Plan prévisionnel

Dataset retenu

Pour évaluer nos modèles de transcription speech to text, nous avons besoin d'un dataset contenant des extraits audio et les transcriptions correspondantes. J'ai choisi le Multilingual LibriSpeech (MLS), téléchargeable à l'adresse : https://openslr.org/94

Nous allons utiliser une partie des fichiers test, en français.

Modèle envisagé

Whisper est un système de reconnaissance automatique de la parole (ou ASR, automatic speech recognition system), open source, publié par OpenAI le 21 septembre 2022. Les articles ci-dessous contiennent notamment :

- La présentation du modèle (kdnuggets)
- Une comparaison de ses performances avec celles de Kaldi et Wave2vec 2.0, notre baseline. (Deepgram – seul article qui n'est pas un article de référence, mais bonne présentation synthétique des résultats + explication de la métrique utilisée)
- Plusieurs cas d'application très récents (arxiv, avril 2024). On peut citer notamment la NASA, qui a choisi d'utiliser Whisper comme ASR pour sa nouvelle génération d'électronique embarquée.

Les cas d'utilisation possibles d'un traducteur automatique sont très variés (voyage, cadre professionnel, etc...), cependant notre objectif est précis :

Nous allons utiliser cet algorithme pour aider le personnel d'un restaurant situé en zone touristique... et se préparant à l'arrivée des jeux Olympiques...
à communiquer avec des clients pouvant venir du monde entier.

Cela implique plusieurs points essentiels, soulignés par la gérante de l'établissement, qui vont guider la construction de l'appli (notre cahier des charges) :

-détection automatique du langage,
-robustesse au bruit,
-rapidité,
-sécurité, respect de la confidentialité des clients,
-fonctionnement hors-connexion,
-facilité d'utilisation, design simple.

Références bibliographiques

- *OpenAI's Whisper API for Transcription and Translation,* by Eugenia Anello, 2 juin 2023

Présentation du modèle. Source : https://www.kdnuggets.com/2023/06/openai-whisper-api-transcription-translation.html

- *Houston we have a Divergence: A Subgroup Performance Analysis of ASR Models,* by Alkis Koudounas, Flavio Giobergia, 31 mars 2024

La Nasa utilise Whisper ! Source : https://arxiv.org/abs/2404.07226

- *WavLLM: Towards Robust and Adaptive Speech Large Language Model,* by Shujie Hu, 31 mars 2024

L'université chinoise de Hong Kong utilise le bloc encoder de whisper pour intégrer le speech to text à leur projet de LLM. Source : https://arxiv.org/abs/2404.00656

- https://deepgram.com/learn/benchmarking-top-open-source-speech-models

Il ne s'agit pas d'un article de recherche, mais on peut consulter cette page pour une explication de la métrique standard pour la transcription, le WER, et un aperçu des performances du model par rapport à la baseline.

Preuve de concept

Nos 2 modèles sont Wav2vec2, la baseline, et whisper, la nouvelle méthode.

Leurs performances et temps de prédiction vont être évalués sur le même jeu de données test, en utilisant le taux de mots erronés (WER) comme métrique objective pour la transcription (voir notebook).

En ce qui concerne la qualité de la traduction, il est beaucoup plus difficile de l'évaluer de manière quantifiée.

(Voir la conclusion du notebook pour une comparaison synthétique.)

Une version en ligne de l'application réalisée pour ce projet est disponible à l'adresse : https://rosetta-stones.streamlit.app/

Cette version est identique à celle qui est utilisée actuellement par le restaurant, à deux exceptions près : l'appli du restaurant est 100% locale (-> sécurité, confidentialité, fonctionne en cas de coupure de courant ou de réseau) et contient uniquement la page de traduction.