ASR data pruning

BTP Spring 2023

Guide: Prof Anil Kumar Vuppala
Mentors: Hasvitha, Shivang, Kowshik, Keshav, Anuprabha
Team: Vaibhav Agarwal (2020101041), Urvish Pujara (2020101032)

Overview

- Our BTP-1 project is about automatic or semi-automatic data collection for ASR. We aim to collect about five thousand hours of data per team member in the education and health domain which will assist in the development of domain adapted ASR.
- We plan to collect domain data in English and 1-2 Indian languages, using our ASRs, data collection platforms and other methods like text scraping.

Objectives

- Domain (healthcare / education) related text data collection by web scraping.
- Interface development and collection of speech samples.
- Collection of speech samples from open-source videos .
- Preprocessing of collected speech samples for domain adapted ASR.

Methodology

For collecting speech data from YouTube videos, the first step is to find relevant videos in the health/education domain, preferring videos with inbuilt captions, which can give us ground truth transcript files. Audio can be extracted from YouTube videos using python libraries such as pyTube and youtube-dl. And YouTubeTranscriptAPI can be used to get the transcript files with timestamps. Additionally, we need to clip the audio files into 15 sec bits and prune the srt files accordingly as well.

Methodology

Web scraping : We plan on collecting the textual data in medicine and education domain, from the curated list of websites using a deployed web scraper built using python modules like selenium, beautiful soup, etc. for extracting the textual data. Once the data is collected, we plan to employ some text to speech APIs or other manual methods for converting the collected text data to speech data..

Work done so far

- Preliminary task - Matching audios to their transcript (srt) files

  https://github.com/KarmanjyotSingh/BTP-1

- Extract audio and transcript data for a pilot dataset of 10 videos links from 3Blue1Brown

  https://colab.research.google.com/drive/1as5m4z3Y8-Y9co9lWk_JEX3qgXc85E7H?usp=sharing

Timeline

Sprint 1 (2 weeks)

Preliminary task - Matching audios to their transcript (srt) files

Sprint 2 (2 weeks)

Collect relevant websites and open-source videos for domain data collection

Sprint 3 (4 weeks)

Get data from websites through web scraping and from videos by extracting audio and in-built transcript files.

Sprint 4 (2 weeks)

Data preprocessing for domain adapted ASR

Sprint 5 (2 weeks)

Interface development

Expected Outcomes

- Ten thousand hours of education/health domain speech samples.

Challenges

- When using the Google speech to text API, we observed that the transcript generated for large files wasn't exact, or completely accurate with lots of sentences missing when comparing the original and the generated transcripts. On the contrary, it worked perfectly for short audio files.
- For now, we don’t have enough storage and computation resources on our systems to extract and store audio and transcript files from a considerable number of video links.

Thank You