STAGE COURT LINGUISTIQUE INFORMATIQUE

Mission : Enrichissement et déploiement sur une infrastructure de recherche (https://www.huma-num.fr/) d'une chaîne de traitement automatique de la transcription, de l'alignement et de l'analyse phonétique de l'anglais.

Contexte

Dans le cadre d'un doctorat, le laboratoire CLILLAC-ARP s'est doté d'une chaîne de traitement pour l'analyse automatique des données de l'anglais oral. Les différentes étapes du projet sont décrites dans les publications suivantes (Ballier & Méli, 2023) et les aligneurs concurrents sont décrits dans (Méli et al., 2023).

A la faveur de l'émergence des LLM, et en particulier de Whisper (Radford, 2022), il est possible d'améliorer les transcriptions, surtout si l'on utilise des versions affinées (*fine-tuned*) de Whisper qui ne corrigent pas les disfluences de la parole. Le schéma suivant récapitule la chaîne de traitement actuellement disponible.

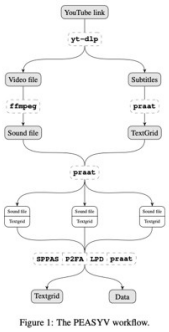

Figure 1: The PEASYV workflow.

Figure 1 : la chaîne de traitement de PEASYV (Méli et al., 2023)

Cette solution n'existe en local, est dépendante des versions des bibliothèques utilisées et n'utile pas les derniers développements des aligneurs et des modèles affinés. Il s'agit de rendre accessible au grand public cette technologie. En entrée, l'utilisateur indiquera une URL de Youtube ou chargera un fichier son et pourra en sortie télécharger une transcription du texte, et une . En mode expert, l'utilisateur pourra choisir l'aligneur et récupérer les rapports d'analyses phonétiques, les formats de fichiers.

Livrables :

- 1. Compléter la chaîne de traitement en python 3.9 en ajoutant l'aligneurs de Montreal des métriques : implémenter les métriques à base de n-grams, et, sous réserve de la disponibilité des inventaires de fréquence dans les corpus de référence de l'arabe (collaborations possibles), intégrer des métriques de sophistication lexicale présentes dans TAALES.
- 2. Dockerisation du système en vue de sa virtualisation
- 3. Déploiement du système sur un serveur de l'infrastructure de recherche huma-num.

Sources :

- Les versions affinées de Whisper pour la transcription de l'oral:
- https://huggingface.co/Transducens/error-preserving-whisper-distilled
- https://huggingface.co/Transducens/error-preserving-whisper
- L'aligneur développé par Montréal https://montreal-forced-aligner.readthedocs.io/en/latest/
- Le code R et Praat utilisé pour la génération des rapports phonétiques et des bases de données

Contacts :

UFR études anglophones : Nicolas Ballier (nicolas.ballier@u-paris.fr)

REFERENCES

Ballier, N., & Méli, A. (2023). PEASYV: A procedure to obtain phonetic data from subtitled videos. In *20th International Congress of Phonetic Sciences* (pp. 3211-3215). Guarant International.

Méli, A., Coats, S., & Ballier, N. (2023). Methods for Phonetic Scraping of Youtube Videos. In *6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)* (Vol. 6, pp. 244-249).