Les disfluences verbales et marqueurs discursifs :
Impact sur le traitement automatique de la parole et améliorations liées

L'équipe R&D (https://labs.linagora.com/) de la société **LINAGORA** (http://linagora.com) développe en open-source des outils d'assistance intelligente pour entreprises, y compris l'assistant vocal LinTO (https://linto.ai/) et la bibliothèque d'outils associée (https://github.com/linto-ai), dont le focus est de mettre à disposition l'état de l'art en Reconnaissance Automatique de la Parole (RAP). La parole conversationnelle intéresse tout particulièrement LINAGORA, avec comme application cible le **résumé automatique de réunion**. Cette application met en scène la RAP ainsi que le Traitement Automatique du Langage Naturel (TALN).

Les systèmes modernes de tâches de TALN, tels la restitution de la ponctuation et la génération automatique de résumé, sont basés sur des modèles statistiques d'apprentissage automatique (**Machine Learning**) entraînés sur de grandes quantités de données textuelles, extraites la plupart du temps de sites web et de recueils numérisés. Or ces données ne sont pas représentatives des transcriptions de **parole spontanée**. De ce fait, les modèles de TALN appris sur ces données ne sont pas adaptés à des applications comme le résumé automatique de réunion à partir de transcriptions de RAP brutes. D'autre part, un système de résumé automatique nécessite d'importants volumes de données pour atteindre des performances acceptables, et il serait trop long et coûteux de recueillir assez de données transcrites de paroles spontanées pour qu'un système de TALN basé sur du Machine Learning puisse généraliser correctement.

La motivation de ce stage s'appuie sur le constat que, si l'on met de côté les erreurs de transcription de la RAP (qui deviennent rares au fur et à mesure des progrès en RAP), deux des plus grandes différences entre la transcription de parole spontanée et le texte écrit est :

- d'un côté, la présence de disfluences verbales, à savoir les hésitations (« euh... », « hmm »), les répétitions, et les faux départs,
- d'un autre côté, la présence de marqueurs discursifs (« eh bien », « alors », « donc »…), qui sont souvent employés pour réguler le flux de parole.

Non seulement les disfluences et les marqueurs discursifs peuvent être importants pour les tâches en aval nécessitant une compréhension du langage (comme la segmentation discursive), mais ils sont présents dans tout corpus utilisé pour former de nouveaux modèles de RAP et devront donc être pris en compte lors de la création d'une vérité terrain. Dans ce qui suit, pour des raisons de simplicité, nous parlons de « disfluences » pour désigner à la fois les disfluences et les marqueurs discursifs, bien qu'il soit de manière générale utile de différencier les deux.

Il n'existe pas ou très peu de bases de données annotant les disfluences au sein de discours. Or, les modèles« Whisper » d'OpenAI, à l'état de l'art en RAP et appris de manière semi-supervisée sur un très grand volume de vidéos sous-titrées, présentent la particularité d'omettre les disfluences, dans certaines conditions qu'il reste à déterminer. Ils font par ailleurs preuve d'une étonnante robustesse dans plusieurs langues, dont l'anglais et le français.

Partant de ce constat, le premier objectif du stage sera de constituer une base de données avec annotation des disfluences dans des transcriptions de parole. La constitution de cette base se fera de manière automatique à partir d'un programme, élaboré par le stagiaire, qui consiste à appliquer les modèles Whisper aux bases de données vocales disponibles à LINAGORA, et à exploiter les alignements de ses transcriptions (incomplètes en termes de disfluences) avec la vérité terrain.

À partir de cette base de donnée, le second objectif sera d'entraîner un modèle « deep learning » de TALN permettant de détecter et supprimer les disfluences dans les transcriptions textuelles. Selon l'avancement du stage, il sera aussi possible d'entraîner un modèle permettant de rajouter des disfluences dans un texte, dont la principale utilité est d'augmenter les bases de données d'entraînement des modèles TALN pour qu'ils soient adaptés au langage parlé.

Le modèle de détection/suppression des disfluences sera utilisé pour améliorer les performances des systèmes de TALN appliqués à la parole. En plus de vérifier et mesurer cette amélioration de performances en TALN, le stagiaire pourra utiliser ce modèle pour analyser l'impact des disfluences sur les systèmes de RAP. En particulier, une des utilités d'un tel modèle est d'améliorer l'estimation des performances des systèmes de RAP, pour comparer les systèmes qui transcrivent les disfluences et ceux qui les omettent.

Encadrement du stage :
Le stagiaire sera encadré par Jérôme Louradour et Julie Hunter de LINAGORA.

Localisation : LINAGORA GSO, Toulouse

Compétences clés recherchées :

- Étudiants de M2 ou d'école d'ingénieur en dernière année, en informatique, avec des compétences en machine learning
- De l'expérience en deep learning serait un plus
- De l'expérience en traitement de la parole et/ou du texte serait un plus

Durée du stage : 5-6 mois

Gratification : à définir selon l'expérience du candidat

Contacts email : jlouradour@linagora.com, jhunter@linagora.com, jplorre@linagora.com

Références :

- « La parole spontanée : transcription et traitement », Thierry Bazillon, Vincent Jousse, Frédéric Béchet, Yannick Estève, Georges Linarès, Daniel Luzzati
- « Auto-interruptions et disfluences en français parlé dans quatre corpus du CID », Bertille Pallaud, Stéphane Rauzy et Philippe Blache
- « Analyse et détection automatique de disfluences dans la parole spontanée conversationnelle », Camille Dutrey

- (Whisper) « Robust speech recognition via large-scale weak supervision », Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey et Ilya Sutskever