An Integrated Framework for Technical Document Summarization and Multiple-Choice Question Generation

[1]Aparna T B, [1]Arun Babu K, [1]Harsha T V, [1]Soja N, [2]Asha Ali, [2]Ratheesh T K
[1]Department of Information Technology, Government Engineering College Idukki,
[2]Department of Information Technology, Government Engineering College Idukki,
[1]aparnababuraj121@gmail.com, [2]arunkathotte2522@gmail.com,
[3]harshavaryath@gmail.com, [4]soja735@gmail.com,
[5]ashaali@gecidukki.ac.in, [6]ratheeshtk@gecidukki.ac.in

ABSTRACT

Exams and assessments are going through a significant shift nowadays. The bulk of assessments are switching to MCQ-based objective tests, which are time-consuming to develop and need to administer daily. It is becoming increasingly important to have an automated MCQ generation system that is cost- as well as time-effective. Another pressing issue arose with the rapid growth of the internet is information overloading which demands systems for summarization. There are many studies being done on text summarization. As a result of the growth of online information these days, these investigations are gaining more and more popularity especially among academics as it simplifies a large text without missing the relevant information. Here we are putting up an integrated framework for summarizing a large academic technical document and for generating Multiple Choice Questions from it. The framework employs extractive text summarization and natural language processing techniques. The intention is to extract vital information from the technical documents. Automating the development of questions using AI-powered technologies is a time- and money-effective methodology that reduces the requirement for human engagement as compared to the traditional form-based method for MCQ generation. In the paper setting, a significant amount of time is saved for both summarization and MCQ generation.

Keywords: Summarization, MCQ, Natural Language Processing, Content parsing, Keyword extraction, Tokenization

INTRODUCTION

In the field of natural language processing, the term "text summarizing" refers to the practice of retaining only the most important information that best conveys the meaning of the complete block of text. The main objective is to extract a section of the text that will help us understand the meaning of the whole body of information. We create and consume data at unprecedented rates every day, making text summarization a critical requirement for many fields. We can quickly acquire the knowledge we generate according to summarization. It enables us to quickly and accurately extract a concise summary of the text that fully encapsulates the information's context. There are several uses for automatic summarization such as Media monitoring,

Newsletters, Search marketing and SEO, Internal document workflow, Books and literature, E-learning and competitive Assessments.

An important instrument for evaluating a learner's knowledge or comprehension is the question. The question is important in the assessment since it is critical to learning. Competitive examinations and assessments are going through a major transformation. The learners can understand and break down a problem more easily with the help of thoughtfully crafted questions. However, manually creating the questions takes a lot of effort and requires subject-matter expertise. An important part of this is played by automatic question generation. Consequently, the automatic creation of questions from text has become a major research field. The most common type of inquiry is the multiple-choice question. MCQ is used in different levels of educational assessment because it is effective at evaluating clearly defined knowledge and concepts that are included in the appropriate text. MCQs have a number of benefits, including quick assessment, shorter testing periods, more reliable results, and the potential for an electronic review. Many exams use computerized environments with MCQ-based question sheets.

Automatic text summarization and MCQ generation has wide scope in Teaching-Learning Process [1]. We are implementing a system which is capable of Automatic Text Summarization and MCQ Generation from Technical documents. The system will be deployed under the Academic area. We deploy an extractive method of summarization, which reassembles or highlights key sentences from the original text [7]. The MCQ question formation is done by extracting the keywords from the precise text content [3] [4] [9].

Extractive and abstractive summarization are the two categories under which text summarization may be classified. In extractive techniques [2] [6], the output summary includes text units from the source text, such as words, sentence segments, or entire sentences, following a quick analysis of the text that solely considers the syntactic level. The analysis performed by the abstractive approaches is more complete [5] [8]; for instance, a semantic analysis is used, and the summary of results may contain additional units that weren't there in the original text. These two categories involve a number of different summarizing techniques. Some of the extractive summarization techniques are K-means clustering, Summcoder, EdgeSumm, and BERT Extractive Summarization. On the other hand, structured and semantic summarization are the two basic categories for abstractive summarization techniques.

With this research, we use BERT (Bidirectional Encoder Representations from Transformers) [6] [8] [10] and NLP techniques to aid in extractive summarization. BERT for Extractive Text Summarization Using the BERT natural language model, the Extractive Summarizer summarizes a document using the main information that has been extracted. Extractive summarization is intended to reduce memory utilization while maintaining the value of the text. Extractive summarizers are used to secure content information. Through the usage of RNN, Attention mechanisms, and Transformers, BERT is used to comprehend human languages. The number of sentences and characters used in the summarizing can be adjusted using BERT extractive summarization. In order to capture a document's overall meaning, extractive summarization selects the most significant phrases and meaningful items.

RELATED WORKS

Studies of automatic text summarization and MCQ generation are well documented. The author brings some information about the background challenges and future improvements. This literature review shows some of the methods employed for Automatic text summarization, Keyword extraction and MCQ generation from technical documents.

Dhawaleswar Rao CH and Sujan Kumar Saha[1] propose an Automatic MCQ's generation system from text book contents. They discussed Generation of MCQ's from Textbook Contents of School-Level Subjects. Main key ideas that deal in this paper are Preprocessing, Sentence Selection, Key Selection and Distractor Generation. Results of a manual evaluation show that the suggested method can produce precise MCQs.

Angel hernández-castañeda , René arnulfo garcía-hernández,Yulia ledeneva, and Christian eduardo millán-hernández et.al,[2] proposed Extractive Automatic Text Summarization Based on Lexical-Semantic Keywords. The generalization issue with artificial text summarization was addressed in this paper's suggested solution. This paper proposes reducing the generalization problem in automatic text summarisation. Feature generation methods, Proximity measures and Cluster validation indexes are the main techniques used in this paper.

Chidinma A. Nwafor and Ikechukwu E. Onyenwe,[3] proposed an automatic multiple choice question generation using natural language processing techniques. This paper proposed an automatic multiple choice question generation using natural language processing techniques. Noise removal and word normalization are mentioned mainly here.

Pritam Kumar Mehta1,et al.[4] have proposed a system of Automated MCQ Generator using Natural Language Processing, According to their proposed system,They have used For producing MCQs, the BERT algorithm and sentence mapping are used. To provide choices for the questions, distractors are formed using wordnet (A lexical database for English).

Gaurav Bhagchandani,et al.,[5] have proposed a system of Abstractive Multi-Document Summarization Using Supervised and Unsupervised Learning.The word graphs, neural networks and visualizing clustering are used For producing effective Solution To Abstractive Multi-Document Summarization. Abstractive Multi-Document Summarization Using Supervised and Unsupervised Learning was the key idea behind this paper.

In the paper by Anirudh Srikanth et al.,[6] Extractive Text Summarization using Dynamic Clustering and Co-Reference on BERT is proposed. BERT model and K-means clustering are mentioned in this paper. By adjusting the summary's size in accordance with the length of the article, the system can get around the drawbacks of a brief summary. In order to create the embeddings, the input paragraph is first tokenized into sentences and then sent to the BERT model.

Rui Liu , Zheng Lin , and Weiping Wang[7] proposed Keyphrase prediction seeks to discover crucial and condensed phrases or words that can effectively capture the essential details of a

document. BiLSTM-CRF architecture and BERT Model are used in this main paper. Extractive techniques try to choose current keyphrases that precisely appear in the content.

Mayank Ramina, Nihar Darnay, Chirag Ludbe, Ajay Dhruv,[8] proposed Topic level summary generation using BERT induced Abstractive Summarization Model. The goal of this paper is to create a framework that enables users to find relevant information about a topic that they are interested in. Bidirectional Encoder Representations from Transformers (BERT) was the key idea behind the paper. Transformer encoders are stacked on top of one another to create multi-layered transformers.Generate text summary on the basis of BERT.

Chonlathorn Kwankajornkiet,et al.[9] focused on the creation of a list of questions from Thai-language input text that are ranked according to acceptability scores. The Thai input text is first divided up into clauses. Second, a collection of question words are produced once each clause has been processed. Third, to create a collection of options, all potential distractors are obtained for each question phrase with a corresponding answer. At this stage, a complete question, including a question phrase and choices, is formed. Finally, the scores from linear regression models are used to rank a set of query words and the related distractions.

Zeinab Borhanifard et.al [10] proposed a natural language understanding model with a specific named entity recognizer for shopping using a BERT transformer. BERT is a machine learning platform for NLP techniques in which the model is initially trained using wikipedia data and then finely pre-tuned using relevant training data sets.

PROPOSED SYSTEM

The proposed automatic production of summary and multiple-choice questions from the technical documents for the assessment process in the academic area is detailed in this section. By summarizing the technical contents using extractive summarization, which is improved with the aid of the BERT algorithm, the approach analyzes the contents. Additionally, the suggested system uses the different Natural Language Processing techniques to generate summary and MCQs. The process flow of the proposed system is shown in figure 2.

Using an extractive summarizer, which entails summarizing the technical document using the extracted key information, is the first step after loading the technical document. Remove all stopwords from the content after it has loaded, then convert it to single sentences. To obtain the extracted summary, merge the sentences after sorting them according to the sentence score. It uses less memory while maintaining the integrity of the contents. The number of summary sentences is controlled via extractive summarization. It always selects the strongest phrases to convey the document's overall meaning. The second phase is to create MCQs after discovering the summary. To determine the precise knowledge hidden in the sentences, tokenize and perform a semantic analysis on the list of sentences from the retrieved summary. After that, take the keywords and create questions and distractions for each of them. To select the appropriate words, significant phrases, and establish the tone of the sentence, sentence mapping is necessary. Natural Language Processing techniques to convert the declarative phrase into an interrogative form after key selection and sentence mapping. The quality of the distractors has a big impact on how well an MCQ performs. If the examinee is not sufficiently perplexed by the distractions,

they can choose the right response. By employing Word Sense Disambiguation to determine the word's right sense as a key, this system then uses Wordnet method to produce distractors.

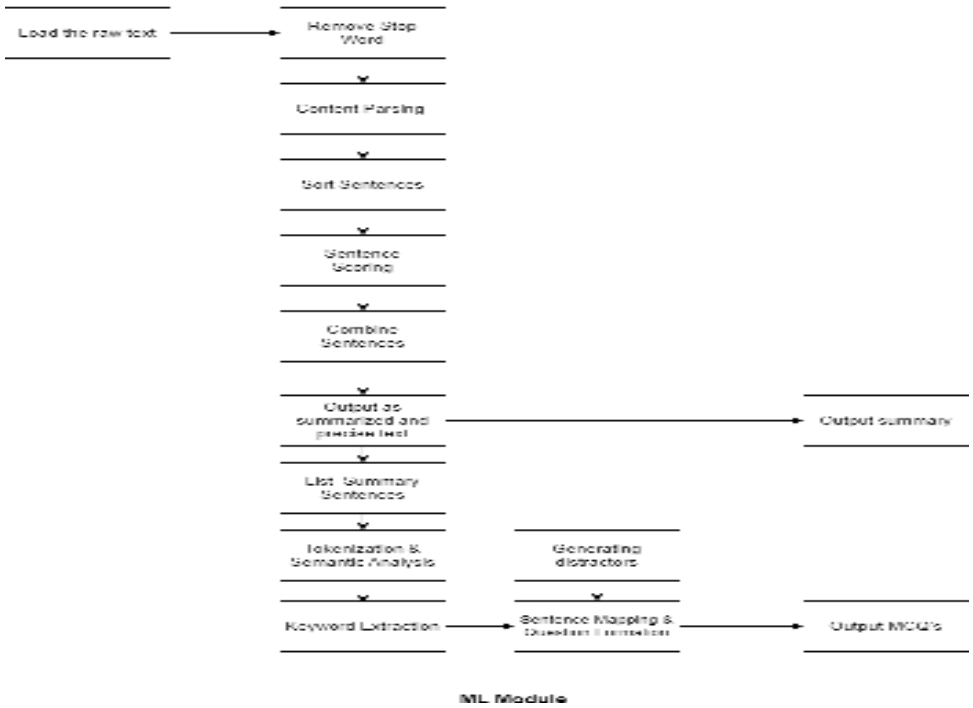

Figure 1. Block Diagram

Summarization

Summarization in this paper does not mean just paraphrasing a text. Specifically, we are focusing on extracting summaries from technical documents. The loading of Technical Content as input kicks off the summary phase. Stopwords are undesired words with no information, and the first step is to find and eliminate them from the input document. Then, determine the Word count for each word in technical content. Verify the word frequency in relation to the overall word count and the word should be marked as a Keyword if it occurs more frequently. Technical content should be interpreted into n sentences. The sentences should be added to the sentence array. Determine whether the words in each sentence in the sentence array are keywords or not. The sentence score should be increased if the word is a keyword. Sort the sentences in descending order of sentence score after giving each sentence a score. Select the first half of the sentences in the sentence array, then combine them to produce the Extracted Summary. The detailed Summarization mentioned below as Algorithm 1:

MCQ Generation

The development of MCQs based on the retrieved summary is the next phase. The Extracted Summary is used as input. Single sentences from the extracted summary are added to the sentence list. Each sentence in the sentence list should be tokenized and its meaning examined. Check to see if the sentence contains any key terms, and define them as Key. Rewrite the sentence as a question instead of a declarative form. The process of creating suitable distractors for the key comes next. Using the wordnet technique, we are creating distractors here and adding

them to the distractor list. Eventually produces MCQ output with questions and distractions. The MCQ generation is mentioned in the following Algorithm 2:

Algorithm 1

```
Input :Technical Content Tc
Output : Extracted Summary ES

Summary(Tc)
{
   For every Sw in Tc
   {
      Remove Sw
   }
   Wordcount(Tc)
   {
      For every W in Tc
      {
         if(Wc(W)>(total_num_words)/50
            Add W to Kw
      }
   }
   ContentParsing(Tc)
   {
      Convert Tc into 'n' sentences(St)
      Add St into SA[]
      For each St in SA[]
      {
         For every W in St
         {
            if (W in Kw)
               St_score++
         }
      }
   }
   SentenceSort(SA[])
   {
      Sort St in SA[] by St_score
   }
   SummaryGeneration(SA[])
   {
      Choose first half of SA[]
      Bind selected St
      Extracted summary
   }
}
```

Algorithm 2

```
Input: Extracted summary ES
Output: MCQs

MCQGeneration(ES)
{
   St_list={list all sentences from summary}
   For each St in St_list
   {
      Do tokenization and semantic analysis
      if(St has any Kt)
      {
         Set Kt as Key
         Convert St from declarative to Ques
         Return Ques
      }
      DistractorGeneration(Key)
      {
         D_list={}
         Generate distractors using wordnet
         Store distractors to D_list
         Return D_list
      }
   }
}
```

We proposed a broad extractive approach for text summarization that is built on the robust architecture of BERT and uses additional topic embedding data to direct the acquisition of contextual information. An accurate representation is crucial for a strong summary. Keywords are extracted with the PKE for MCQ generation from factual knowledge embedded within the technical document. The quality of Distractors is always an important metric for evaluating the MCQ's generated. Here, Wordnet Approach is used to generate best quality distractors for the appropriate questions generated by the system. Under the Academic Sections, this system will prove useful. Simply said, by saving the extremely valuable time, the proposed method will be extremely beneficial and time-saving for the educational institutions and academicians in many ways.

LIMITATIONS AND FUTURE SCOPE

The assessment of the summary and subject identification are the significant difficulties in a text summarization system and so for the proposed architecture. A human evaluation team is needed to evaluate the correctness of the results given by the proposed system as there is no common accessible dataset for the evaluation of the MCQs created by the system. In order to evaluate the system using human reviewers, the developers need to develop private test data. In order to process or manage multiline facts, we must employ hybrid NLP approaches.

We noticed that there are a few drawbacks with the automated MCQ creation that need to be resolved. In order to strengthen the field, we should concentrate on these problems. The capacity to handle complicated knowledge content, the development of complex distractions, standard assessment methodologies, and data from gold standard tests are among the list. MCQ from multi-line facts is another. Therefore, there are still many opportunities for further research in the field.

CONCLUSION

In order to summarize and reduce the volume of text, automatic text summarization is a prominent area of research that is widely applied and integrated into various applications. Also Competitive examinations and assessments are going through a major transformation. The creation of effective multiple-choice questions requires the use of superior distractions. It is no longer necessary for humans to create the question paper and response. Things being so, we propose an integrated framework for summarizing a large academic technical document and for generating Multiple Choice Questions from it. Our system attempts to improve existing text-summarizing methods. Additionally, it is important to develop technical document summary methods in order to create summaries that are more reliable and of higher quality. With the use of NLP, the suggested system generates automated questions that minimize human intervention. It is also a time and money-efficient method, with good distractor accuracy. This method aids students who are studying for competitive exams as well as teachers who use it for electronic evaluations. Both their problem-solving skills and concept comprehension can be tested by students using the questions. Overall with the proposed system, a significant amount of time is saved for both summarization and MCQ generation.

REFERENCES

[1] D. R. CH and S. K. Saha, "*Generation of Multiple-Choice Questions from Textbook Contents of School-Level Subjects,*" in IEEE Transactions on Learning Technologies, 2022.

[2] Á. Hernández-Castañeda, R. A. García-Hernández, Y. Ledeneva and C. E. Millán-Hernández, "*Extractive Automatic Text Summarization Based on Lexical-Semantic Keywords,*" in *IEEE Access*, vol. 8, pp. 49896-49907, 2020.

[3] Chidinma A. Nwafor and Ikechukwu E. Onyenwe "*An Automatic Multiple Choice Questions Generation using Natural Language Processing Techniques,*" in International Journal on Natural Language Computing(IJNLC) Vol.10, No.2, April 2021.

[4] Pritam Kumar Mehta1, Prachi Jain, Chetan Makwana, Dr. C M Raut "*Automated MCQ Generator using Natural Language Processing"*, in International Research Journal of Engineering and Technology (IRJET) Vol. 08 Issue: 05 May 2021.

[5] G. Bhagchandani, D. Bodra, A. Gangan and N. Mulla, "*A Hybrid Solution To Abstractive Multi-Document Summarization Using Supervised and Unsupervised Learning,*" 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019

[6] A. Srikanth, A. S. Umasankar, S. Thanu and S. J. Nirmala, "*Extractive Text Summarization using Dynamic Clustering and Co-Reference on BERT,*" 2020 5th International Conference on Computing, Communication and Security (ICCCS), 2020

[7] R. Liu, Z. Lin and W. Wang, "*Addressing Extraction and Generation Separately: Keyphrase Prediction With Pre-Trained Language Models,*" in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3180-3191, 2021

[8] M. Ramina, N. Darnay, C. Ludbe and A. Dhruv, "*Topic level summary generation using BERT induced Abstractive Summarization Model,*" *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2020

[9] C. Kwankajornkiet, A. Suchato and P. Punyabukkana, "*Automatic multiple choice question generation from Thai text,*" 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2016.

[10] Zeinab Borhanifard, Hossein Basafa, Seyedeh Zahra Razavi, "*Persian Language Understanding in Task-oriented Dialogue System for Online Shopping*", 11th International Conference on Information and Knowledge Technology (IKT) December 22-23, 2020; Shahid Beheshti University - Tehran, Iran.