Extraction d'informations sémantiques dans des transcriptions de résumés oraux d'histoires par des enfants

Informations générales
Durée : 6 mois Début : à partir de janvier 2025, au plus tard avril 2025
Lieu : Université d'Avignon – LIA – Campus
Gratification selon la grille réglementaire
Perspectives : Programme de doctorat de 3 ans

Contexte :
Ce stage s'inscrit dans le cadre du projet ANR Chica-AI (2024-2028), qui vise à concevoir un environnement informatique capable d'analyser automatiquement les résumés oraux d'enfants pour évaluer leur compréhension d'un texte à la suite d'une tâche de lecture. L'étude PISA 2019 montre, en effet, que 20 % des élèves français de 15 ans présentent des difficultés sévères en lecture, et que les écarts socio-économiques accentuent les disparités de niveau. Le projet a pour ambition de réduire les difficultés de lecture des enfants du cycle 3 en proposant des méthodes basées sur l'apprentissage artificiel, permettant un accompagnement personnalisé pour l'élève et un retour informatif pour les enseignants.

La compréhension de la lecture est un enjeu fondamental, et elle peut être entrainée grâce à des activités telles que le résumé de texte. Pour analyser automatiquement la compréhension du texte par l'enfant, il s'agit d'évaluer sa production orale du résumé. Pour cela, il faut extraire un ensemble d'informations sémantiques du résumé mais aussi fournir un ensemble d'indicateurs pertinents et différenciés, tant pour les élèves que pour les enseignants.

Pour atteindre ces objectifs, plusieurs modules seront développés : un module de reconnaissance de la parole, adapté aux voix d'enfants ; un module de compréhension de la parole permettant d'extraire des informations sémantiques ; et un module de mise en correspondance de ces informations sémantiques avec une grille d'évaluation fondée sur des critères psycho-cognitifs, pour juger de la qualité des résumés produits.

Travail du stagiaire
L'objectif du stage est de développer un premier système d'extraction d'informations sémantiques à partir des transcriptions de résumés oraux. Dans un premier temps, le stagiaire explorera différentes méthodes de traitement automatique du langage naturel (TAL) pour détecter et extraire des entités nommées (comme les lieux ou personnages) à l'aide de modèles de type CamemBERT et Flaubert. Dans un deuxième temps, ces méthodes seront adaptées à la détection d'évènements pertinents dans le résumé (différentes actions). Il s'agira de se familiariser avec des techniques avancées de TAL et d'apprentissage automatique (architectures neuronales et grands modèles de langues – LLM). Une des difficultés à lever consistera à faire le lien entre les informations sémantiques issues du résumé et celles recherchées pour son évaluation.

Candidatures
Le candidat doit être en master 2 informatique avec des connaissances en intelligence artificielle. Des connaissances en traitement automatique du langage naturel seront appréciées. Le candidat doit montrer un intérêt pour le travail en équipe et interdisciplinaire.

Les candidatures (CV et lettre de motivation, relevé de notes Bac+4) sont à envoyer nathalie.camelin@univ-avignon.fr avant le 15/12/2024.