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 2027, au plus tard avril 2027
Lieu : Université d'Avignon – LIA – Campus
Gratification selon la grille réglementaire

Contexte :

Ce stage s'inscrit dans le cadre du projet ANR Chica-AI (2024-2028), qui vise à concevoir un environnement informatique capable d'évaluer automatiquement une tâche de lecture. En effet en 2018, les résultats de l'étude PISA soulignaient les difficultés des élèves français en compréhension de l'écrit : 21 % des élèves de 15 ans se situaient sous le niveau minimal de compétences. Cette proportion atteint désormais 33 % selon les résultats PISA 2025.

Face à cet enjeu, le projet ambitionne de contribuer à la réduction des difficultés de lecture des enfants du cycle 3 (CM1 à 6e), en développant des méthodes fondées sur l'intelligence artificielle pour proposer un accompagnement personnalisé aux élèves et des retours informatifs aux enseignants. Le résumé de texte, après sa lecture, constitue un exercice particulièrement intéressant pour le travail de compréhension de l'écrit. Le projet vise ainsi à analyser automatiquement les résumés oraux produits par les enfants, en extrayant des informations sémantiques et en construisant des indicateurs pertinents et différenciés, à destination des élèves et des enseignants.

Pour atteindre ces objectifs dans le projet ANR Chica-AI, 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 prototype d'analyse automatique des transcriptions de résumés oraux d'enfants, en s'appuyant sur des architectures de type LLM (Large Language Models).
Dans un premier temps, le stagiaire explorera plusieurs architectures de modèles de langage (Qwen, Llama, ...) et étudiera différentes stratégies de *prompt engineering* pour extraire automatiquement les critères définis dans la grille d'évaluation des résumés. L'objectif sera notamment de comparer les modèles et les stratégies d'interrogation afin d'identifier les approches les plus pertinentes. Les différentes expérimentations devront également permettre d'évaluer l'impact des erreurs de transcription sur la qualité de l'analyse automatique.
Dans un second temps, le système pourra être enrichi en combinant les LLM avec d'autres approches. Plusieurs pistes pourront être explorées :

- RAG (Retrieval-Augmented Generation) : apporter au modèle des connaissances complémentaires, notamment à partir du texte source à résumer et de la grille d'évaluation.
- Extraction d'informations sémantiques : exploiter des systèmes de reconnaissance d'entités nommées (par exemple spaCy ou des modèles de type CamemBERT) ou des méthodes d'alignement sémantique (développées par la doctorante).

Une difficulté particulière consistera à établir le lien entre les informations sémantiques extraites du résumé de l'enfant, le contenu du texte source et les critères attendus pour l'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 15/12/2026.