

# Intelligent Questions and Summary Generator - Examally

J. A. T. C. Jayakody  
Undergraduate  
Department of Software Engineering  
Sri Lanka Institute of Information  
Technology  
Malabe, Sri Lanka

A. D. I. U. Amarathunga  
Undergraduate  
Department of Information  
Technology  
Sri Lanka Institute of Information  
Technology  
Malabe, Sri Lanka

W. A. S. L. Jayarathne  
Undergraduate  
Department of Information  
Technology  
Sri Lanka Institute of Information  
Technology  
Malabe, Sri Lanka

## ABSTRACT

Automatic question generation and text summarization can be used to make the process of studying or teaching a lesson much more efficient and also effective at the same time. With the advancement of technology, everything in the world is automating, and the education sector represents a major character in that development to make things more comfortable and meaningful. Therefore, a perfect tool for students and teachers could be created with the combination of converting speech to text with modern technologies. This paper presents an intelligent question generation and lesson summarization system (Examally) which allows the student to get a summarization or a set of questions based on a specific lesson or chapter and study in a much more efficient manner by managing their study time. This system will make studying easier and save the valuable time of a student and also help the lecturers/teachers to save their time and effort when they have to give a small test, or a summary based on a lecture/lesson they deliver. Therefore, in this system, a user can input a textual document or an audio recording of a lesson to get it summarized or generate questions out of it. Generated questions will be in WH, binary, gap-fill and MCQ formats. With the failure of existing systems to provide such a wide variety of functionalities with a sense of completeness as in the proposed system, current products have not been able to make a considerable impact on our lives. Our solution is expected to fill the gap between conventional teaching or studying methods and the application of new technologies of NLP and ML in studying or teaching.

## General Terms

WH Questions - Questions starting with WH-words including what, where, when, who, which, whom, whose, how and why

MCQ - Multiple choice questions

NLP - Natural Language Processing

ML - Machine Learning

## Keywords

NLP, Text summarization, WH, Gap-fill, MCQ, Automatic question generation

## 1. INTRODUCTION

Once, students learn the concepts, the knowledge they obtained are evaluated through exams. Therefore, preparing for exams is very important for students and question paper standard is another important factor. Nowadays, Teachers and students have a heavy workload. Due to teachers haven't much time to create an exam paper and review the standard of the papers. Also, students have little time to study the whole syllabus. Due to not enough time, they missed to learn

important factors, trouble to memorize knowledge, study necessary lessons/concepts. Because of those reasons, many students can't get success in exams.

The proposed system will be very beneficial and helpful for teachers and students with their busy schedule in teaching and learning processes. This system saves valuable time of students by helping them to memorize their knowledge easily. Teachers have much time to review the generated paper by using this system. This paper presents a solution with an automated system that can generate questions from the extracted text of the uploaded document or voice clips that included a lesson. Also, this system generates a summary from the given text by extracting and filtering the key information regarding the topic. Generated questions in different types such as WH type, fill-in-the-blanks type, yes/no type and MCQ type. Even though there are many existing systems which implemented through this research area, they used only a textual document as the input. But, in the proposed system, audio files will also be taken as an input to overcome this limitation. The proposed system (Examally) uses natural language processing (NLP) for the implementation of question generation, and knowledge base is used to generate answers for the generated questions and also to validate those questions to provide an accurate idea relevant to the uploaded content. The knowledge base has become an important means of structuring knowledge and building knowledge-intensive systems. Hence, main intention in this research is to, come up with a solution by implementing intelligent questions and answer generating and also a summarizing system for teachers and students via a user-friendly web application which has better procedure to automatically generate various kind of questions and a summary to get a quick understand about the lesson.

## 2. LITERATURE REVIEW

Various kind of early experiments and researches have been conducted related to the Intelligent Questions and Summary Generation system and several research papers, and articles were studied and analyzed to understand the properties and requirements of a new system by creatively thinking and grasping about the features in existing running systems. Even though there are some similar applications, there is no such proper application as the proposed system to generate multiple types of questions and also a summary from the uploaded document.

Automatic Question Generation System from Paragraph was a research done to present an approach to generate questions from a paragraph, and the size of the input paragraph is defined by its scope. Hence, a mix of syntax and semantic-based approach to natural language processing is used to generate the questions from the paragraph. It involves two

main procedures, such as generating questions based on simple sentences and generating questions based on complex sentences. Also, they have highlighted the necessity of generating different types of questions like the yes/no questions and summary questions. In this research, questions were generated based on the selected sentences only [1].

Automatic Question and Answer Generation from Course Materials was the system based on the approach of Natural Language Processing to analyze the text sentence by sentence and generate all possible questions. System functionalities have divided into four main steps as identifying key phrases and words, forming a proper question, validating generated questions and find answers for the validated question. In here the question, a generation has limited into two types of questions and knowledge base has not highly considered [2].

Automatic Question Generation System Using Software Agents for Technical Institutions was developed and implemented was to take input in the form of a text file from user which contains the text upon which the user aspires to fetch questions; the output is produced in the format of a textual file including questions based on Bloom’s taxonomy in this system. But in here, they have created only one type of question set, and answers for all these questions were not generated according to that [3].

Automatic Question Generation from Text, extracted and simplified multiple statements from complex sentences including operations for various syntactic constructions encoded with pattern matching rules for trees. Researchers have used rules to extract and simplify sentences which are motivated by linguistic knowledge as their implementation method. Answers for the generated questions are not generated here, and users are unable to get the set of questions by uploading the whole document in this system [4].

### 3. METHODOLOGY

Natural Language Processing (NLP) is the major technology has been used in the implementation of this intelligent question generation system. Therefore, different kind of methods and approaches such as algorithms related to the selected research area have been applied in the development process of this research to provide a high quality of accurate output effectively. The system consists of the four main functions such as audio analysis, text recognition and question generation, MCQ generation and question validation and summary generation.

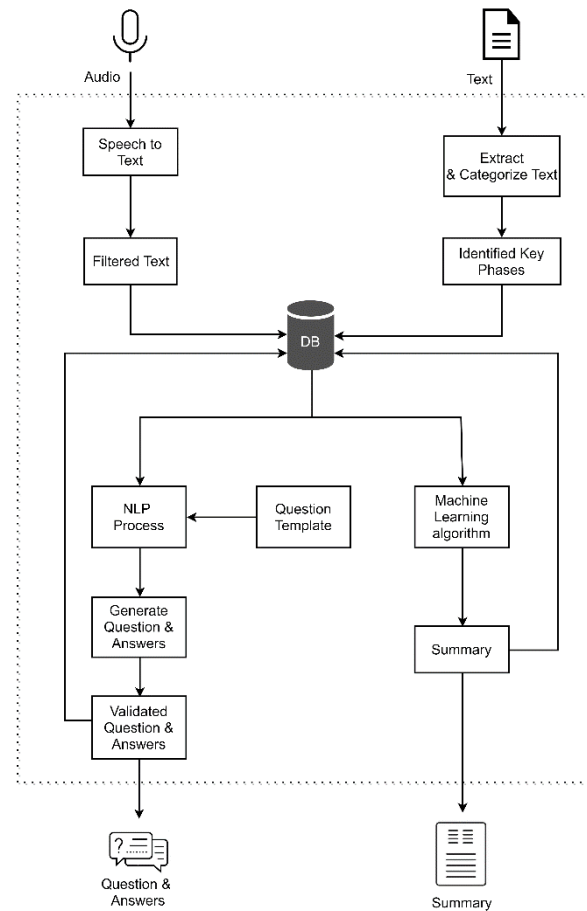


Fig 1: High-Level System Architecture

### 3.1 Text Extraction and Question Generation

This is the key step in the automatic questions and summary generation system. According to the main flow, when a user uploads a word or pdf file to the system, the text is extracted from the uploaded content, and it is processed to identify and filter the content where questions can be generated. Then the system generates “WH”, binary and gap-fill questions from the identified content and provides answers to them. The created knowledge base is used to complete this question and answer generation key procedure by text extraction. In between the unstructured input and questions and answers as the input, there are some critical functions done by the system.

Text extraction is the initial process that uploaded file should confront. In this system, the process of extracting text from the uploaded word or pdf file automatically can be introduced as the text extraction process. After extracting the text, preprocessing techniques are used to identify the most suitable phrases or sentences to generate meaningful questions with matching answers. Preprocessing can be considered as an important procedure in transforming an unstructured or semi-structured text into something useful and relevant like algorithms, etc. For the first step of sentences disuniting, sentences splitter is used. In here, the extracted text content will be tokenized into sentences and proceed into the next step of the tokenizer. Disintegrating an array of text into smaller components such as phrases, words, symbols or any other meaningful portions called tokens can be identified as the tokenization process [5]. Therefore, separated sentences will be moreover tokenized into words and proceed into the

adjacent stage of part of speech tagger. POS tagging could be useful for syntactic and semantic analysis [6]. Here, NLTK is used to map every tokenized word with the corresponding part of speech tag. Basic parts of speech in the English language are verb, noun, adjective, preposition, conjunction, adverb, etc. Specific tag for a particular word can be provided by using the POS tagger in Natural Language Tool Kit.

Sentence = [(Computer - NN), (is - VBZ), (a - DT), (fast - JJ), (electronic - JJ), (device - NN)]

NN: noun, VBZ: verb, DT: determiner, JJ: adjective

Then the system generates a question from the extracted and identified content in the uploaded file. What, when, where, and who are the words that are considered in generating WH questions. System checks for the answer phrases which have a high probability of targeting to WH questions and simply convert them to the question phrases to create meaningful questions. If the sentence is simple, the system identifies the sentence pattern and divides it into subject, verb and object according to the universal formula of a sentence in English. After the classification of words, the system checks the relationship between those words and generates suitable WH questions from the sentence. If there is a complex sentence beginning with the words of because, since, etc., those discourse connectives are highly considered in question generation. Therefore, according to the type of discourse connective, the pattern or type of the WH question to be generated can be recognized. Sentences with words or phrases like because, since and as a result, will generate question type of “Why”. Here, by transforming the definition of sentence directly to the question or converting the verb of the sentence to its negative form, yes/no type questions are generated randomly. Fill in the blanks or gap-fill questions are also produced by considering the definition of the sentence depending on text syntactic annotation, by eliminating the subject of sentence directly or any other word. Question template database which contains identified patterns of sentences can be used in this question generation procedure. Then answers for those questions are generated by considering the type of question. Answer construction for the yes/no type questions and fill in the blanks questions are a very straightforward process. When the system takes the extracted sentence as it is for the binary question, the answer is “YES”, and if it converts to the negative form, the answer is “NO”. A removed word from the sentence is directly taken as the answer in gap-fill questions, and the knowledge base is used in answer generation of WH type questions to provide an answer with high accuracy.

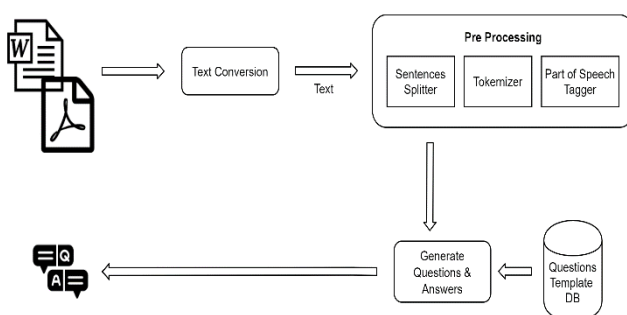


Fig 2: Text Extraction & Question Generation Process

### 3.2 MCQ Generation

The first step of the MCQ generation process is sentence selection. Many sentences are long, complex and compound. In the proposed system, first, it aims to identify meaningful

sentences. To convert the sentence into a simple format ‘Stanford CoreNLP Suite’ is used. This tool is not directly converting complex sentences into simple sentences. It provides the parse result of the example sentence in Stanford typed dependency (SD) notations. Then the system analyzes the dependency structure provided by the tool to convert it. ‘Stanford Deterministic Co-Reference Resolution System’, which is a module of the ‘Stanford CoreNLP Suite’ is used for that procedure. Finally, simple sentences from the compound sentence can be gained as output after this procedure.

Then sentences have been selected and proceed into the next step of keyword identification and MCQ generation. Classifying keywords is the main process of it and system checks for each and sentence that matches with question templates in here [7]. After the keyword identification, questions are generated by replacing that identified keyword with a blank. Extracted sentences usually start with don’t, do, must like rules described sentences patterns, the questions like choose the correct sentence [8].

Distractor Generation is another key step in the MCQ generation criteria. Distractors are also very close to the keywords. Based on the knowledge base relationships, distractors are generated for suitable keywords. If it doesn’t have knowledge base relationships for the keyword, those distractors can be identified as synonyms of the keyword. Rules described sentence patterns questions distractors are the opposite word of do, must, could like words. Therefore, the final output of meaningful MCQ questions can be generated using question rule types and knowledge base, which created according to the uploaded content.

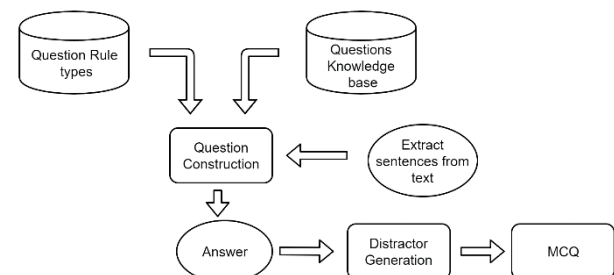


Fig 3: MCQ Generation Process

### 3.3 Question Validation and Summary Generation

The system generates more than one question from the input content after extracting the sentence from it. Therefore, the possibility of generating invalid and incorrect questions which have grammatical errors is high. But, for the best user experience, the system must provide correct and valid questions for users to make the procedure successful. For that requirement, questions must be validated using proper techniques. Under the question validation part, inappropriate questions will be removed by considering two scenarios. Therefore, invalid questions which are not relevant to the content and also the questions with grammatical errors will be removed to display the most suitable and accurate questions to the user. Most of the times, errors occur between the verb and the auxiliary verb in WH questions. When the question is checked with its noun, incorrect verbs or auxiliary verbs can be there in the questions.

What did the computer performed?

Why do humans uses computers?

Therefore, to check the above grammatical errors in the

questions, Language Check Module, which is free for use in Python. This module suggests the correct verb when it finds the grammar mistake in the question, and incorrect verb can be replaced with the correct verb can be done by using this module. Also, questions with low accuracy can be identified and removed to choose better questions that match with the generated answers. In gap fill and binary questions, there might be some adverbs, prepositions and conjunctions like also, because, but at the beginning of the question. In that kind of cases, those words must be removed to validate the question. The scenario is a bit different in checking whether the question is valid with the uploaded content or the topic. Hence, the knowledge base is used to identify appropriate questions which are associated with the relevant content.

In NLP, there are two types of text summarization methods such as extraction based summarization and subtraction based summarization [9]. Intelligent summary generator in this system uses the method of extraction based summarization and the most important key points will be identified and filtered from the uploaded text and combined them to generate a summary in that method. Suitable algorithms written in python are used in the summarization process. Transforming a textual content into something complicated like algorithms is not a straightforward process.

As the first step, the text will be extracted from the uploaded word or pdf file and proceed into the next step called pre-processing. Preprocessing techniques are widely used to identify the positions that should be filtered and converted. Sentence segmentation is the initial process that the content has to cope with. Here, content is segmented into individual sentences by identifying the end of the sentence. Stop word removal is used to remove the connecting parts of sentences except for subject, verb and object of it. Words like “the”, “a”, “an” which are commonly used as the connecting words in a sentence can be introduced as stop words, and they can be removed by comparing the text with the list of pre-mentioned stop words [10].

Input sentence = Computer is a fast electronic device

Output = Computer is fast electronic device

Stop the word “a” removed in the above example by using the removal techniques. After removing those words, the text will proceed to the next step of stemming. Stemming is a process where words are abbreviated to the root form of it by removing inflection through dropping unnecessary characters which are broadly used as a suffix. Snowball and porter are some of the available stemming models and result after stemming are very useful to identify the relationships and connections through the large set of data. Feature extraction is the most important procedure to construct a meaningful summary from the specific content. Here, word frequency checker and the sentence position checker is used in the extraction process. Position of the sentences that they located in the document can be found using the sentence position checker, and it provides guidelines to identify the most relevant sentence in the very first of document and other sub sentences. Word frequency checker or term frequency checker finds that how often a particular word displays in the document. Hence, term frequency (TF) of a term t and a document d can be calculated using the below formula [11].

$$tf(t, d) = \frac{\text{number of occurrences of term in document}}{\text{total number of all words in document}}$$

Sentences ranking based on the scoring method is the final

step of text summarization in this system. By utilizing all the results from the above processes, the score is given to the sentences according to their differences and similarities to rank them from the highest value. A summary can be generated as a final result by combining the definite number of a top-ranked sentence with the highest score.

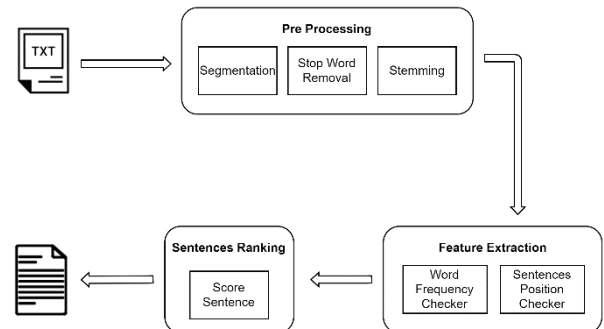


Fig 4: Text Summarization Process

### 3.4 Audio Analysis

Audio Analysis is a novel part of this system which converts speech to text by analyzing the audio input relevant to the specific course content. In here, voice signal which recorded via quality microphone can be taken and the input signal in the initial step. Then it moves to the feature extraction process. Afterwards, this audio signal should retain useful and important information of the signal, deduct redundant and information which is unwanted, show less variation from one speaking environment to another, occur normally and naturally in speech. Feature extraction will be done using MFCC technique, and feature matching will be done using Vector Quantization technique. Steps of the MFCC technique are,

1. Framing and Blocking
2. Windowing
3. FFT (Fast Fourier Transform)
4. Mel-Scale
5. Discrete Cosine Transform(DCT)

After the completion of those processes, audio input is transferred to the next procedure called decoding. The decoder is used to decode the input signal after feature extraction and will show the desired output. Previously created models can input in this decoding procedure. Therefore, Language model and the Acoustic model will be applied here. The acoustic model is an application in automatic speech recognition, which can be applied to an audio signal and the phonemes or any other linguistic units to express the relationship between them that make up speech [12].it contains statistical representations of each distinct sounds that make up a word. A Language Model is a file format used by a Speech Recognition Engine (SRE) to recognize speech [13]. It assigns a probability to a sequence of words using a probability distribution. Finally, interpreted text that given by the computer can be obtained and it will be passed as another input file to the question generation and summary generation process.

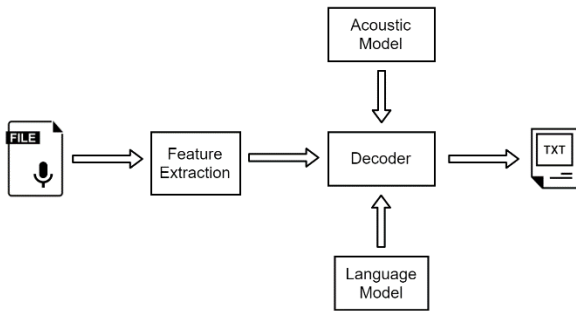


Fig 5: Audio Analysis Process

## 4. RESULTS

Different samples regarding the specific content were tested using the intelligent question and summary generator by comparing with the manual outputs according to the human choices and knowledge.

### 4.1 Questions Generation

In question generation functionality, several kinds of subjective contents were uploaded to the system and processed them to check the accuracy level and validity of the generated questions as the final output. Some of the constructed results are as below after the evaluation.

Computer can finish any given task within a very short time with high efficiency and \_\_\_\_\_  
Answer is: speed

Computer can provide correct information when correct \_\_\_\_\_ and data are given.  
Answer is: Instructions

Fig. 6. The output result of gap-fill questions

Computer can perform tasks for a long time with the same speed.  
Answer is: YES

The basic functions of a computer are not entering data, processing them and producing processed (information).  
Answer is: NO

Fig. 7. The output result of binary questions

Question: What device Computer is?  
Answer: Computer is a fast electronic device made to fulfill human needs.

Question: How can Human do complex tasks?  
Answer: Human can do complex tasks efficiently and accurately with the help of computer.

Fig. 8. The output result of WH questions

Finally, all the results that were generated via sample contents according to the question types such as WH question, binary questions, gap –fill questions and MCQ questions were put into the comparison process to obtain an exact percentage of accuracy.

Table 1. Comparison of Generated Questions

No	No of Sentences	No Of Questions	Valid Question	Invalid Questions
1	5	5	4	1
2	8	8	5	3
3	10	8	7	1
4	10	4	4	0

5	8	5	4	1
6	10	4	4	0
7	10	6	5	1
8	2	1	1	0
9	4	3	3	0
10	6	5	5	0

Total No of sentences = 73

Total No of question = 49

Total No of Correct Question = 42

Total No of Invalid Question = 7

The results in the above table proved that the system Examally is performing comparatively valid results with the accuracy of over 85%.

### 4.2 Summary Generation

Summary generation function was checked with the relevant course content, and it provided a better combination of the most important key factors as an output. It reduced the total number of sentences in the input file into a considerably low number of sentences to represent the exactness of the feature extraction procedure.

Computer is a fast electronic device made to fulfill human needs. Human can do complex tasks efficiently and accurately with the help of computer. Computer can perform tasks for a long time with the same speed. The basic functions of a computer are entering data, processing them and producing processed data (information). Computer can finish any given task within a very short time with high efficiency and speed. Computer can provide correct information when correct instructions and data are given. So the accuracy of a computer is high. Computer produces consistent output when the same input is given. Computer has an ability to store a large amount of data. It can obtain them at any given time for any process. Also, computer can act according to the given instructions. But it cannot take decisions on its own like a human being. Equipment like washing machines, mobile phones, modern motor vehicles and modern televisions are operated by computer programs. Computer programs included in such equipment are known as Embedded Computer Systems. Main components of a computer are display, speaker, keyboard, mouse, internal router and system unit. A computer is a unit which consists of many parts. We can categorize those computer parts into components according to the nature of the functions they do. Those parts are input devices, output devices, central processing unit, main memory, storage devices and communication devices.

Fig. 9. Input text for the summarization

Computer can act according to given instructions.  
Computer can perform tasks for long time with same speed.  
Computer can finish any given task within very short time with high efficiency and speed.  
Human can do complex tasks efficiently and accurately with help of computer.  
Computer is unit which consists of many parts.  
Main components of computer are display, speaker, keyboard, mouse, internal router and system unit.  
Accuracy of a computer is high.  
Computer is fast electronic device made to fulfill human needs.  
Computer can provide correct information when correct instructions and data are given.  
Computer produces consistent output when same input is given.

Fig. 10. Output text after summarization

Generated summaries were compared and evaluated to precisely measure the accuracy level of text summarization process introduced by this system.

**Table 2. Evaluation of Text Summarization**

Test	Attribute	MIN	MAX	AVERAGE
Source Text	Number of Sentences	5	95	50
	Number of Words	64	1618	841
Manual Summary	Number of Sentences	3	32	18
	Number of Words	32	512	272
System Summary	Number of Sentences	2	37	20
	Number of Words	28	619	324
Evaluation	Precision	0.4	0.89	0.65
	Recall	0.5	0.75	0.63
	F-Score	0.45	0.82	0.64
	Compression Rate	50%	71%	60.5%
	Response Time	3 sec	5 sec	4 sec

Precision =  $\frac{\text{\# of sentences in manual \& system summary}}{\text{total \# of sentences in system summary}}$

Recall =  $\frac{\text{\# of sentences in manual \& system summary}}{\text{\# of sentences in manual summary}}$

F Score =  $2 * (\text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall}))$

The results in the above table proved that the system is performing comparatively valid results with the accuracy of over 80%.

## 5. CONCLUSION & FUTURE WORKS

The core objective of this research study is to provide a better solution for the recent problem found in the education sector while studying a research problem by implementing automatic questions and summary generation system which helps both students and teachers to achieve their critical paths in exams very easily and effectively. Therefore, our effort and aim became true attestation to prove what NLP can perform. Intelligent question and summary generator system (Examally) abled to represent a major place in comparison with the existing systems, and experimental results of the research indicate that its importance by reducing the processing time and computational complexity.

After the efficacious implementation of this research project for the particular module or a topic, it will guide to the generation of questions and summary from all the modules or any content automatically. Due to the complexity of the input content in the human language, achieving more than 95% of accuracy is a challenging task. But, our system acquired that matter of contention up to a considerably high percentage, and it can further expand with the highest accuracy rate in the future. Also, the English language is the main consideration for the system and functionalities can deeper extend to make the system supports many languages as future work.

## 6. REFERENCES

[1] Dhaval Swali, Jay Palan and Ishita Shah, " Automatic Question Generation from Paragraph ", International

Journal of Advance Engineering and Research Development, Vol. 3, December 2016.

- [2] A.S.M Nibras, M.F.F Mohamed, I.S.M Arham, A.M.M Mafaris and M.P.A.W Gamage, " Automatic Question and Answer Generation from Course Materials", International Journal of Scientific and Research Publications, Volume 7, Issue 11, November 2017.
- [3] Shivank Pandey and K.C. Rajeswari, " Automatic Question Generation Using Software Agents for Technical Institutions", International Journal of Advanced Computer Research Vol. 3 No. 4 Issue. 13 December 2013.
- [4] Himansh Jethwani, Mohd Shahid Husain and Mohd Akbar, " Automatic Question Generation from Text", International Journal for Innovations in Engineering, Science and Management, Volume 3, Issue 4, April 2015.
- [5] Andrius Velykis, " Tokenization and sentence splitting", [Online] Available: <http://tint.fbk.eu/tokenization.html>.
- [6] Gianpaul Rachiele, "Tokenization and Parts of Speech (POS) Tagging in Python's NLTK library", [Online] Available: <https://medium.com/@gianpaul.r/tokenization-and-parts-of-speech-pos-tagging-in-pythons-nltk-library-2d30f70af13b>.
- [7] JunAraki, Dheeraj Rajagopal, Sreecharan Sankaranarayana n, Susan Holm, Yukari Yamakawa and Teruko Mitamura, " Generating Questions and Multiple-Choice Answers using Semantic Analysis of Texts", International Conference on Computational Linguistics: Technical Papers, pages 1125–1136, Osaka, Japan, December 11–17 2016.
- [8] Mukta Majumder and Sujun Kumar Saha, " A System for Generating Multiple Choice Questions: With a Novel Approach for Sentence Selection", Workshop on Natural Language Processing Techniques for Educational Applications, pages 64–72, Beijing, China, July 31, 2015.
- [9] Dr Michael J. Garbade, " A Quick Introduction to Text Summarization in Machine Learning", [Online] Available: <https://towardsdatascience.com/a-quick-introduction-to-text-summarization-in-machine-learning-3d27ccf18a9f>.
- [10] GeeksforGeeks, " Removing stop words with NLTK in Python", [Online] Available: <https://www.geeksforgeeks.org/removing-stop-words-nltk-python/>
- [11] William Scott, "TF-IDF from scratch in python on real world dataset", [Online] Available: <https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089>.
- [12] Samir Tanfous, "Tuning of parameters for decoding in automatic speech recognition", [Online] Available: <https://medium.com/linagoralabs/tuning-of-parameters-for-decoding-in-automatic-speech-recognition-4bf4705488c6>.
- [13] Wikipedia [Online] Available: [https://en.wikipedia.org/wiki/Language\\_model](https://en.wikipedia.org/wiki/Language_model)