

Automatic Quiz Generation

Ishita Mandal
MT21029
ishita21029@iiitd.ac.in

Lasani Hussain
MT21042
lasani21042@iiitd.ac.in

Sanchita Saha
MT21075
sanchita21075@iiitd.ac.in

Shaguftha Zuveria Kottur
MT21079
shaguftha21079@iiitd.ac.in

Madiha Tariq
MT21125
madiha21125@iiitd.ac.in

1 UPDATED PROBLEM STATEMENT

Quizzes are a great way to test and aid a student's understanding of a topic, as demonstrated by various studies [2, 8, 10]. Setting up quizzes requires considerations such as relevance to the topic, unambiguity, originality of the questions, and generation of distractors (choices) for MCQs. So, setting up quizzes is a time-consuming task and more so since quizzes are conducted regularly. This project aims to determine the level of automation which can be introduced into the task of quiz generation and to build an end-to-end system which can generate different formats of questions (MCQ, short answer).

2 MOTIVATION

The task of automatic quiz generation has two target audiences; students and instructors. For instructors, the existence of such a system allows the manual task of quiz generation to be offloaded to the model, which ultimately allows them to focus their time on other aspects of the pedagogical process.

For students, this system can be used, independent of the involvement of any instructor. Thus basically serving as a tool to help them quiz themselves and evaluate their understanding of any material they are studying. For this, the model would require the additional feature of answer checking. Most question banks available to the student would have repeated questions, and questions maybe outside the material they are studying. The existence of a system which generates questions based on the material provided, would ensure the generation of new/original questions based only on the material provided.

3 UPDATED LITERATURE REVIEW

Several works have been done in the field of Question Generation (QG) and Question Answering (QA). Cheng et al.[2] define the difficulty level of a question as the number of inference steps required to answer it. Existing QA systems perform substantially worse in answering multi-hop questions than single-hop ones [16]. To achieve DCQG with the above definition, QG model should have strong control over the logic and reasoning complexity of generated questions, for which graph-based models are better suited. The authors first transform the given raw text into a context graph, from which answers/reasoning chains are sampled. A question generator and a question rewriter generates an initial simple question and step-by-step rewrites it into more complex ones. HotpotQA [16] dataset was utilized where most questions require two inference steps to answer and can be decomposed into two 1-hop questions. Having learned how to rewrite 1-hop questions into 2-hop ones with this dataset, this framework can easily extend to the generation

of (n+1)-hop questions from n-hop ones. BLEU3, BLEU4, METEOR and CIDEr metrics were used for automatic evaluation.

Gao et al. [6] introduced the concept of difficulty controllable question generation for reading comprehension. For their work they classified the SQuAD dataset into Easy and Hard question by performing automatic labeling protocol. For the actual model, they first converted the input to appropriate representation using bidirectional LSTM encoder. Then the decoder generates the actual question considering the difficulty.

The authors in [1] focussed on the generation of sophisticated high-level question from multiple sentences rather than factoid questions which are retrieved from single sentences. This helped them in generating questions that were more semantically oriented. To achieve this objective they used the ProcessBank corpus, which comprises of 200 paragraphs from high school biology textbooks containing information on biological processes along with some multiple choice questions per paragraph. Their methodology to generate question and phrase level distractors was heavily dependent on the question template and the text annotation (such as entity, event triggers, event-event relations) done by the domain experts. The model generated 200 questions along with 3 distractors per question and its significance was evaluated by two human annotators.

Another aspect that holds equal importance while generating questions is the relevance of the distractors being generated. To study this relevance, Liang et al., 2018[9], performed comparative and comprehensive analysis of various feature-based, neural network based and unsupervised models. The models are trained and tested with two datasets viz. SciQ dataset(13.7k MCQs) and self compiled MCQL dataset(7.1k MCQs). A two stage cascaded framework is proposed in order to make the process of ranking effective and efficient. While the first stage ranker significantly reduces the candidate size, the job of the second stage ranker is to learn from the top predictions of the first stage. They concluded that a 2-stage framework with feature-based ensemble methods (LambaMART and Random Forest) as second stage, yielded better results in comparison to other models based on evaluation metrics like recall@10, precision@1, precision@3, MAP@10 and mean reciprocal rank.

In the paper by Gao et al. [7], their goal is to generate context and question related, grammatically consistent wrong options, i.e. distractors, for the question, given an article, a pair of question and its correct option originated from the article. The dataset used by them is RACE. They randomly divide the dataset into the training (80%), validation (10%) and testing sets (10%). The framework used by them is hierarchical encoder-decoder. In encoder part they obtain hierarchical contextualized representations for the whole article,

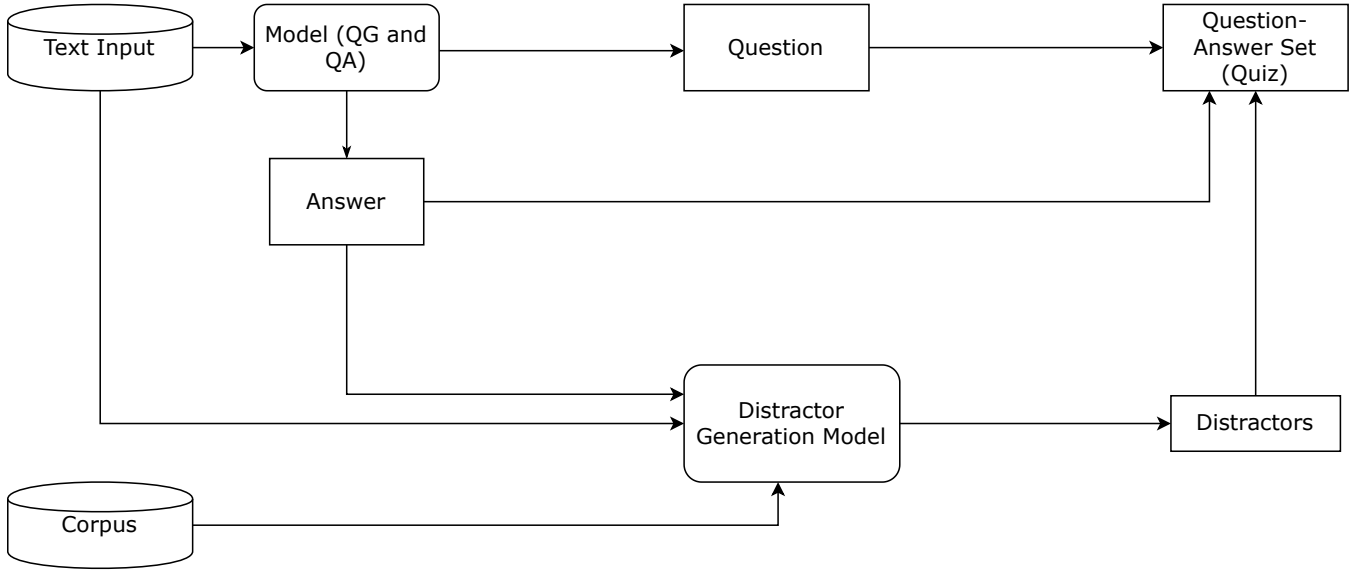


Figure 1: High Level Abstract Model

namely, word-level representation and sentence-level representation. In decoder part, they employ a language model to compress the question information into a fixed length vector to initialize the decoder state, making the distractor grammatically consistent with the question. To evaluate the similarity of generated distractor with the ground truth they employ BLEU and ROUGE scores to evaluate the similarity. The limitation of this paper is that answer should be provided as input.

The paper by Katherine Stasaski et al.[14] aims to generate cause-and-effect questions, given text as input, which test knowledge of the relationship between these two events or states. For this they use, pre-existing Causal extraction systems, which uses a series of syntactic rules to extract causes and effects from unstructured text. They pass two or three sentence passages into the causal extractor, since a cause and effect may span across adjacent sentences. Multiple causal relationships may be found in a single passage; the sliding window captures the multiple relationships across different passages. For evaluation, they use crowd workers. The only drawback is that they generate cause and effect relationship for three passage at a time, but what if relationship spans for more than three passage at a time.

Question answer pairs from multi paragraph text was explored by Roemmele et al. [12]. They ranked the topmost 4 relevant paragraphs, followed by encoder-decoder system for question answering and question generation.

The paper by Ho-Lam Chung et al.[3] tried to overcome the existing problem of distractor generation by presenting a new distractor generation scheme which removes major two drawback of previous distractor generation method i.e the poor quality of distractor and generation of single distractor instead of multiple. To improve the quality of distractor author has used BERT model. The input representation to BERT is done by summing corresponding token, segment and position embedding. After this input embeddings

travel through fine tuned BERT for prediction. For distractor generation input given are paragraph, answer and question. The BDG model takes multiple iterations to generate distractor. To enhance the quality of BDG, author jointly train BDG and a parallel MLM by using multitasking loss function. To make sure that generated distractors are not similar, answer negative regularization has been used. The core structure for distractor generation is based mainly on sequential recurrent MLM decoding mechanism. MRC model has been used for ranking/selecting distractor set. The input given is passage, question and answer and goal is to find triplet to form option set such that entropy function is maximized. Race dataset has been used. For performance evaluation BLUE and ROUGE scores are used.

4 DATASET

Three datasets have been used for question-answer and distractor generation. **SQuAD + NewsQA** [11] dataset has been used for question-answer pair generation and Gao’s[7] processed **RACE** dataset in which any semantically irrelevant distractors were removed from the original dataset. There are two version of SQuAD, the model we are using uses SQuAD-2.0 that contains a total of 130,319 and 11,873 samples respectively in train and test set. NEWSQA contains 107,674 and 5,988 train and test items, respectively. The processed RACE dataset has 96,501 training samples and 12,284 test samples.

5 PROTOTYPE/ SYSTEM

We have tried two approaches for MCQ generation. In the first approach, the input text was first summarized using BERT extractive summarizer to find important and potential sentences from which question can be framed. Next, keyword extraction is done from the original text using python’s Multipartite graph keyphrase extraction model and only those keywords were kept that were present

in the above extracted summarizer. This was followed by sentence mapping in which the sentence which contained that keyword was retrieved. Finally, the distractors were generated using Conceptnet. General pre-processing steps like removing stopwords, tokenization, conversion to lower case etc were used. However, the scope of this approach is limited to generation of one word distractors only and the length of the input text can't be too long.

The second approach involves MCQ generation from multi-paragraph text and their corresponding sentence level distractors. This uses a pipeline of two models, first being BERT (Bidirectional Encoder Representations from Transformers) model which does the Question Generation(QG) and Question Answering(QA), while the second model is BDG (BERT-based Distractor Generation) which is used for distractor generation.

Here, we are proposing a web-based end to end system which accepts the text from which MCQ is to be generated and outputs the questions and its corresponding answer and distractors.

Code and models can be found here. Figure 2 and Figure 3 show the proposed UI for our application.

6 PROPOSED SOLUTION

We present an end-to-end system that applies QA and QG to multi-paragraph documents for the purpose of user content understanding [13]. The existing approach of predicting indices of answer spans in the text, performs well only when reference text is limited to a single paragraph. The authors address the task of performing QA for multiple-paragraph documents By adapting an existing method to additionally leverage a pre-trained text encoding model. They have adapted Clark and Gardner's [5] shared-normalization approach by replacing their GRU BiDAF encoder with the BERT-BASE-UNCASED encoder. Also they evaluate QA with reference to a single document, unlike Wang et al. [15] used a similar approach for open-domain QA, where answers are mined from the entirety of Wikipedia.

Authors here [4] first propose to jointly train BDG and a parallel MLM (P-MLM), architecture for distractor generation to enhance the quality of BDG. They find that original BDG scheme may overfit in sentence writing and underfit in learning in passage semantics, they propose and validate through experiments that incorporation of multi task learning setting prevents potential overfitting. They also propose a loss(answer negative loss) to penalize predicting tokens in answer set A when predicting distractors, expecting to generate distractors to use words different from A. Existing work selects results on different beam search paths to generate multiple distractors, this lowers power of distracting a reader for MCQ preparation. Authors propose to select distractor set by considering semantic diversity by incorporating a multi-choice reading comprehension (MRC) model for ranking/selecting distractor sets. M_{MRC} takes a passage P , a question Q , and a set of options (including an answer A and distractors D_1, D_2, \dots, D_n) as input and outputs $[p_A, p_{D_1}, \dots, p_{D_n}]$ as the answer probabilities of the options. M_{MRC} is trained by maximizing the answer probability p_A while minimizing the probabilities $[p_{D_1}, \dots, p_{D_n}]$.

The **novelty** of this project lies in generating MCQs as well as short answer type questions from multi-paragraph text as input.

REFERENCES

- [1] Jun Araki, Dheeraj Rajagopal, Sreecharan Sankaranarayanan, Susan Holm, Yukari Yamakawa, and Teruko Mitamura. 2016. Generating Questions and Multiple-Choice Answers using Semantic Analysis of Texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 1125–1136. <https://aclanthology.org/C16-1107>
- [2] Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. 2021. Guiding the Growth: Difficulty-Controllable Question Generation through Step-by-Step Rewriting. In *ACL/IJCNLP*.
- [3] Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. 2020. A BERT-based distractor generation scheme with multi-tasking and negative answer training strategies. *arXiv preprint arXiv:2010.05384* (2020).
- [4] Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. 2020. A BERT-based Distractor Generation Scheme with Multi-tasking and Negative Answer Training Strategies. <https://doi.org/10.48550/ARXIV.2010.05384>
- [5] Christopher Clark and Matt Gardner. 2017. Simple and Effective Multi-Paragraph Reading Comprehension. <https://doi.org/10.48550/ARXIV.1710.10723>
- [6] Yifan Gao, Lidong Bing, Wang Chen, Michael R Lyu, and Irwin King. 2018. Difficulty controllable generation of reading comprehension questions. *arXiv preprint arXiv:1807.03586* (2018).
- [7] Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R Lyu. 2019. Generating distractors for reading comprehension questions from real examinations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6423–6430.
- [8] Jennifer A Jones. 2019. Scaffolding self-regulated learning through student-generated quizzes. *Active Learning in Higher Education* 20, 2 (2019), 115–126.
- [9] Chen Liang, Xiao Yang, Neisarg Dave, Drew Wham, Bart Pursel, and C. Lee Giles. 2018. Distractor Generation for Multiple Choice Questions Using Learning to Rank. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, Louisiana, 284–290. <https://doi.org/10.18653/v1/W18-0533>
- [10] Annelies Raes, Pieter Vanneste, Marieke Pieters, Ine Windey, Wim Van Den Noortgate, and Fien Depaepe. 2020. Learning and instruction in the hybrid virtual classroom: An investigation of students' engagement and the effect of quizzes. *Computers & Education* 143 (2020), 103682.
- [11] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2383–2392. <https://doi.org/10.18653/v1/D16-1264>
- [12] Melissa Roemmele, Deep Sidhpura, Steve DeNeefe, and Ling Tsou. 2021. AnswerQuest: A system for generating question-answer items from multi-paragraph documents. *arXiv preprint arXiv:2103.03820* (2021).
- [13] Melissa Roemmele, Deep Sidhpura, Steve DeNeefe, and Ling Tsou. 2021. AnswerQuest: A System for Generating Question-Answer Items from Multi-Paragraph Documents. <https://doi.org/10.48550/ARXIV.2103.03820>
- [14] Katherine Stasaski, Manav Rathod, Tony Tu, Yunfang Xiao, and Marti A. Hearst. 2021. Automatically Generating Cause-and-Effect Questions from Passages. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Online, 158–170. <https://aclanthology.org/2021.bea-1.17>
- [15] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. <https://doi.org/10.48550/ARXIV.1804.07461>
- [16] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2369–2380. <https://doi.org/10.18653/v1/D18-1259>


QuizGen

Enter Quiz Material

When you're having a holiday , one of the main questions to ask is which hotel or apartment to choose . However , when it comes to France , you have another special choice : tree houses . In France , tree houses are offered to travelers as a new choice in many places . The price may be a little higher , but you do have a chance to fulfill your childhood memories . Alain Laurens , one of France's top tree house designers , said , "Most of the people might have the experience of building a den when they were young . And they like that feeling of freedom when they are children." Its fairy-tale style gives travelers a special feeling . It seems as if they are living as a forest king and enjoying the fresh air in the morning . Another kind of tree house is the star cube . It gives travelers the chance of looking at the stars shining in the sky when they are going to sleep . Each star cube not only offers all the comfortable things that a hotel provides for travelers , but also gives them a chance to look for stars by using a telescope . The glass roof allows you to look at the stars from your bed .

Generate Questions

Figure 2: Prototype Input



Generate Questions

ABOUTGENERATE QUIZ

Q: What is one of the main questions to ask?

A: which hotel or apartment to choose

D: How to choose a new choose

Q: What do you're having?

A: a holiday

D: a new house

Q: What are tree houses offered to travelers as?

A: a new choice

D: a new house

Q: What may be a little higher?

A: The price

D: The time you tree houses

Q: Why do you have a chance?

A: to fulfill your childhood memories

D: to see the new house in the house

Q: What might Laurens have?

A: the experience of building a den when they were young

D: the experience of living a new house

Q: What is Laurens?

A: one of France's top tree house designers

D: one of the best visitors in America

Figure 3: Prototype Output