Automatic QnA Generation from YT Videos

Mohammad Aflah Khan (2020082) & Neemesh Yadav (2020529)
Group 1

Motivation

Modern age of fast-paced media has led to decrease in human attention span

Implications in education: learners disengage from lengthy lectures, struggle to recall material

Potential mitigation strategy: implementation of a feedback loop with regular questions on these youtube videos

Problem Statement

This study aims to implement deep learning (DL) models for generating question-and-answer pairs from transcripts extracted from YouTube videos and develop a user interface for interactive prompts and real-time feedback while watching YouTube videos.

Overarching aim is to develop end-to-end user application that generates unique question-answer pairs from YouTube video transcript which are presented at fixed intervals of time

No existing works that provide similar end-to-end pipeline, while other works focus on isolated aspects of the project

Related Works

A neural question generation approach was proposed in a study that generates answer-aware input representations using an encoder-decoder architecture. The study demonstrated that the proposed approach could generate fluent and diverse questions without relying on rigid heuristic rules.

Another study introduced 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 of existing models for question generation analyzed different methods, from traditional rule-based approaches to advanced neural network-based methods. The survey provides a valuable reference for researchers and identifies promising future directions.

In a study focused on Turkish question answering and generation tasks, the authors proposed a multi-task learning framework using a fine-tuned multilingual T5 transformer. The proposed approach streamlines the generation of exam-style questions and achieves state-of-the-art performance on various datasets.

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

A recent study 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 affecting model performance and identified possible reasons for model failure.

Despite the potential benefits of video-based question generation methods, the decision was made against using them due to their computational complexity and the high accuracy of transcript-based representations for the targeted educational videos.

Dataset

All of our experiments comprised the selection of five educational YouTube videos from the “CrashCourse” channel, chosen randomly. These videos were deemed suitable for the task at hand due to their inherent level of detail and structural organization.

We tried to extract the transcript using both the automated transcription provided by YT and Assembly AI’s API

Video Transcription Analysis

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".

Due to these reasons we decide to use the automated transcripts only as retrieving transcripts via API calls is very slow

Baselines (1/2)

A. End-to-End Question Generation (Answer Agnostic)

The approach used is based on the end-to-end methodology proposed by Lopez et al. It utilizes the pre-trained "valhalla/t5-small-e2e-qg" model from HuggingFace, specifically designed for end-to-end question generation. The model is executed using sample code from the original repository. This approach is answer-agnostic, meaning it does not rely on any information about the answer to generate questions. However, it may generate questions for irrelevant filler text or introductory sections in videos, which do not contribute to meaningful content.

Baselines (2/2)

B. Traditional Linguistics Based Question Generation - B2

This baseline approach relies on traditional linguistics and hardcoded rules to transform sentences into questions. We utilize an existing implementation present on Github at dipta-dhar/Automatic-Question-Generator. However, this approach has limitations as it can only generate questions and cannot provide correct answers. Additionally, its performance may be suboptimal due to reliance on linguistic rules that may not cover the full range of unstructured input data.

Frontend

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.

Backend

Since Streamlit code is essentially a Python script, the creation of distinct REST APIs for the purpose of fetching is 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 however for comparison sake we've added an option to use both OpenAI and Cohere's API in the app

Evaluation

We conducted a human evaluation process to assess the efficacy of our experiments, using three metrics: Adequacy, Fluency, and Relevance. Adequacy was measured as a binary score (yes/no) based on the appropriateness of the question for the provided text. Fluency was rated on a scale of 1-5, indicating the linguistic fluency of the question without considering its relevance to the text. Relevance was rated on a 1-5 scale, indicating the degree to which the question was germane to the text and worthy of being posed.

To facilitate the evaluation, we distributed a Google Form to 15 individuals aged 18-22 who were frequent viewers of educational content on YouTube. One paragraph was randomly selected from each video, and the top-k questions generated from these paragraphs were used for evaluation. This demographic was chosen to mimic a larger audience for the evaluation process.

Results

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.


Thank you!