Generating Questionnaire from Lecture Transcripts using Keyword Detection

Subrat Roy and Saksham Rohatgi

April 29, 2024

1 Abstract

This project presents an innovative approach to enhance the efficiency of generating multiple-choice questions (MCQs) from lecture videos. The process begins with extracting keywords from the lecture videos using advanced natural language processing techniques. These keywords serve as pivotal elements in understanding the core concepts discussed in the lectures. Subsequently, the lecture transcripts are segmented into coherent sections, ensuring that each segment encapsulates a distinct topic or concept. This segmentation step is crucial for organizing the content and facilitating the generation of targeted MCQs. Utilizing the extracted keywords and segmented transcripts, the system employs an algorithmic framework to formulate MCQs dynamically. The generated questions are designed to assess the comprehension of specific topics covered in the lectures, thereby promoting effective learning and retention. By automating the MCQ generation process, this project aims to alleviate the burden on educators and instructional designers while ensuring the availability of diverse assessment materials for learners. The proposed system demonstrates promising results in efficiently producing high-quality MCQs aligned with the content of lecture videos, thereby contributing to the advancement of automated educational content generation.

2 Introduction

In summary, the background for this project encompasses the challenges associated with assessing student learning from lecture videos, the potential benefits of automating the generation of MCQs, and the interdisciplinary nature of the research area, drawing from NLP, educational technology, and assessment theory. For the keyword extraction since a good dataset was not available, we generated the dataset with the help of Gemini along with creating a custom dataset using a few books which have been cited in the referrence section. Then the MCQ questions were obtained using a pretrained AI model.

3 Literature Review

1. The paper proposes a system that automatically creates and inserts self-assessment questions within online video lectures. The system analyzes the lecture transcript to identify key points and generates questions at appropriate intervals, potentially improving student engagement and knowledge retention. This paper looks at detecting topic boundaries lexically using text-tiling algorithms, which is a good idea if we are looking at naive data, but if we look at technical data as lecture transcripts belonging to CS domains, we need to look out for semantic relations in the text.It also

looks at generating assessment items at every topic change to check the attention of the students, but this has a huge flaw given the fact that we are only using data about the CS domain, the number of keywords is very low, and they tend to repeat very often, which would lead to the questions getting repeated a lot, and also the fact that Wikipedia is being used for gathering concepts because the data in Wikipedia is too diverse for question generation for a technical topic. Relying on the Wikipedia category taxonomy for ontology-based distracted generation leads to limited coverage, especially in specialized or emerging domains where Wikipedia may lack comprehensive information.[1]

2. The research describes a system for automatically transcribing speeches in the Polish Senate. It uses a combination of techniques, including speech recognition, language modeling, and acoustic analysis, to convert spoken words into written text. This can potentially improve accessibility of Senate sessions and facilitate information retrieval.

3. The paper addresses the challenge of navigating long lecture videos efficiently. It proposes a method that analyzes both the visual elements (slides) and audio (speech) of a lecture to create a text-based index. This index allows viewers to jump directly to relevant sections of the video based on keywords, improving navigation and focus during learning. This paper aims at using OCR for keyword generation which is not possible for such a large corpus of data.[2]

4 Assignment Problem

Generating Questionnaire from Lecture Transcripts using Keyword Detection. First we created a dataset for the keywords using gemini. Then a custom data set was created using several books. Then both of these were compared and jointly used to get a dataset which was further used. Then using selenium all lecture transcripts were extracted and then with the help of this questionnaire was generated using gemini api which is a pre-trained AI model

5 Implementation

Coming to the implementation of this project. Due to the unavailability of a good data set we had to first generate a data set. The online data set which was available used wikipedia as it's only source of information so it would not include everything and moreover a lot of keywords were going missing. So we created a bot using selenium web driver which enters a prompt on the gemini model and returns the output it produces. The gemini api was not used as it was returning absurd outputs to the prompts. A python program was created

which gave all the video links present in a playlist and stored it into a text file. Then using these text files one by one the keyords for each video were extracted A custom dataset was created using a few famous AI/ML books which provided us with a small but a good dataset with all keywords being directly about the topic and relevant overall as well. We used this custom data set as our primary dataset and the one generated with gemini our alternative data set.

- We start by gathering a list of important words or phrases we want to find (keywords) from a file.
- The transcript (text we want to analyze) gets cleaned up: we break it down into smaller pieces (words or even short groups of words) and remove common, unimportant words. Then, we put the remaining pieces back together into a single string.
- We use a special tool (CountVectorizer) to convert the transcript into a kind of code based on the keywords. This code tells us if each keyword is present in the transcript (like a 1) or not (like a 0).
- We create labels (tags) for each piece of the transcript (word or phrase). If a piece matches a keyword, it gets a label of 1 (keyword). Otherwise, it gets a label of 0 (not a keyword).
- We train a system (Random Forest Classifier) using the coded transcript and its labels. This system learns to recognize patterns that distinguish keywords from other words.
- Once trained, we can use the system to analyze a new transcript. It predicts labels (keyword or not) for each piece of the new transcript.
- Finally, we show the predicted keywords as a list, removing any duplicates, to give a clear picture of the important words or phrases the system found in the new transcript.

Other than random forest we have also extracted keywords using natural language processing methods in which we found keywords from a new transcript that matched keywords from an external data set which had been earlier created by us. Both the transcripts file and datasets were lemmatized which made sure that found keywords are based on the same lemma as those in the dataset helping to capture variations of the same word.

Then the transcripts were extracted using an official youtube api which required us to give it the video id and it returned the transcript with time stamps as a dictionary. A separated file was created for each playlist which had a transcript for each of the video lecture involved. The playlists of the lectures we chose were cs221,cs229,cs230,cs234,cs231,cs4780,ocw6.036 and ocw6.s191.

Coming to questionnaire generation. The same problem persisted for this as well. The data set was not available on the internet so creating one would be tough. Rather than that we tried using Llama which is again a pre-trained model but it caused a problem since it cannot be run locally due to the fact that the newer version has just been released and the older ones have been discontinued and the newer one is not available locally as mentioned earlier. So we proceeded with using gemini api which in this case returned quality MCQ questions utilizing the keywords and since multiple keywords were provided to it. The questions would stick to the topic of interest.

6 Result and Analysis

- For results we used performing metrics namely precision, recall, frequency and accuracy.

  Precision measures the proportion of correctly identified positive cases among all cases identified as positive.

  Recall measures the proportion of correctly identified positive cases among all actual positive cases.

  Frequency represents the occurrence rate of a specific event or item within a dataset or population.

  Accuracy quantifies the proportion of correctly predicted outcomes among all predictions made.

- Random forest model we used when trained on cs221 transcripts and tested on a new transcript gave us the following metrics:-

  Precision: 0.058823529411764705

  Recall: 0.009230769230769232

  Frequency: 0.009230769230769232

  Accuracy: 0.9962514642717688

- The NLP script we used to extract keywords in a brute force way gave varying returns based on transcripts ranging from 15 percent accuracy to 85 percent accuracy as a result of the unclean ground truth labels.

- The question generation from the gemini api was very successful resulting in question answer pair pertaining only to the keywords given to the model.

  example:These questions and answers have been derived using the script after giving it a list of 20 keywords extracted through our random forest model.

  1. Which term refers to a type of neural network that can generate new data from existing data?

2. What is the name of the optimization algorithm commonly used in deep learning for minimizing loss functions?

3. What type of environment is often used for training autonomous vehicles?

4. Which subfield of deep learning focuses on generating new images?

5. What is the process of changing the weights and biases of a neural network to reduce error?

6. Which mathematical concept is used to describe the rate of change of a function?

7. What type of software is used to create deep learning models?

8. Which field of deep learning deals with generating natural language text?

9. What is the name of the competition that challenges teams to develop deep learning models for solving real-world problems?

10. What is the historical origin of machine learning?

Answer Key:

1. Generative Deep Learning

2. Gradient Descent

3. Synthetic Environments

4. Image Generation

5. Gradient Descent

6. Derivative

7. Deep Learning Frameworks

8. Natural Language Generation

9. Kaggle Deep Learning Competition

10. Early 20th century, with the work of Alan Turing and Claude Shannon

7 Future Works

This is a very crude model due to the fact that we do not have any well defined dataset to work with. If we had more data to work with we could extract keywords using deep learning methods and segment the transcript based on keywords using RNN's and LSTM's.

8 Referrences

1. Automatic Generation and Insertion of Assessment Items in Online Video Courses by the Department of Computer Science and Engineering, Centre for Education Technology, Indian Institute of Technology Kharagpur, West Bengal, India[1]

2. Automatic Keyphrase Extraction and Segmentation Video Lectures by the Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India

3. System for Automatic Transcription of Sessions of the Polish Senate The Polish-Japanese Institute of Information Tech nology

4. Erratic Navigation in Lecture Videos Using Hybrid Text-based Index Point Generation by the Department of CSE, VTU[2]

5. Pattern recognition and Neural Network by Yoshua Bengio

6. Fundamentals of deep learning-designing next generation machine intelligence by Nikhil Buduma

7. Fundamentals of Machine Learning by Thomas Trappengerg

8. Stanford cs221,cs229,cs230,cs234,cs231,cs4780,ocw6.036 and ocw6.s191 lecture notes