatly vary depending on the text. As with the size of the input text, the number of questions generated has also increased significantly.

5.5.4 Vocabulary Based

Since the vocabulary based questions had similar results on all case studies tried out, all the results have been compiled under a common section here. The quality of questions generated of synonyms and antonyms was quite good, as they were fairly guessable when the paragraph was made available to the user. Identifying the appropriate word from the passage given, requires a bit of application in a few cases, as the meaning of the word in the context used can vary significantly from the one provided in the question. In the case of antonyms, a few of the antonym words can be easily obtained by adding appropriate prefixes like *ir*, *un*, etc. However, in other cases, obtaining the answer word from the passage given the word opposite in meaning to the answer, shall test the vocabulary of the user appropriately.

Chapter 6

Machine Learning Approaches

The QG-Net model is not a complete end to end question generation model. One of the preprocessing steps as required by the model, of tagging possible answer words in a given test sentence is not performed by the model. To perform this preprocessing task, various machine learning models were tried. Some of the approaches giving relatively the best results have been discussed below:

6.1 Dataset generation:

For the training, validation and testing the machine learning models SQUAD dataset was used. Dataset contains files which have context paragraphs, questions, the location of starting and ending characters which form the answer and few more fields in json format. The dataset was parsed into csv files for model development and data analysis purposes.

6.2 Models:

6.2.1 Model using LSTM

For generating questions, a model is required which can tag possible answer words in a sentence. To accomplish this task, the understanding the context is necessary and hence a RNN based approach was finalised. To overcome the common problem of vanishing gradient, a LSTM neural network was also applied. LSTM traverses the sentence word by word and can understand the complete context of the sentence - a fundamental requirement of the task. 2 dense layers using ReLU and Softmax activations, respectively, are used. The last dense layer has 2 nodes as the output of this model. The two numbers for each word represent the probability of that word being the answer. Sum of these probabilities is 1 so categorical cross entropy loss function is used in this case.

Softmax activation function

The softmax function is a function that takes as input a vector of K real numbers, and normalizes it into a probability distribution consisting of K probabilities. That is, prior to applying softmax, some vector components could be negative, or greater than one; and might not sum to 1; but after applying softmax, each component will be in the interval $(0, 1)$, and the components shall add up to 1, so that they can be interpreted as probabilities.

Categorical Cross Entropy

The results of this model were not encouraging as the probability that a word is an answer word was far too less as compared to probability that it is not. So no word of the test sentence was tagged as an answer word. On

Figure 6.1: Softmax Activation Formula

$$L(y, \hat{y}) = - \sum_{j=0}^M \sum_{i=0}^N (y_{ij} * \log(\hat{y}_{ij}))$$

analysis it was found that this model was extremely heavy for a simple task of tagging answer words. Another reason as identified was that training a LSTM requires very large dataset. Hence, it was decided to try with a simple linear layer model.

6.2.2 Simple Dense Layer Model

A simple model containing Dense layers and Sigmoid activation function is built. The output of the model is a binary string which contains 0s and 1s where 1 indicates that the word is a possible answer word and 0 being not an answer word.

This is a multiclass multilabel task which means more than one output classes (multi-word output) are present and more than one word can be tagged as answer i.e. sum probabilities of words being tagged as answer should not be 1. To achieve this sigmoid activation function with binary cross entropy loss are used.

Adam optimizer

Adam is an optimization algorithm that is used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data. In classical stochastic gradient, a single learning rate for all weight updates are maintained and the rate does not change during training. But in adam optimizer, the learning rate is maintained for each network weight and separately adapted as learning unfolds.

Figure 6.2: Sigmoid Function

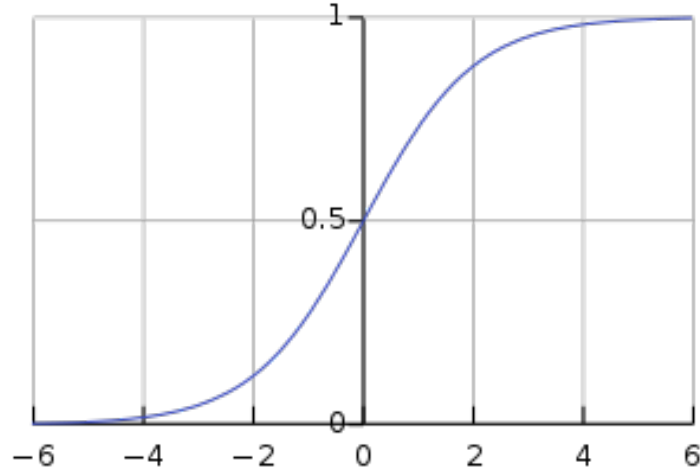


Figure 6.3: Binary Cross-Entropy Formula

$$H_p(q) = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))$$

6.2.3 Results of Simple Dense Layer Model

On testing the model on various inputs, it was found out that some words of the sentences were tagged as answer words and this was given as input to the QG-Net algorithm. On analysis of the generated questions, it was found that the questions generated were better than the previous model. On trying out this model on few more topics it was found that around 60% of the questions were answerable and grammatically correct.

Figure 6.4: Results of Simple Dense Model

```
'what is the ancient city of the city of lahore ?'
'what is another name for bahu lona wala ?'
'what kind of music do mehdi hassan and malika pukhraj have ?'
'how far is richmond from charlottesville ?', 'what is the capital of richmond ?'
'what kinds of governmental agencies does richmond have ?'
'what did patrick henry do to give me death ?'
'what was the basis of the virginia state capitol ?'
'what did george washington design to bypass the fall line ?'
'what was the outcome of the confederate confederacy ?'
'what is one of the largest markets in the confederacy ?'
```

Chapter 7

References

During the literature summary an exhaustive survey of the research already conducted in Automatic Question Generation Systems was conducted. Following are a few of the research papers, we referred -

1. Deep Guessing: Generating Meaningful Personalized Quizzes on Historical Topics by Introducing Wikicategories in Doc2Vec <https://ieeexplore.ieee.org/document/8501891/>
2. Automatic Question Generation for Intelligent Tutoring Systems <https://ieeexplore.ieee.org/document/8066538/>
3. Context Aware Restricted Tourism Domain Question Answering System <https://ieeexplore.ieee.org/document/7877473/>
4. Automatic Question Generation System <https://ieeexplore.ieee.org/abstract/document/6996216/>
5. Automatic Question Generation from Childrens Stories for Companion Chatbot <https://ieeexplore.ieee.org/document/8424749/>
6. Computational Intelligence Framework for Automatic Quiz Question Generation <https://ieeexplore.ieee.org/document/8491624>

7. QG-net: a data-driven question generation model for educational content <https://dl.acm.org/citation.cfm?id=3231654>

Among other things, we would also like to add a few of the online resources we used to learn concepts of Deep Learning in Natural Language Processing, which were of foremost importance in the design and implementation of the model we have proposed.

1. <https://www.coursera.org/learn/neural-networks-deep-learning?specialization=deep-learning>
2. <https://www.coursera.org/learn/nlp-sequence-models>
3. <https://rajpurkar.github.io/SQuAD-explorer/explore/v2.0/dev/>
4. https://cloud.google.com/storage/docs/gsutil_install

Chapter 8

Publication Details

Our research work over the year, has been drafted in the form of a journal paper, and has been submitted for review at the
Conference.