Offre de stage (Master/Ingénieur)

Construction automatique d'un graphe de connaissances géo-historiques à partir d'entrées encyclopédiques

Mots clés

LLM génératifs, graphes, dictionnaires anciens, extraction de connaissances

Contexte

Le projet ECoDA financé par la FIL[1] (Fédération Informatique de Lyon) s'inscrit dans le cadre d'une collaboration entre les laboratoires d'informatique ERIC[2] et LIRIS[3] et en partenariat avec le laboratoire de linguistique ICAR[4]. Le projet a pour objectif le développement d'outils pour l'analyse de l'évolution des connaissances dans les encyclopédies et dictionnaires anciens. Plus particulièrement ce stage s'intéresse à la construction automatique de graphes de connaissances géo-historiques. Un deuxième stage est également disponible, sur le thème de l'identification automatique de domaines de connaissances et l'analyse de leur évolution. Ce sujet est disponible ici :

https://partage.liris.cnrs.fr/index.php/s/trMm6LMoZNa7DLZ

Objectifs du stage

L'extraction et la structuration automatiques des informations géographiques (et leur temporalité) permettent de créer des graphes de connaissances (Tual et al. 2023). L'utilisation de ces bases de connaissances permet de faciliter l'analyse et la visualisation des dynamiques géographiques mais aussi historiques d'un territoire. Dans le cadre de ce projet, nous nous intéresserons en particulier à identifier les structures récurrentes et les types d'informations présents dans les entrées géographiques de deux ouvrages du XVIIIeme siècle : l'Encyclopédie de Diderot et d'Alembert et le Dictionnaire Universel Français-Latin de Trevoux DUFLT (Vigier et al., 2022), afin de **structurer les informations extraites sous forme de graphes**. Une partie du travail sera donc consacrée à la **définition du schéma de données pour la construction du graphe** (définition et modélisation des entités et des relations). Nous utiliserons des frameworks et langages orientés graphe comme RDF ou Neo4j.

Pour peupler le graphe de connaissances, la personne recrutée **expérimentera et évaluera l'utilisation de LLMs pour la construction automatique du graphe** à partir du texte (Lairgi et al., 2024) ou à partir de données annotées par un modèle de classification de tokens (Gabay et al., 2022 ; Brenon et al., 2022). En particulier, elle pourra s'appuyer sur un modèle BERT fine-tuné (Devlin et al. 2018) et disponible sur HuggingFace[5]. Ce modèle permet entre autres le repérage des noms de lieux et des relations spatiales. Concernant les

[1] https://fil.cnrs.fr
[2] https://eric.msh-lse.fr
[3] https://liris.cnrs.fr
[4] https://icar.cnrs.fr
[5] https://huggingface.co/GEODE/bert-base-french-cased-edda-ner

LLMs, les frameworks LangChain et LlamaIndex pourront être testés et évalués. L'objectif est d'évaluer les performances des LLMs génératifs (des modèles GPTs d'OpenAI mais également des *Open source small* LLMs locaux tels que Llama, Mistral ou Phi3) en *zero* ou *few-shot learning*, c'est-à-dire sans affinage (*fine tuning*). Enfin, une étape importante du travail sera consacrée à développer une méthodologie de validation et d'évaluation du graphe ainsi qu'à la valorisation des résultats obtenus.

Bibliographie

Brenon A., Moncla L., McDonough (2022). Classifying encyclopedia articles: Comparing machine and deep learning methods and exploring their predictions. Data and Knowledge Engineering, Elsevier. https://doi.org/10.1016/j.datak.2022.102098

Devlin, J. Chang, M.-W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gabay, S., Suarez, P. O., Bartz, A., et al. (2022), "From FreEM to D'AlemBERT: a Large Corpus and a Language Model for Early Modern French". arXiv preprint arXiv:2202.09452, 2022.

Lairgi, Y., Moncla, L., Cazabet, R., & Benabdeslem, K. (2024) Knowledge Graph Construction Using Large Language Models. Journée Nationale sur la Fouille de Textes. International Web Information Systems Engineering conference (WISE) Doha, Qatar arXiv preprint arXiv:2409.03284

Tual, S. T., Abadie, N., Duménieu, B., Chazalon, J., & Carlinet, E. (2023). Création d'un graphe de connaissances géohistorique à partir d'annuaires du commerce parisien du 19 ème siècle: application aux métiers de la photographie. In *IC 2023, 34es journées francophones d'Ingénierie des connaissances*.

Vigier D. Moncla L. & al. (2022). Geography articles in Trévoux's Dictionnaire Universel and Diderot and d'Alembert's Encyclopédie, Langue Française, n° 214 (2), 59-80.

Déroulement du stage

Profils recherchés : Master 2 Informatique / Ingénieur

Des compétences sont attendues en programmation, en web sémantique, en science des données (Machine Learning et Deep Learning). Des connaissances en traitement automatique de la langue (TAL) seront appréciées.

Rémunération : environ 630€ par mois

Lieu : Laboratoire LIRIS – INSA Lyon, Bâtiment Blaise Pascal, Campus La Doua, Villeurbanne.

Date de début : février/mars 2024

Durée : 5 à 6 mois

Candidature : Envoyer un mail présentant votre parcours, vos motivations ainsi que votre CV et vos derniers relevés de notes à : ludovic.moncla@insa-lyon.fr, fabien.duchateau@univ-lyon1.fr et frederique.laforest@insa-lyon.fr

Date limite de candidature : 17 novembre 2024