Natural Language Processing

Lab 9

Speech to Text and Translation

Students:
Daniel Vidal Guerra, 1634599
Joan Samper, 1631430
Neil de la Fuente, 1630223

Professor:
Xim Cerdá Company

$31^{th} May, 2024$

Contents

1 Introduction

This report focuses on using OpenAI's Whisper model to transcribe audio from the "librispeech" dataset into text. We then use the Phi-3 LLM to translate the English text into Spanish. Our goal is to test how well these pre-trained models perform on this dataset without any fine-tuning or retraining. By evaluating their accuracy and efficiency, we aim to understand the strengths and limitations of these models in handling real-world audio-to-text transcription and Translation tasks.

2 Librispeech Dataset

The "LibriSpeech" corpus is a collection of approximately 1,000 hours of audiobooks that are a part of the LibriVox project. For this practice we will only use the validation split of the data, as we don't have to train or finetune any model, only test the capabilities of whisper. This dataset include the text and the written transcription, so a perfect ground truth is given to get transcription metrics for our model.

3 Whisper Transcriptions

The whisper-large model is used for this test, loaded from huggingface, achieving very good and coherent results in transcription. The large version is used because it get the best results, and as the model is only used for inference the resources cost are not very high.

3.1 Results

We used Whisper to transcribe 30 audio samples and analyzed both the qualitative and quantitative results. The transcriptions were compared to the ground truth to calculate ROUGE scores, which measure the precision, recall, and F1 score. These scores indicate how well the transcriptions match the original text.

Average ROUGE-S Scores

- **Precision:** 0.5373
- **Recall:** 0.6009
- **F1 Score:** 0.5647

These averages show that the transcriptions are generally accurate, with good alignment to the ground truth text.

3.2 Example Transcriptions and ROUGE Scores

Example 1

- **Predicted:** Mr. Quilter is the apostle of the middle classes, and we are glad to welcome his gospel.

- **Ground Truth:** mister quilter is the apostle of the middle classes and we are glad to welcome his gospel
- **ROUGE-S Scores:** Precision: 0.571, Recall: 0.645, F1: 0.606

Example 2

- **Predicted:** he tells us that at this festive season of the year with christmas and roast beef looming before us similes drawn from eating and its results occur most readily to the mind.
- **Ground Truth:** he tells us that at this festive season of the year with christmas and roast beef looming before us similes drawn from eating and its results occur most readily to the mind.
- **ROUGE-S Scores:** Precision: 1.0, Recall: 1.0, F1: 1.0

Example 3

- **Predicted:** He has grave doubts whether Sir Frederick Leighton's work is really Greek after all, and can discover in it but little of rocky Ithaca.
- **Ground Truth:** he has grave doubts whether sir frederick leighton's work is really greek after all and can discover in it but little of rocky ithaca.
- **ROUGE-S Scores:** Precision: 0.549, Recall: 0.595, F1: 0.571

Example 4

- **Predicted:** linnell's pictures are a sort of up guards and adam paintings and mason's exquisite idylls are as national as a jingo poem mr burkett foster's landscapes smile at one much in the same way that mr carker used to flash his teeth and mr john collier gives his sitter a cheerful slap on the back before he says like a shampooer in a turkish bath next man.
- **Ground Truth:** linnell's pictures are a sort of up guards and at em paintings and mason's exquisite idylls are as national as a jingo poem mister birket foster's landscapes smile at one much in the same way that mister carker used to flash his teeth and mister john collier gives his sitter a cheerful slap on the back before he says like a shampooer in a turkish bath next man.
- **ROUGE-S Scores:** Precision: 0.540, Recall: 0.532, F1: 0.536

Example 5

- **Predicted:** he laments most bitterly the divorce that has been made between decorative art and what we usually call pictures makes a customary appeal to the last judgment and reminds us that in the great days of art michael angelo was the furnishing upholsterer.
- **Ground Truth:** he laments most bitterly the divorce that has been made between decorative art and what we usually call pictures makes the customary appeal to the last judgment and reminds us that in the great days of art michael angelo was the furnishing upholsterer.

- **ROUGE-S Scores:** Precision: 0.976, Recall: 0.976, F1: 0.976

These examples demonstrate that while some transcriptions are perfect matches, others show small discrepancies that affect the ROUGE scores. Despite these variations, the overall performance of Whisper is impressive, achieving a high level of accuracy in transcribing audio to text.

It's important to note that the presence of commas in the text significantly affects the final value of the metrics. Even when the words are the same, the number of matching n-grams and the longest matching subsequences are reduced, leading to lower scores. This is because the ROUGE metric is sensitive to punctuation, which can break up otherwise identical sequences of words and the ground truth usually does not include punctuation marks but the Whisper transcription does.

4 Phi Translations

For the LLM we choose Phi-3 as it's open source and permission is not required. Llama models were an option, but as permission is required we decided to go with other models that also perform very well on this task.

In this case, no ground truth is available so the results will be analyzed qualitatively. To be able to correctly asses the goodness of the model the Translation is done from English to Spanish, as these are 2 languages spoken by all the group members.

4.1 Results

Original Transcription: Mr. Quilter is the apostle of the middle classes, and we are glad to welcome his gospel.
Translation: Sr. Quilter es el apóstol de las clases medias, y estamos encantados de recibir su evangelio.

Original Transcription: Nor is Mr. Quilter's manner less interesting than his matter.
Translation: Ni el modo de Mr. Quilter no es menos interesante que su asunto.

Original Transcription: he tells us that at this festive season of the year with christmas and roast beef looming before us similes drawn from eating and its results occur most readily to the mind.
Translation: Él nos cuenta que en esta temporada festiva del año, con Navidad y asado de cerdo ante nosotros, las comparaciones relacionadas con la comida y sus resultados ocurren más fácilmente en la mente.

Original Transcription: He has grave doubts whether Sir Frederick Leighton's work is really Greek after all, and can discover in it but little of rocky Ithaca.
Translation: Tiene serias dudas sobre si el trabajo de Sir Frederick Leighton realmente es griego, y puede encontrar en él poco de la Ithaca rocosa.

Original Transcription: linnell's pictures are a sort of up guards and adam paintings and mason's exquisite idylls are as national as a jingo poem mr burkett foster's landscapes smile at one much in the same way that mr carker used to flash his teeth and mr john

collier gives his sitter a cheerful slap on the back before he says like a shampooer in a turkish bath next man.
Translation: Las obras de Linnell son un tipo de retratos de guardaespaldas y pinturas de Adán y los idílios exquis.

Original Transcription: It is obviously unnecessary for us to point out how luminous these criticisms are, how delicate in expression.
Translation: Obviamente es innecesario para nosotros señalar cuán luminosas son estas críticas, cuán delicadas en su expresión.

5 Conclusions

In this report, we explored how well OpenAI's Whisper model transcribes audio and how Phi-3 handles translations. Our tests show that Whisper does a pretty solid job with transcriptions, often getting quite close to the actual text. Even though there are small mistakes in some transcriptions, the overall accuracy is impressive.

For the translations, Phi-3 gave us good results too. Since we didn't have a perfect translation to compare against, we relied on our own knowledge of English and Spanish and the pure qualitative evaluation. The translations generally made sense and captured the meaning well, though there were a few awkward phrases.

To sum up, both models performed well for our purposes. Whisper proved to be a reliable tool for turning speech into text, and Phi-3 handled translations competently. While there is always room for improvement, these models are definitely useful for real-world applications like transcribing and translating spoken content.