Search in presentations
Functional Specification

Michael Wall
13522003

Thursday 24th November, 2016

Contents

1 Introduction

The idea for this search system is to enable users to search presentation materials, specifically videos, with greater ease. At present, if a user wants to look for information on a topic, they may come across a video presentation on the topic. The video may be too long to watch all of if the user is only looking for a specific piece of information. The system aims to allow a user to search theses styles of video to get straight to the section which is relevant to their information need.

The need for this system comes from researching college material using external sources. Sometimes these come in the form of lectures and screencasts from other colleges. These lectures may have the information that you are looking for but are too long to warrant as a usable resource.

These lectures are usually in video format with presentation slides as the video and audio from the lecturer overlaid on the video. The system will allow the user to search the contents of the lecture slides used by the lecturer and also to search the voiceover presented by the lecturer for extra information that may not have been included in the slides.

The plan for this system is to use image processing to recognise text boxes or lecture slides which contain strings of text from keyframes in the video. The system will then use Optical Character Recognition (OCR) to convert the images into searchable text. In addition, the audio will be parsed using (insert name for audio search) to retrieve searchable text. The users search request will then be applied to the output of the above functions to find sections in the video which may be relevant. These sections will be highlighted to the user so that they can jump to these points.

2 Functional Description

Our system will use a combination of Optical Character Recognition and Automatic Speech Recognition to process a selected video. These are detailed further below.

2.1 Automatic Speech Recognition (ASR)

[1]

With current technology in automatic speech recognition it is not possible to generate a perfect transcription of audio. The generated index data can be searched

through, however it will have errors which will cause some search terms not to be found, even in a relevant piece of dialogue. This will in turn reduce the Mean Average Precision of the overall system. However, this might be tolerable if not all of the data is required to be relevant in a search, as long as we can have some keywords that are important identified. For example, if the lecturer speaks about a topic and mumbles through some of the sentences, but mentions one of our keywords. If this is recognised then it is irrelevant whether or not the rest of the sentence was heard correctly. In our case, the direct translation of the audio is not what we are trying to obtain either, just a pointer to a section that might be relevant. The user can then navigate to this section and determine if it is of relevance themselves.

For our videos, we will be processing them directly through YouTube, as this is where the majority of our target data is located. YouTube provides an application programming interface to access a caption track[2] on a given video. This API allows the captions to be retrieved using a HTTP GET request.

This is how we will access captions for the majority of our videos. These captions are generated using Google's speech to text API and works for over 80 languages[3]. In my experience, this API provides a reasonably good level of accuracy, even in noisy or imperfect audio environments.

In some cases, videos can be captioned by the uploader or voluntarily by a viewer, and this will help our search vastly in these cases.

2.1.1 Algorithms

We will process the audio data as follows:

- The search request will be broken into terms using porter stemming algorithm.
- The caption data will also be passed over using porter stemming to get all of the terms.
- We any appearances of the search terms will be marked in the caption data.
- The marked sections will be compressed in a 0 - 100 range in the form of a heat map.
- This heatmap will allow the user to see where their search terms occur most frequently in the caption data, and use this to navigate to relevant sections in the video.

2.1.2 Pros and cons

Trusting Google's Machine Learning has it's advantages and disadvantages. For one, they are using Machine Learning to constantly improve their accuracy in translation, and they also have access to possibly the largest set of training data for such a task. This is also a built in solution to the videos, and does not require extra computational resources for the user to process a video. This preprocessing is done by YouTube when a video is uploaded.

On the other hand, the accuracy of this method is not as accurate as if we were to have videos manually transcribed. This would vastly improve our search capability and accuracy, but at the expense of costly man hours. It would also limit the number of videos our search application could be used for.

2.1.3 Assumptions

We are making assumptions about the data which a user want's to search. Such assumptions are that the videos the user wants exist on YouTube and not on a college website. It also assumes that the API provided by YouTube will be available indefinitely in the future.

3 Implementation

3.1 Architecture diagram

Insert High Level diagram of the system here.

3.2 Test effectiveness of the system

How we would evaluate the system

Notes

[1]ASR is the recognition of human speech and converting it to text.

[2]Put reference to: https://developers.google.com/youtube/v3/docs/captions

[3]https://cloud.google.com/speech/