Détection d'anomalies basée sur les ontologies et l'apprentissage automatique

Noa Barbosa [1]

[1] Laboratoire d'informatique de Bourgogne, Université de Bourgogne, France
noa.barbosa@ube.fr

Contexte et problématique

L'analyse de données environnementales repose sur des sources **hétérogènes**, combinant des **données structurées** (mesures, campagnes d'échantillonnage, indicateurs quantitatifs) et des **documents textuels non structurés** (cf Figure 1). Si les données structurées peuvent être exploitées par des approches statistiques ou d'apprentissage automatique, l'extraction de connaissances à partir des textes demeure plus difficile. Dans ce contexte, les **ontologies** offrent un cadre sémantique pertinent pour représenter, formaliser et relier les connaissances d'un domaine. Le **peuplement d'ontologies** permet d'extraire automatiquement des entités et des relations à partir de sources variées afin d'instancier les concepts du domaine. Ce travail de thèse vise à **extraire** et **structurer** des connaissances relatives aux polluants à partir de textes, puis à exploiter ces connaissances dans une approche hybride combinant **raisonnement ontologique** et **apprentissage automatique** afin d'améliorer la **détection d'anomalies dans le domaine environnemental**.

Questions de recherche

- **Q1.** Comment effectuer de la **reconnaissance d'entités nommées** et de l'**extraction de relations** à partir de textes dans un domaine spécifique (p. ex. domaine environnemental) ?
- **Q2.** Comment **peupler** une ontologie avec les entités et relations extraites afin de **structurer des connaissances** spécifiques à un domaine ?
- **Q3.** Comment exploiter l'ontologie peuplée dans une approche hybride combinant **ontologies** et **apprentissage automatique** pour la **détection d'anomalies** dans un domaine spécifique ?

Verrous scientifiques et positionnement

Quatre familles d'approches peuvent être distinguées :

- **Approches à base de règles** : interprétables, mais coûteuses à construire et peu généralisables.
- **Apprentissage automatique** : bonne capacité de généralisation, mais forte dépendance aux données annotées.
- **Apprentissage profond** : performances élevées, mais nécessite de grands corpus annotés et reste sensible aux changements de domaine.
- **LLM** : réduit l'effort d'annotation, mais encore faiblement intégrés aux ontologies.

Approche

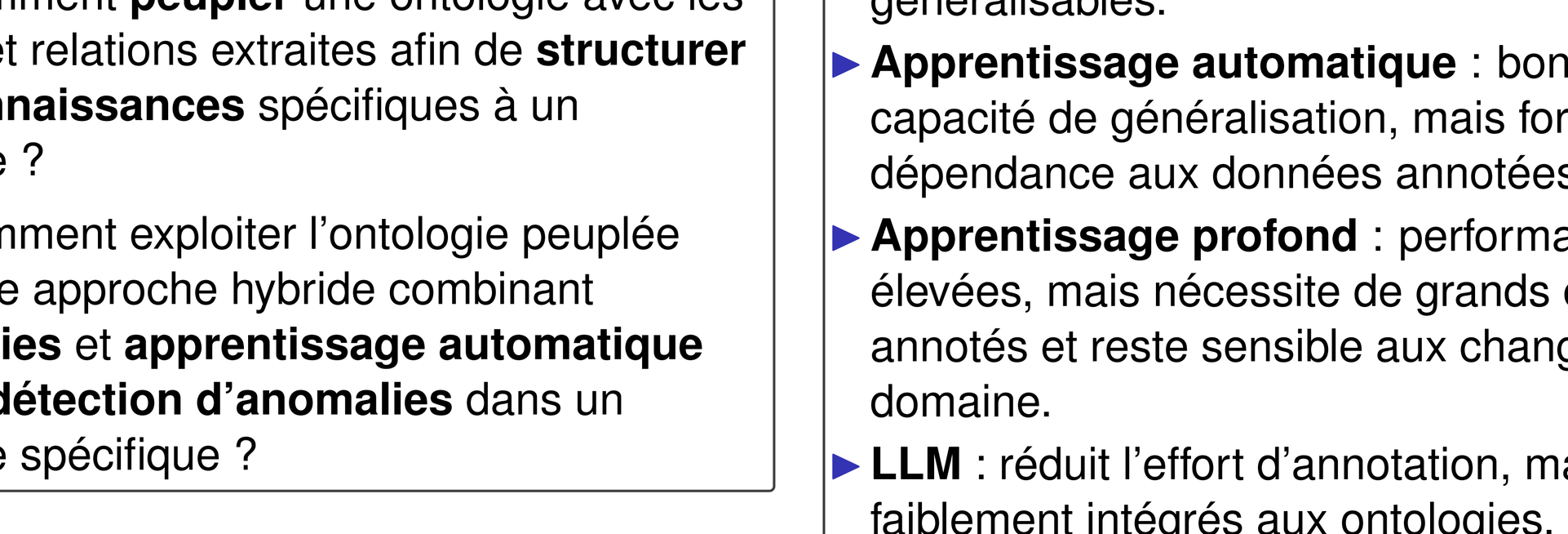

Figure 1: Partie haute: Peuplement de l'ontologie, Partie basse : Détection d'anomalies