Détection automatique de questions spontanées vs questions préparées

Encadrants : Iris Eshkol-Taravella et Angèle Barbedette

Données : transcriptions de l'oral

Méthodes : apprentissage supervisé de surface (et si le temps le permet apprentissage profond)

Hypothèse de travail : les questions posées d'une manière spontanée ont des caractéristiques différentes de celles des questions préparées (p. ex. présence de disfluences au milieu de la question pour les questions spontanées vs présence de mots interrogatifs pour les questions préparées)

Étapes du travail :

1. Constitution de corpus (récupération, mise en forme des données)

2. Analyses manuelles du corpus pour repérer les caractéristiques des questions spontanées vs préparées, annotation de questions, évaluation de l'annotation manuelle

3. Vectorisation du corpus (tf-idf, word embeddings, n-gram) et autres prétraitements

4. Expériences combinant différents modèles statistiques proposés par Scikit-learn, différentes mesures de vectorisation et différentes features linguistiques sur un extrait du corpus de référence

5. Choix de la meilleure combinaison, évaluation sur un autre extrait (ou validation croisée à N-plis).

6. Analyse des erreurs

7. Expériences en deep learning (s'il reste du temps)