Stage pour Ingénieurs/Étudiants de Master II : Développement d'un Modèle de Langage pour la Classification et l'Analyse de Contenus Pédagogiques en Odontologie

Contexte

Les modèles de langage sont de plus en plus utilisés dans des applications éducatives, notamment en médecine et en odontologie. Cependant, ces modèles généralistes comme GPT-4 peuvent parfois manquer de précision et d'adaptabilité lorsqu'ils sont appliqués à des domaines spécifiques et suscitent des craintes (propriété intellectuelle, contrôle des données, respect de la vie privée). Récemment, des modèles de plus petite taille comme CamemBERT-bio, entraînés sur des corpus médicaux français, ont montré de meilleures performances pour classer des termes médicaux par rapport à des modèles comme BERT-7B, avec des temps de calcul et des émissions de carbone jusqu'à 30 fois inférieurs (Touchent et al. 2024). Ces avancées ouvrent des perspectives pour développer des modèles de langage accessibles, adaptés à des domaines spécifiques, et utilisables localement avec des ressources limitées.

Ce stage s'inscrit dans un projet visant à entraîner et évaluer des modèles de langage spécialisés pour classer, regrouper et organiser des contenus pédagogiques en odontologie, tout en soutenant l'analyse des curriculums et l'extraction des notions clés.

Objectif du projet

Évaluer les performances d'un modèle de langage spécialisé dans la classification et le regroupement de contenus pédagogiques (e.g., QCM, glossaires, chapitres de cours), dans l'extraction des notions clés et dans la hiérarchisation de ces contenus selon les niveaux de compétences définis par les directives européennes. Une analyse comparative des coûts environnementaux des modèles employés sera intégrée aux métriques d'évaluation.

Étapes du projet

1. **Classification et clustering des contenus pédagogiques**
    - Effectuer une analyse exploratoire des données (QCM, glossaires, chapitres).
    - Utiliser des approches de classification supervisée ou de clustering non supervisé pour regrouper les contenus selon des thématiques ou niveaux de compétences.
    - Extraire des notions clés et identifier les chapitres les plus pertinents pour chaque thématique.
2. **Entraînement du modèle de langage**
    - Pré-traiter les données (découpage en fragments, vectorisation).
    - Adapter des modèles existants (CamemBERT-bio, Mistral) via affinage ou génération augmentée par recherche (RAG).
3. **Évaluation des performances et des impacts**
    - Calculer des métriques de classification et de clustering (F1 score, précision, cohérence des clusters).
    - Comparer les performances des modèles spécialisés avec des modèles généralistes comme GPT-4.

- Évaluer l’impact environnemental des modèles via des indicateurs tels que la consommation énergétique lors de l’entraînement et l’inférence, en collaboration avec des outils comme CodeCarbon ou équivalents.

4. **Exploration de l’explicabilité et des représentations latentes**
    - Étudier les représentations internes du modèle pour comprendre les regroupements réalisés et justifier les classifications.

Compétences requises :

- Connaissances en traitement du langage naturel (NLP), apprentissage automatique et clustering.
- Maîtrise de Python et des bibliothèques de deep learning (PyTorch, TensorFlow).
- Intérêt pour les questions d’impact environnemental et d’analyse de données éducatives.

Perspectives

Ce stage permettra de développer des compétences en IA appliquée à l’enseignement de l’odontologie et de contribuer à l’innovation pédagogique. En intégrant une réflexion sur l’impact environnemental, il s’inscrit dans une démarche responsable et durable, en phase avec les exigences actuelles des industries technologiques et éducatives.

Contacts

Thomas Grenier (thomas.grenier@creatis.insa-lyon.fr
Sébastien Valette (sebastien.valette@creatis.insa-lyon.fr)
Raphaël Richert (raphael.richert@insa-lyon.fr)

Références

Touchent R, Romary L, de La Clergerie E. 2024. CamemBERT-bio: Leveraging Continual Pre-training for Cost-Effective Models on French Biomedical Data. 2024 Jt Int Conf Comput Linguist Lang Resour Eval Lr 2024 - Main Conf Proc.:2692–2701.