Ontology alignment classifier based on NLP

Applied Data Science Project

Emanuele Pietropaolo
emanuele.pietropaolo@repertorio.ai

Who we are

Understanding the patient's voice is crucial, but Patient Preference Studies are fragmented, complex to analyze, and difficult to apply.

Repertorio is the intelligent platform that transforms this data into strategy. We accelerates the discovery, analysis, and application of preference studies for informed R&D, market access, and clinical decisions.

We are an innovative start-up based in Turin, born as an academic spin-off from the University of Turin and the University of Padova.

Repertorio

UNLOCKING PATIENT PREFERENCES

Value-driven project

Despite significant advances in AI and data engineering, aligning structured data semantically remains a persistent and computationally demanding task.

Organisations routinely encounter data from different sources that represent identical concepts in different ways.

This project investigates whether modern language models (such as Transformer-based architectures) can improve the automation of schema and ontology alignment.

Data

All experiments will be conducted using the **Ontology Alignment Evaluation Initiative (OAEI)** benchmark, and established international framework for evaluating semantic matching algorithms.

Data (called Track here):
- **Biodiversity and Ecology track (biodiv)**
  - It provides a benchmark for ontology matching within the environmental and biological sciences
  - It's a domain characterized by rapid data evolution and conceptual heterogeneity

Students are encouraged to explore alternative OAEI tracks if they wish to examine the generalisability of their approach, conduct comparative evaluations or investigate domain-specific modelling challenges.

Task

Main objective:
To design and implement a semantic alignment model capable of identifying conceptual correspondences between heterogeneous knowledge structures across different domains

Research focus
The project aims to explore whether modern language models, particularly *Transformer-based architectures*, can enhance the automation of schema, taxonomy, and ontology alignment

Evaluation Framework
The students will have access to a dedicated endpoint connected to the the **Matching EvaLuation Toolkit (MELT)** on their prediction to provide them with the same evaluation used in the conference

Light mentoring

- Paola Berchialla - Department of Clinical and Biological Sciences, University of Torino

- Emanuele Pietropaolo - Data Science Engineer at Repertorio

CENTRE FOR BIOSTATISTICS,
EPIDEMIOLOGY AND PUBLIC HEALTH

Privacy policy

Both project descriptions and implementations will be part of a repository group published on GitHub

The repositories will be public unless particular requests from the organizations that will be discussed

Ideally, the projects should be conceived open in the design

Thank you for your attention

Questions?

Repertorio
UNLOCKING PATIENT PREFERENCES