Introduction

In the last several years, researchers have made significant progress on transcribing speech, but general speech recognition still has not been solved yet for any language. That is because speech recognition is fundamentally hard.

Speech recognition models training from scratch is both complex with a steep learning curve and exhaustive in time, data and computing resources. Industry-standard is to first use already available STT tools/services, customize and fine-tune them for a specific problem. Only then, having a good grasp of available systems capabilities and shortcomings, dive into extensive research, design, and optimization process for developing new models and algorithms.

During this assignment, the student will be placed in the shoes of an ML engineer/data scientist in-charge of implementing and optimizing general speech recognition pipeline given a restricted time frame.

Things you will learn

- Speech data preparation and validation.
- Basics of Amazon Web Service (AWS) cloud platform.
- Working with Amazon Transcribe automatic speech recognition system.
- Make critical insights into speech recognition results and propose improvements.

Goals

At the end of this assignment student will be evaluated on the quality of completion of the following goals:

- A well-justified in-depth analysis of acquired speech recognition results. (40%)
- Implementation of Amazon Transcribe speech-to-text (STT) pipeline. (30%)
- Attempt to improve over baseline results. (30%)

Requirements

- Use Python for all parts of the homework.
- Present your analyses using notebooks (Jupyter Lab/Notebook, Google Collab, etc.)
- Well structured and clean code.

Steps

1. Download and prepare Descript's Youtube dataset

Descript's Youtube dataset can be found in their Medium post. This dataset contains audio files (their Youtube URL links) and hand-labeled transcriptions.

The way Descript presents their data in their post makes it harder to retrieve it. Therefore, extract and download this data (audio and transcripts) and make sure it's valid and does not have errors.

The goal of this step is to prepare a Python script that prepares Descript's Youtube audio transcriptions dataset by:

- Downloading audio files (only within given timestamps).
- Building information index file (.csv, .json, or any other format) that links audio files to their transcripts, Youtube URLs, timestamps, and additional optional metadata.

This ensures that the collected data is easily usable and **reproducible**.

2. Setup Amazon Transcribe speech-to-text pipeline

Following Amazon Transcribe documentation and any additional sources, set up a speech-to-text (STT) pipeline for transcribing audio files.

The goal of this step is to have a functional pipeline (as a Python script) that needs to have the following sequence of processes:

1. Upload the audio file to Amazon cloud storage (Amazon S3).
2. Transcribe uploaded audio files using Amazon Transcribe (without customizations).
3. Retrieve transcription result data from the cloud.
4. Parse transcription and extract its text output.

3. Transcribe the prepared dataset

Transcribe audio files in the prepared dataset using Amazon Transcribe service as follows:

- Run the previously described STT pipeline (Step 2) on the dataset prepared in Step 1.
- Parse transcription results for all audio samples in the dataset.
- Save output transcripts in a single file from which further analysis could be made.
- Ensure that individual transcripts in the transcripts file can be mapped back to the dataset index file.

The goal of this step is to successfully prepare the output transcripts file.

Note: In this step only use default Amazon Transcribe parameters.

4. Analyze intermediate results

Analyze transcription results manually or analytically. Make insights into them and answer the following questions:

- What problems in the resulting transcriptions can be seen?
- What actions could improve transcription results?

5. Implement suggested improvements

Implement suggested changes to the speech-to-text (STT) pipeline. Choosing what improvements should be made and how is up to personal preference.

These improvements can be original Amazon Transcribe pipeline modifications, the addition of data pre-processing steps, implementation of new services, or even custom STT models/pipelines.

The overall goal of this step is to demonstrate the ability to implement theorized changes in speech recognition processes.

6. Transcribe the prepared dataset with an improved STT pipeline

Transcribe audio files in the Descript's Youtube dataset using the improved speech-to-text pipeline.

7. Analyze final results

Analyze final transcription results manually or analytically just like in step 4. Make insights into them and answer the following questions:

- How results have changed in comparison to step 3 results? Have improvements in transcription quality been achieved?
- What problems can be observed in the final results?
- What further changes to the STT pipeline could result in improved transcription quality by addressing the observed problems?

Career

About the company

Biomapas IT development department is Biomapas newest addition that is privileged to enjoy both: startup-like work environment, drive for new ideas and achievements along with backing and benefits of being part of a stable and well-established company. By joining us you will join a small, growing team of ambitious developers working on data science and IT infrastructure tasks, where you will get to experience and be part of many different aspects of software development along the way.

Biomapas as a company is more than 20 years old and is well established. Biomapas is a functional and full outsourcing solution provider to the global life science industry, with key expertise in Clinical Trials, Regulatory Affairs, and Pharmacovigilance. With headquarters in Lithuania and offices in Switzerland, Russia, Georgia, Ukraine, Poland, Kazakhstan, and Sweden, Biomapas operations are spread over 5 continents, concentrated in Europe, Russia, and the former CIS region.

Here, you will find a supportive work environment with a guarantee for professional and personal development, as well as a competitive salary, benefits, and many more initiatives that will make your daily office life comfortable.

For more about the company visit our website.

About the project

Our current project is to create a voice assistant for medical information that would aid in informing people about medicines and collecting information about their unforeseen side effects.

Expected candidate

Requirements

- Ability to work both in a team and individually;
- Problem solving skills;
- Experience in coding with Python;
- Experience in researching;
- Understanding ML core concepts and algorithms, experience working with them;
- Fluent in written and spoken English.

Responsibilities

- Data science research and machine learning models training/testing;
- Leading data science decisions and working on speech recognition, NLP, and NLU problems;
- Working with training data;
- Researching ML, speech and natural language topics;
- Augment, clean, and handle audio and textual datasets;
- Test effectiveness of different machine learning algorithms;
- Perform other duties relevant to the company activities and objectives as assigned.