Lyon 1

Découverte de relations spatiales dans l'encyclopédie de Diderot & d'Alembert

Cahier des charges

Binôme : Trontin Dylan & Bruniquel Thomas
Encadrants : Fabien Duchateau & Ludovic Moncla

Contexte

De nombreuses connaissances sont actuellement stockées dans des documents textuels, ce qui rend leur exploitation par des machines difficile. L'**extraction d'information** permet de repérer automatiquement des entités et leurs relations dans ces textes [2].

Dans la phrase exemple suivante, on peut identifier Lyon (entité nommée) comme étant le chef lieu (type de relation) de la région Auvergne Rhône Alpes (entité nommée).

"Lyon est le chef-lieu de la région AURA"

Pour que ces données soient plus exploitables, il est important de relier les entités détectées à des bases de connaissances comme Wikidata ou GeoNames afin de lever les ambiguïtés et d'assurer une interprétation correcte des informations (Lyon étant aussi le libellé d'une ville ou de comtés aux États-Unis). Par exemple, la Figure 1 montre les entités correspond à la ville de Lyon (France) sur 2 bases de connaissances. Il est également crucial de classer les relations extraites (sous forme de texte) en un type de relation général et non ambigu (e.g., wikidata:capital-of pour la notion de chef-lieu).

Ce projet vise à créer des méthodes pour extraire et classer les relations entre entités, en s'appuyant sur des ontologies existantes et en évaluant l'efficacité des

algorithmes. Le jeu de données sera l'encyclopédie de Diderot et d'Alembert, un objet d'étude des Humanités Numériques de par son apport à la diffusion des connaissances pendant le siècle des Lumières [4].

Figure 1. Entité "Lyon" sur Wikidata (gauche) et Geonames (droite)

Objectifs

L'objectif principal de ce projet est d'associer automatiquement des relations extraites de textes à des types de relations prédéfinis, afin de rendre ces informations plus structurées et plus exploitables par des algorithmes de traitement des connaissances. Un prototype pour l'extraction d'entités existe déjà (disponible ici).

Le projet consistera en l'élaboration et la mise en œuvre de modèles de classification pour ces relations spatiales. Le projet sera majoritairement réalisé en Python car le langage est adapté et de nombreuses bibliothèques (notamment de *machine learning*) utiles à ce projet existent. Ce dernier suivra les étapes suivantes :

- **Construire une taxonomie des types de relations** en s'inspirant des ontologies spatiales existantes comme celles de Wikidata ou les relations spatiales DE-9IM, afin de pouvoir catégoriser les relations extraites de manière cohérente et systématique [1, 3].

- **Construire un jeu de données expertisé** (i.e., ensemble de relations avec leurs entités et un type de relation validé manuellement), qui servira d'entrée pour des algorithmes de classification supervisés.

- **Ajuster un algorithme déjà existant**. On utilisera dans un premier temps des algorithmes déjà existants de classification (la bibliothèque scikit-learn propose par exemple plusieurs algorithmes destinés à l'apprentissage automatique que l'on pourra exploiter). On essaiera alors d'adapter ces algorithmes pour répondre aux besoins de notre projet. L'idée est d'ensuite développer notre propre algorithme de classification comme décrit dans le point suivant.

- **Développer un algorithme de classification** qui permet d'identifier les relations connues. Les relations insolites ou peu fréquentes peuvent dans un premier temps être classifiées par défaut comme "autre" puis classifiées de façon pertinente avec une intervention humaine ou un algorithme plus précis. Cela permettrait de potentiellement enrichir la taxonomie avec de nouveaux types de relation. Dans un second temps on peut réfléchir à connecter l'algorithme à une ontologie plus large comme YAGO ou Wikidata. Cette approche est plus contraignante (recherche dans l'ontologie en temps réel pour identifier le type le plus proche en fonction du contexte) mais serait plus flexible si la taxonomie devait constamment évoluer.

- **Évaluer les performances de l'algorithme** à l'aide du jeu de données expertisé pour vérifier si l'algorithme est capable de correctement classer les relations. Les métriques standard d'évaluation comme la précision et le rappel seront utilisées pour mesurer ces performances. Cette évaluation nous permettra également d'ajuster les paramètres de l'algorithme pour optimiser les résultats.

Ces objectifs sont amenés à évoluer au cours de la réalisation du projet en fonction des difficultés techniques rencontrées ou des opportunités d'amélioration identifiées.

Calendrier

La date limite du projet étant le 23 février 2025, le projet sera donc mené de début novembre 2024 jusqu'à février 2025. Voici ci-dessous un diagramme de Gantt présentant un aperçu du déroulement du projet.

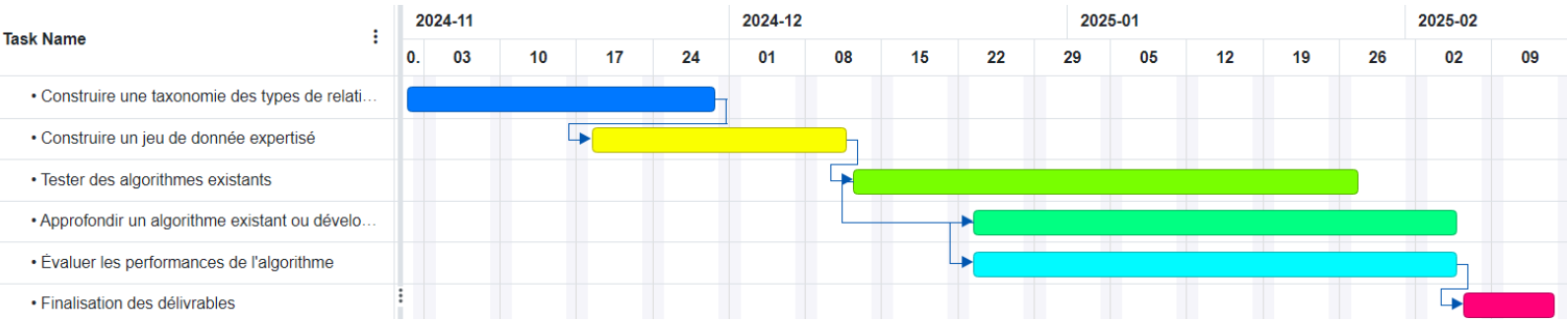

Diagramme de Gantt

Bibliographie

[1] C. Claramunt. Ontologies for geospatial information: Progress and challenges ahead. Journal of Spatial Information Science, 1(20):35–41, 2020.

[2] X. Han, T. Gao, Y. Lin, H. Peng, Y. Yang, C. Xiao, Z. Liu, P. Li, M. Sun, and J. Zhou. More data, more relations, more context and more openness: A review and outlook for relation extraction. JCNLP, 2020.

[3] F. Zhang, Q. Lu, Z. Du, X. Chen, and C. Cao. A comprehensive overview of rdf for spatial and spatiotemporal data management. The Knowledge Engineering Review, 36:e10, 2021.

[4] D. Vigier. *L'esprit encyclopédique moderne en France entre 1690 et 1902*. *Langue française*, 2022.