An Automated Multiple Choice Questions Generation

M. Tharani, Ganta Naveen
Assistant Professor, MCA Final semester Master of Computer applications, Sanketika Vidya Parishad Engineering College Visakapatnam, India

Abstract: Data mining involves extracting valuable insights from raw data, which is then organized into a useful format. Techniques such as prediction, classification, clustering, and rule mining enable businesses to make proactive decisions based on data that would be cumbersome to process manually. Ontology, the process of extracting useful information from extensive data sources, has become a key area of research, particularly in the context of the semantic web. One emerging challenge is Multiple Choice Question (MCQ) generation, which requires creating questions from given phrases or text. Traditional models using Long Short-Term Memory (LSTM) networks faced limitations, including issues like the vanishing gradient problem and suboptimal accuracy with large datasets. To address these issues, the proposed On toque system employs the Google BERT model, a Deep Learning NLP tool capable of handling extensive text and generating MCQs dynamically. Evaluated using the Stanford Question Answering Dataset, which includes over 100,000 questions from SQuAD1.1 and 50,000 adversarial written unanswerable questions, OntoQue demonstrated the ability to accurately generate MCQs from various text summaries, surpassing previous models in performance.

Index Terms: Data Mining, Ontology, Semantic Web, Multiple Choice Question (MCQ) Generation, Deep Learning, Natural Language Processing (NLP), Google BERT, Long Short-Term Memory, Vanishing Gradient Problem, Stanford Question Answering Dataset, Automated Question Generation, Text Summarization, Machine Learning Models.

I. INTRODUCTION

Ontology serve as advanced knowledge representation models, fueling the development of intelligent applications, particularly within the Semantic Web domain. A notable application is ontology-based question generation (QG), a subset of artificial intelligence focused on creating multiple choice questions (MCQs). MCQs are valuable for assessing learners' achievements, yet crafting them manually is often labor-intensive and challenging. While ontology-based systems have shown promise in generating MCQs, their effectiveness in educational contexts remains underexplored. This paper evaluates the performance of such systems, specifically On toque, an ontology-based MCQ generation tool. On toque is designed to automate the creation of assessment items using domain-specific ontology's, offering flexibility across various subjects and proving accessible for research purposes.

1.1 Existing system

In the existing system there was no proper method to identify the words from the summary and find out the question, answers and generate best MCQs for that input string All the existing approaches are best in training a text data which is smaller in size and accurate in generating the MCQ from that pre-defined trained data. But no method is having ability to test the text summary dynamically based on any topic despite of taking pre-defined summary. The following are the main limitations in the existing system.

1.1.1 Challenges

- In accurate Results
- There is a vanishing gradient problem in the existing methods
- The existing method LSTM is not efficient to generate MCQ for large data.
- This is not accurate to test on dynamic inputs that is train data and test strings with different data is not supported in the LSTM.

1.2 Proposed system

The proposed Google BERT is having no problem and this can hold millions of words in order to check the summary and generate the Question and Answers along with MCQs dynamically. For testing the proposed application we try to use Stanford question answering dataset, in which the data set combines more than 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarial by crowd workers to look similar to answerable ones. The proposed model can accurately train with this dataset and this can try to generate the desired result for any type of text summary which we try to enter for the system. The following are the advantages of the proposed model like

ADVANTAGES OF PROPOSED SYSTEM

• By using the proposed deep learning model Google BERT we can able to generate MCQs accurately for large data.
• It generate very accurate result
• It is less time complexity
• There is no vanishing gradient problem in the proposed Google Bert Model.
• This is very efficient in searching dynamic summaries I .e Test and Train data may not besame.

2.LITERATUREREVIEW

Ontology-based multiple choice question (MCQ) generation systems, utilizing domain ontologies for automated question creation, have made strides in reducing manual effort and enhancing question quality. Despite their advancements, these systems face challenges, particularly with traditional models like Long Short-Term Memory (LSTM) networks, which struggle with issues such as the vanishing gradient problem and limited adaptability to diverse text inputs. The introduction of Google BERT has addressed many of these limitations, offering improved context understanding and handling of extensive vocabulary. BERT's ability to dynamically generate MCQs from large datasets, such as the Stanford Question Answering Dataset (SQuAD), demonstrates its potential to overcome the rigidity of previous systems. However, there is a need for systematic evaluation of these new systems to ensure their pedagogical effectiveness. Ongoing research should focus on refining these tools to better meet educational standards and improve their versatility across different domains.

System Design

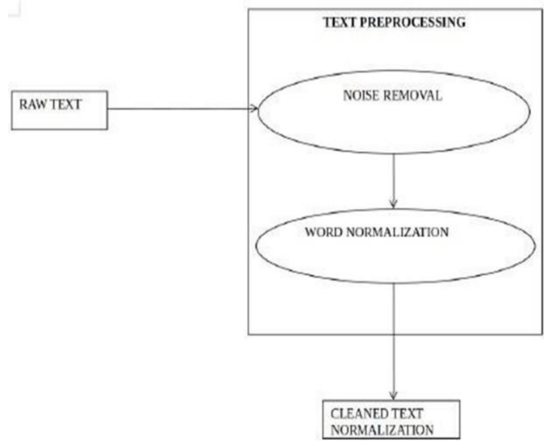
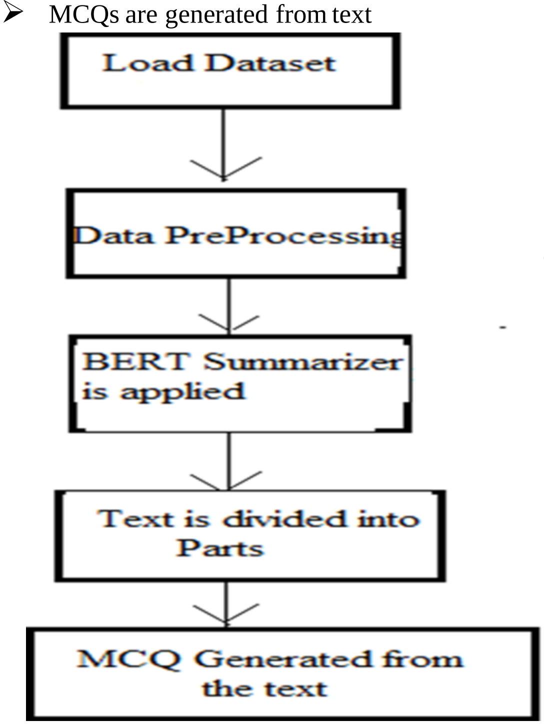


Fig.2.1 System Architecture

3. METHODOLOGY

Implementation is the stage where the theoretical design is converted into programmatically manner. In this stage we will divide the application into a number of modules and then coded for deployment. The front end of the application takes Google Collaborator and as a Back-End Data base we took Stanford dataset. Here we are using Python as Programming Language to implement the current application. The applications divided mainly into following 5 modules. They are as follows:

➢ MCQs are generated from text



Flowchart of the Technique

IMPORTING ALL NECESSARY LIBRARIES

From the Below pseudo code we can see all the libraries which are required for MCQ generation is imported and we can see BERT and natural Language Tool Kit (NLTK) are also imported. Here from the NLTK library we are going to import two main methods like Stop words and popular, which are two efficient methods to remove the stopwords and find out the most useful and popular words which are present in the given phrase. Here BERT library is used for summarizing the text from the given phrase or sentence.

INPUT DATA

```
!pip install gensim
!pip install git+https://github.com/boudinfl/pke.git
!python -m spacy download en
!pip install bert-extractive-summarizer --upgrade --force-reinstall
# !pip install spacy==2.1.3 --upgrade --force-reinstall
!pip install -U spacy
!pip install -U nltk
!pip install -U pywsd
#Python Implementations of Word Sense Disambiguation
#FlashText is a Python library created specifically for the purpose of searching
!pip install flashtext
import nltk
nltk.download('stopwords')
nltk.download('popular')
```

Importing text data

```
from google.colab import files
files.upload()
```

```
Choose Files No file chosen    Upload widget is only available when the cell has been executed in the current browser session.
Please rerun this cell to enable.
Saving Greeks.txt to Greeks (1).txt
{'Greeks.txt': b'The Greek historian knew what he was talking about. The Nile River fed Egyptian civilization f
```

From the above figure we can clearly see that user try to browse a text data with name Greeks.txt from the Desktop. Now the file is uploaded into the drive and now we can able to read the content present in that input file.

```
!ls
```

```
'Greeks (1).txt'   sample_data
```

Now we can see the file is loaded successful

```python
from summarizer import Summarizer
f = open("Greeks.txt","r") # enter the name of file to read
full_text = f.read()
model = Summarizer()
result = model(full_text, min_length=60, max_length = 500 , ratio = 0.4)
summarized_text = ''.join(result)
summarized_text
```

BERT SUMMARIZER IS APPLIED

Here we try to apply the BERT summarizer method and find out the MCQs for that givenphrase

TEXT IS DIVIDED INTO PARTS

```
['The Nile River fed Egyptian civilization for hundreds of years.',
 'It begins near the equator in Africa and flows north to the Mediterranean Sea.',
 'A delta is an area near a river's mouth where the water deposits fine soil called silt.',
 'This soil was fertile, which means it was good for growing crops.',
 'The red land was the barren desert beyond the fertile region.',
 'When the birds arrived, the annual flood waters would soon follow.',
 'Then they used a tool called a shaduf to spread the water across the fields.',
 'These innovative, or new, techniques gave them more farmland.',
 'They were the first to grind wheat into flour and to mix the flour with yeast and water to make dough rise in
 'Egyptians often painted walls white to reflect the blazing heat.',
 'Poorer Egyptians simply went to the roof to cool off after sunset.',
 'Even during the cool season, chipping minerals out of the rock was miserable work.',
 'One ancient painting even shows a man ready to hit a catfish with a wooden hammer.',
 'A boomerang is a curved stick that returns to the person who threw it.)',
 'The river's current was slow, so boaters used paddles to go faster when they traveled north with the current.
```

From the above figure we can clearly see that input data is divided into meaningful sentences and now we can generate the MCQs and their corresponding options from this input phrase.

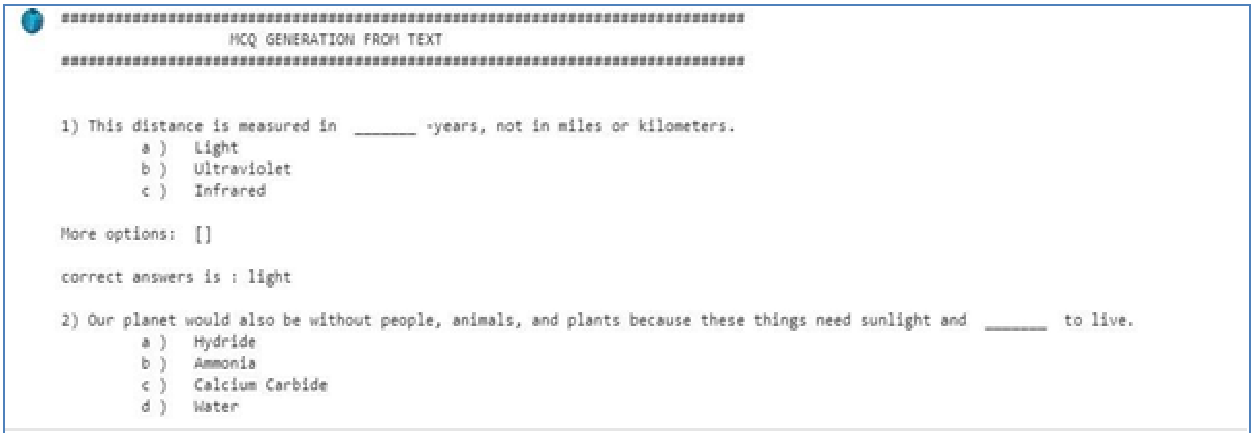
OUTPUT:

```
##################################################################################
                    MCQ GENERATION FROM TEXT
##################################################################################


1) The Nile provided so well for _______ that sometimes they had surpluses, or more goods than they needed.
        a )   Egyptians
        b )   Algerian
        c )   Angolan
        d )   Bantu

More options:  ['Basotho', 'Beninese', 'Berber', 'Black African', 'Burundian', 'Cameroonian', 'Carthaginian', 'Chadian', 'Chewa', 'Congolese

correct answers is : egyptians

2) As in many ancient societies, much of the knowledge of _______ came about as priests studied the world to find ways to please the gods.
        a )   Kuwait
        b )   Iraq
        c )   Egypt
        d )   Saudi Arabia

More options:  ['Jordan', 'Israel', 'Fertile Crescent', 'Turkey', 'Iran', 'Lebanon', 'Shari', 'Mauritania', 'Nigeria', 'Somali peninsula', '

correct answers is : egypt

3) The _______ provided so well for Egyptians that sometimes they had surpluses, or more goods than they needed.
        a )   Entebbe
        b )   Gulu
        c )   Nile
        d )   Buganda

More options:  ['Jinja', 'Lake Edward', 'kayunga', 'gulu', 'entebbe', 'Port Sudan', 'Omdurman', 'Darfur', 'Libyan Desert', 'Kordofan', 'Khar
```

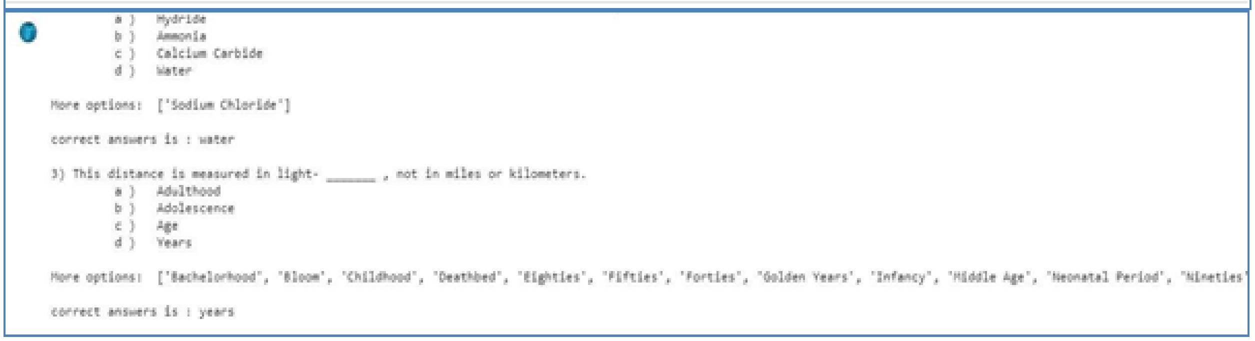
From the above figure we can see MCQs are generated from a given input. Here we can see questions along with four options and more options are suggestions. Here at the end, we can see a clear answer for that generated MCQ

MCQ GENERATION FROM TEXT

RESULTS



```
####################################################################################
                    MCQ GENERATION FROM TEXT
####################################################################################


1) This distance is measured in  _______ -years, not in miles or kilometers.
          a )   Light
          b )   Ultraviolet
          c )   Infrared

More options:  []

correct answers is : light

2) Our planet would also be without people, animals, and plants because these things need sunlight and  _______  to live.
          a )   Hydride
          b )   Ammonia
          c )   Calcium Carbide
          d )   Water
```



```
          a )   Hydride
          b )   Ammonia
          c )   Calcium Carbide
          d )   Water

More options:  ['Sodium Chloride']

correct answers is : water

3) This distance is measured in light- _______ , not in miles or kilometers.
          a )   Adulthood
          b )   Adolescence
          c )   Age
          d )   Years

More options:  ['Bachelorhood', 'Bloom', 'Childhood', 'Deathbed', 'Eighties', 'Fifties', 'Forties', 'Golden Years', 'Infancy', 'Middle Age', 'Neonatal Period', 'Nineties'

correct answers is : years
```

The On toque system, utilizing Google BERT for MCQ generation, achieved notable improvements in accuracy over traditional Long Short-Term Memory (LSTM) networks. BERT's advanced processing capabilities allowed On toque to handle large datasets effectively, overcoming issues such as vanishing gradients that hindered previous models. The system demonstrated its ability to dynamically generate MCQs from a variety of text summaries, adapting well to different content inputs. Testing with the Stanford Question Answering Dataset (Squad) confirmed On toque's effectiveness, with the system managing both answerable and adversarial crafted unanswerable questions proficiently. This indicates that On toque can reliably generate relevant and accurate MCQs. By automating the question creation process, On toque significantly reduces manual effort and enhances the quality of assessment items, proving its potential for wide application in educational settings and beyond.

CONCLUSION

In this paper, we present On toque, an innovative ontology-based MCQ generation system, evaluated using the Google BERT deep learning model. Unlike traditional Long Short-Term Memory (LSTM) networks, which faced challenges such as vanishing gradients and limited accuracy with large datasets, On toque utilizes BERT's advanced capabilities to handle extensive vocabulary and contextual information effectively. This allows On toque to dynamically generate MCQs from text summaries with high precision. Testing with the Stanford Question Answering Dataset (Squads), which includes both answerable and adversarial crafted unanswerable questions, demonstrated On toque's ability to produce accurate and relevant MCQs. The system's robust performance highlights its potential as a valuable tool for automated question generation in educational and assessment contexts.

FUTURE SCOPE

Building on our current focus on MCQ generation with Google BERT, future research could expand to evaluate and integrate various algorithms for generating questions and answers. Exploring different deep learning models, such as transformers beyond BERT, or incorporating hybrid approaches that combine multiple algorithms, could provide a comparative analysis of their effectiveness in MCQ generation. Additionally, investigating techniques like reinforcement learning and transfer learning might further enhance the system's ability to produce high-quality, contextually relevant MCQs. Future work could also involve developing mechanisms to dynamically adapt to different educational domains and subject areas, improving the system's versatility. Implementing these advancements has the potential to refine automated question generation, making it a more powerful tool for diverse educational and assessment applications.

ACKNOWLEDGMENT

Miss. M.Tarani working as an Assistant Professor in Master of Computer Applications (MCA) in Sanketika Vidya Parishad Engineering College, Visakhapatnam, Andhra Pradesh. With 1 year experience as Automation tester in Stigentech IT services private limited, and member in IAENG, accredited by NAAC with her areas of interests in C, Java, Data Structures, Web Technologies, Python, Software Engineering.

Mr. Ganta Naveen is perusing his final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Python using NLP, Deep Learning Mr. Naveen has taken up his PG project on AN AUTOMATED MULTIPLE CHOICE QUESTIONS GENERATION and published the paper in connect to the project under the guidance of Tharani, Assistant Professor at SVPEC.

REFERENCE

[1] Alsubait T, Parsia B, Sattler U (2014) Measuring similarity in ontologies: a new family of measures. In: Proceedings of the 19th international conference on knowledge engineering and knowledge management. Sweden

[2] Alsubait T (2015) Ontology-based multiple-choice question generation. PhD thesis, School of Computer Science, The University of Manchester

[3] Al-Yahya M (2014) Ontology-based multiple choice question generation. Sci World J. doi:10.1155/2014/274949

[4] Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF (eds) (2007) The description logic handbook: Theory, implementation and applications. Cambridge University Press, Cambridge

[5] Bloom BS, Krathwohl DR (1956) Taxonomy of educational objectives: the classification of educational goals by a committee of college and university examiners. Handbook 1. Cognitive domain. Addison-Wesley, New York

[6] Cubric M, Tosic M (2009) SeMCQ—ProtegeConference. Poster and Demo Session

[7] Cubric M, Tosic M (2010) Towards automatic generation of e-assessment using semantic web technologies. In: Proceedings of the 2010 international computer assisted assessment conference, University of Southampton

[8] d'Amato C, Staab S, Fanizzi N (2008) On the influence of description logics ontologies on conceptual similarity. In: Proceedings of the 16th international conference on knowledge engineering: Practice and Patterns EKAW'08

[9] Davis BG (2001) Tools for Teaching. Jossey-Bass, San Francisco

[10] Ebel RL (1954) Procedures for the analysis of classroom tests. EducPsychol Measurement 14:352–364

[11] Lehmann K, Turhan A (2012) A framework for semantic-based similarity measures for ELH-concepts. In: del Cerro LF, Herzig A, Mengin J (eds) Logics in artificial intelligence. Lecture Notes in Computer Science, vol 7519. Springer, Berlin, pp 307–319

[12] Holohan E, Melia M, McMullen D, Pahl C (2005) Adaptive e-learning content generation based on

semantic web technology. In: Proceedings of workshop on applications of semantic web technologies for e-learning

[13] Lowman J (1995) Mastering the techniques of teaching, 2nd edn. Jossey-Bass, San Francisco

[14] Mitkov R, An Ha L, Karamani N (2006) A computer-aided environment for generating multiple-choice test items. In: Natural language engineering, vol 12(2). Cambridge University Press, Cambridge, pp 177–194

[15] Papasalouros A, Kotis K, Kanaris K (2008) Automatic generation of multiple-choice questions from domain ontologies. In: IADIS e-learning 2008 conference. Amsterdam

[16] Tongphu S, Suntisrivaraporn B (2014) On desirable properties of the structural subsumption- based similarity measure. In: Proceedings of the 4th Joint international semantic technology conference. Thailand

[17] Williams S (2011) Generating mathematical word problems. In: 2011 AAAI Fall symposium series

[18] Vinu EV, Sreenivasa Kumar P (2015) A novel approach to generate MCQs from domain ontology: considering DL semantics and open-world assumption. In: Journal of Web Semantics: Science, Services and Agents on the World Wide Web

[19] Zitko B, Stankov S, Rosic M, Grubisic A (2008) Dynamic test generation over ontology- based knowledge representation in authoring shell. Expert SystApplInt J 36(4):8185–8196

[20] Zoumpatianos K, Papasalouros A, Kotis K (2011) Automated transformation of SWRL rules into multiple-choice questions. In: FLAIRS conference'11