Mémoires recherche : Iris Taravella

Trois sujets

Analyse linguistique du profil sociologique du locuteur et sa prédiction automatique à partir de transcriptions

- 4 catégories étudiées : âge, sexe, CSP, niveau d'étude
- Travaux antérieurs
  - Détection autom. de l'âge (apprentissage de surface, profond, LLM)
  - Comparaison de caractérisation automatique (LLM) et humaine
- Travaux à prévoir
  - Caractéristiques linguistiques saillantes pour chaque catégorie
    - Relevées par l'annotation manuelle/autom
  - Rôle de stéréotypes sociaux dans la détection
  - Détection automatique de trois catégories (CSP, niveau d'études, sexe) avec l'apprentissage de surface et profond

Prédictibilité en lien avec les pauses dans les données d’écriture en temps réel (financé)

- Prédictibilité sub- et supra-lexicale
  - syllabe, morpho et syntaxe
- En utilisant des outils comme BERT, concernant les régularités combinatoires et donc l'attente d'un mot dans un contexte donné (= le degré de prédictibilité), il s'agirait de voir si ce qui vient après une pause correspond aux attentes ou si le degré de prédictibilité est plus bas (ce qui expliquerait les pauses).

Excessivité (potentiellement financé)

- Annotation manuelle
  - Mise en place de la plateforme/de la méthodologie d'annotation collaborative en stockant des informations sociologiques sur les annotateurs
    - Réflexion comment attirer le public
    - Surveiller la variabilité d'annotateurs
    - Lien avec Qualtrics (outil d'enquêtes/sondages)

- Annotation automatique
  - Deep : Camembert (fine tuning)
  - LLM génératifs (différents paramètres : température, langue, contexte, exemple, définitions, formatage de sortie)
  - Apprentissage active