Automatic QnA Generation from YT Videos

Mohammad Aflah Khan (2020082)
aflah20082@iiitd.ac.in

Neemesh Yadav (2020529)
neemesh20529@iiitd.ac.in

1 Introduction

According to scholarly observations, the modern age of fast-paced and concise media has given rise to a noticeable decrease in human attention span. This decrease in attention span has significant implications in education, where learners frequently disengage from lengthy lectures and may struggle to recall previously covered material. To address this challenge, one potential mitigation strategy is the implementation of a feedback loop, which presents learners with a set of questions at regular intervals to help them identify and reinforce their retention of the material. This approach is particularly beneficial for young learners and individuals seeking to learn on-the-go.

Several proprietary software options are available that offer similar functionality to the feedback loop approach. For instance, the Zapiens AI Question Maker is capable of generating questions and answers from any text. However, access to this software is restricted to paid users only. Another comparable option is Quetab AI, which also offers limited API availability. However, like Zapiens, this software is also paywalled and requires a subscription to access its full capabilities. Nonetheless, these software options are inadequate, as they are primarily designed for educators to build question papers and take in text content only.

In summary, the decrease in attention span resulting from the prevalence of fast-paced and brief media has implications in the field of education. One potential solution to this challenge is the implementation of a feedback loop that presents learners with questions at regular intervals. While several software options are available that offer similar functionality, they are primarily designed for educators and require a subscription to access their full capabilities, making them unsuitable for learners seeking a more accessible solution.

2 Problem Statement

This study aims to implement deep learning (DL) models for generating question-and-answer pairs from YouTube videos using textual features such as transcripts or transcriptions generated by tools like Whisper. Additionally, the study aims to develop a user interface that provides interactive prompts and real-time feedback to users while watching videos on YouTube. This user interface will also allow users to engage in discussions with a chatbot that can answer questions about the video being watched. Our current objective is to conduct a human evaluation for our final application and evaluate the Deep Learning (DL) models intended for use via intrinsic evaluation techniques utilizing metrics such as the contextual similarity between the generated questions and answers in relation to the corresponding video context. The overarching aim is to develop an end-to-end user application that can take in a YouTube video link and generate unique question-answer pairs using the transcript of the video, which are presented at fixed intervals. There are no such works that exist, and several works take on different aspects of this project in isolation while we focus on creating an end-to-end pipeline for the same.

3 Related Work

Several research studies have focused on generating questions from text or video content. One such study proposed a neural question generation approach that generates answer-aware input representations using an encoder-decoder architecture (Zhou et al., 2018). The study demonstrated that the proposed approach could generate fluent and diverse questions without relying on rigid heuristic rules. Another study (Yuan et al., 2017) proposed a machine comprehension model that employs supervised and reinforcement learning to generate high-quality questions that can benefit from

a question-answering system's performance. The proposed model is trained and evaluated on the SQuAD dataset.

A review survey (Zhang et al., 2021) of existing models for question generation analyzed the underlying ideas, major design principles, and training strategies, from traditional rule-based methods to advanced neural network-based methods. The survey provides a valuable reference for researchers in question generation and identifies promising future directions. In another study, the authors proposed a multi-task learning framework for Turkish question answering and generation tasks (Akyon et al., 2021) using a fine-tuned multilingual T5 transformer. The proposed approach streamlines the generation of exam-style questions and achieves state-of-the-art Turkish QA and QG performance on various datasets.

A novel approach called Video Question-Answer Generation (VQAG) was introduced in another study, which generates question-answer pairs based on videos (Su et al., 2021). The proposed network includes Joint Question-Answer Generator (JQAG) and Pretester (PT) components and achieved state-of-the-art performance, outperforming supervised baselines using generated questions only.

Finally, a recent study (Lopez et al., 2021) proposed a transformer-based fine-tuning technique for question generation in NLP that outperformed previous RNN-based Seq2Seq models and performed on par with Seq2Seq models that employ answer-awareness and other special mechanisms. The study analyzed various factors that affect the model's performance, such as input data formatting, the length of the context paragraphs, and the use of answer-awareness. Additionally, the study identified possible reasons why the model fails.

Despite the potential benefits of video-based question generation methods, we decide against using them due to their computational complexity and the high accuracy of transcript-based representations for the educational videos we plan to target.

4 Dataset

The Baseline Experiments comprised the selection of five educational YouTube videos from the "CrashCourse" channel, chosen randomly and without relevance to the research endeavor. These videos were deemed suitable for the task at hand due to their inherent level of detail and structural organization. Subsequently, the transcriber tool provided by AssemblyAI was utilized to generate transcripts for each of the selected videos, which in turn formed the basis for conducting the experiments.

5 Baseline Implementation

5.1 Video Transcription

In order to transcribe videos, we utilized the API provided by AssemblyAI. This was deemed necessary due to the absence of captions in certain videos, as well as the existence of errors within captions which cannot be rectified without manual intervention. One common error pertained to the amalgamation of multiple words into a single word, for which automated methods yielded suboptimal accuracy. Nonetheless, despite the utility of the API, certain challenges were encountered. Specifically, when confronted with Non-English names, the API produced some minor inaccuracies. For example, the name "Aurangzeb" was transcribed as "Orangzeb".

5.2 End-to-End Question Generation (Answer Agnostic) - B1

This approach is based on the end-to-end methodology proposed in the work of Lopez et al. (Lopez et al., 2021). Specifically, we employ the pre-trained "valhalla/t5-small-e2e-qg" model from HuggingFace, which is tailored for the task of end-to-end question generation. To execute the model, we rely on sample code provided by the original repository[1]. Notably, this approach is answer-agnostic, implying that it does not rely on any information about the answer to generate questions. Nonetheless, this method cannot distinguish irrelevant filler text from the essential content. Consequently, it generates questions even for introductory sections and user interactions in videos, which do not contribute to the meaningful content.

5.3 Traditional Linguistics Based Question Generation - B2

This baseline is based on traditional linguistics and employs hardcoded rules and parsers to transform sentences into questions. An existing implementation provided at[2] is used in our study. However,

[1] https://github.com/patil-suraj/question_generation

[2] https://github.com/dipta-dhar/Automatic-Question-Generator

Method	Avg. Adequacy	Avg. Fluency	Avg. Relevance
Our Method	**0.95714**	**4.7**	**4.6**
B1	0.9733	4.3601	3.9067
B2	0.8267	2.6267	2.8133

Table 1: Human Evaluation Results averaged out per method over all the paragraphs. Average Adequacy is out of 1, and Average Fluency and Relevance are out of 5.

one major limitation of this approach is that it cannot provide the correct answer since it can only generate questions. Moreover, its performance is suboptimal due to its reliance on linguistic rules which do not cover the full range of input data, particularly when dealing with unstructured data.

6 Our Approach

6.1 Video Transcription

Upon the implementation of the baseline models, it became apparent that the utilization of a video-transcription service constitutes a bottleneck in terms of both time and compute resources. Consequently, the decision was made to discontinue its use and instead concentrate solely on video content possessing English transcriptions, whether they be automated or user-uploaded. Subsequently, the extracted text will be employed to generate relevant questions. Hence our input is this transcript file.

6.2 Front End

In the front-end architecture, the StreamLit application framework, specifically designed for machine learning engineers, is utilized. Its implementation enables the creation of a two-page setup. The initial page serves as a landing page, where the user is prompted to provide the relevant link. Upon completion of the pre-processing phase, the user is automatically redirected to the second page. Herein, a video player and a corresponding set of questions are presented to the user.

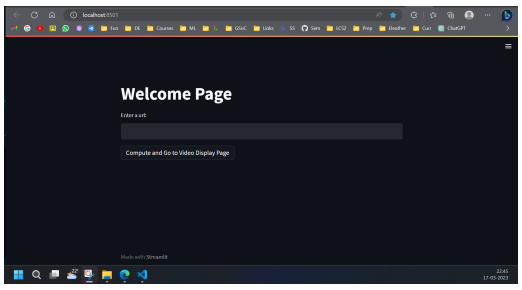

Figure 1: Landing Page

Figure 2: Result Page - A

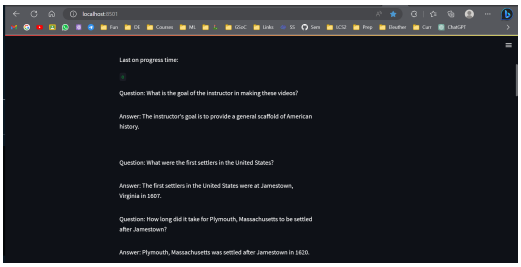

Figure 3: Result Page - B

6.3 Back End

It is noteworthy that since StreamLit code is essentially a Python script, the creation of distinct REST APIs for the purpose of fetching is deemed unnecessary. Rather, a separate backend file with functions can be created, which can subsequently be imported and invoked through callbacks and event handlers.

The question generation process utilized by our system is facilitated through the employment of GPT-3.5 by OpenAI. As an API endpoint for ChatGPT, a diverse range of prompts are employed to establish an optimal set. The identified set is subsequently utilized to prompt the model, generating a predetermined quantity of five questions per transcript chunk. The generated questions are rendered on the front-end for user convenience. To explore alternative methods, we experimented with Cohere's API, FLAN, Quantized Llama, and other instruction-tuned models from HF. However, our evaluation revealed that these methods were deficient in either instruction-following or generation length.

7 Evaluation

In order to assess the efficacy of our experiments, we employed a human evaluation process that entailed the establishment of three distinct metrics: Adequacy, Fluency, and Relevance. Adequacy was operationalized as the extent to which a given question was appropriate for the provided text, and was quantified using a binary score (i.e., yes/no). Fluency, on the other hand, was defined as the degree to which a question was linguistically fluent, without regard to its relevance to the text, and was rated on a scale of 1-5, with 1 denoting the lowest level of fluency and 5 denoting the highest. Lastly, Relevance was construed as the degree to which a question was germane to the text, was worthy of being posed, and was rated on a 1-5 scale.

To facilitate the evaluation process, a Google Form was disseminated to a sample of 15 individuals aged between 18-22 who were frequent viewers of educational content on YouTube. This demographic was deemed most appropriate for mimicking a larger audience. Within the Google Form, one paragraph was randomly selected from each video, and the top-k questions generated from these paragraphs were used as the basis for evaluation.

8 Results

Our observations in Table 1 indicate that Baseline B1 outperforms B2, which was to be anticipated given that it utilizes a sophisticated transformer-based T5 architecture, albeit in a reduced form. The model is pre-trained on copious amounts of data and is also fine-tuned on Squad for question-generation purposes. In contrast, B2 frequently generates incoherent text, thereby suffering in terms of the overall rating. Nevertheless, it exhibits a reasonably high level of adequacy, as the evaluators were instructed to designate it as adequate even if the generated questions were only tangentially related to the given topic (i.e., containing relevant keywords).

Our results are better in all respects except Avg. Adequacy from our baselines. The relevance and fluency is significantly better and unlike our baselines our proposed methods also generated answers making it a better choice altogether.

9 Future Work

In our future work, we aim to extend our research in two different directions. Firstly, we aim to improve the transcription process by correcting the automatically retrieved transcripts using some contextual error correction pipeline.

Secondly, we plan to incorporate better prompting techniques in our approach. With the advent of language models trained on large volumes of data, we believe that we can harness their power through Zero-Shot learning. This involves prompting the models with transcribed texts and expecting them to generate questions and answers in a template format. To achieve this, we intend to use Cohere and OpenAI's Instruction Tuned LLM APIs and HuggingFace's open source instruction tuned models, in addition to OpenAI's ChatGPT API which we are currently using.

References

Fatih Cagatay Akyon, Devrim Cavusoglu, Cemil Cengiz, Sinan Onur Altinuc, and Alptekin Temizel. 2021. Automated question generation and question answering from turkish texts using text-to-text transformers. *arXiv preprint arXiv:2111.06476*.

Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. 2021. Simplifying paragraph-level question generation via transformer language models. In *PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part II 18*, pages 323–334. Springer.

Hung-Ting Su, Chen-Hsi Chang, Po-Wei Shen, Yu-Siang Wang, Ya-Liang Chang, Yu-Cheng Chang, Pu-Jen Cheng, and Winston H Hsu. 2021. End-to-end video question-answer generation with generator-pretester network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11):4497–4507.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Sandeep Subramanian, Saizheng Zhang, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. *arXiv preprint arXiv:1705.02012*.

Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. 2021. A review on question generation from natural language text. *ACM Trans. Inf. Syst.*, 40(1).

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2018. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6*, pages 662–671. Springer.