Hybrid Ontology Alignment using DragonAI and BERTMap

Capstone Midterm Presentation

Gabriel Nixon, Harshit Soni, Aditya Desai

Project Group: Clover (Group 14)
Mentors: Benjamin M. Gyori, Buz Galbraith (Northeastern University)

Fall 2025

Research Question / Objective

- How can we improve ontology alignment by combining Retrieval-Augmented Generation (RAG) and structure-aware deep models?
- Objective: Achieve high-precision, high-recall mappings between biomedical ontologies (e.g., SNOMED, FMA) with minimal manual intervention.
- Approach: A two-stage hybrid pipeline combining DragonAI (RAG) for completion and BERTMap for structure-aware matching.

Methodological Overview

Pipeline Stages

1. **DragonAI (RAG):** Impute/complete missing ontology descriptions via retrieval-augmented generation over a vector index.
2. **Candidate Generation:** Mix lexical and embedding top-k candidates for coverage and diversity.
3. **Pre-cleaning:** Drop disjoint/type-incompatible pairs (BERTMap-style constraints).
4. **Scoring:** Compute mapping probabilities using BERTMap.
5. **Thresholding & Validation:** Keep top-n/above-threshold; validate with LLM context and BioRegistry cross-checks.

Planned Experiments and Evaluation

- **Prior:** SeMRA (Raw Semantic Mappings Database).
- **Benchmarks:** OAEI ontology mapping datasets(Gold Standard).
- **Primary Metrics:** P@1, R@1/3, F1, Accuracy.
- **Secondary:** Unsatisfiable classes, cycle counts, expert acceptance rate for 1→N mappings.
- **Validation:** Compare to gold standards; ablations on RAG-only, BERTMap-only, and hybrid.

Progress and Timeline

Progress so far

- Baselines established for DragonAI and BERTMap, preliminary thresholds explored.
- Dataset for MVP selected:
  - Human Phenotype Ontology (HPO):
    Classes: 31,860 Roots: 522 Max Depth: 17
  - Mammalian Phenotype Ontology (MP):
    Classes: 36,182 Roots: 540 Max Depth: 34
- Candidate generation pipeline (lexical + embedding) prototyped.

Next Steps

- Integrate BERTMap and DragonAI Pipeline for integration test.
- Modify the BERTMap pipeline to address our research question.
- Run OAEI benchmarks; perform error analysis and expert review.

Summary

- Hybrid RAG + BERTMap pipeline to improve recall and reliability in ontology alignment.
- Designed for reproducibility and reduced manual effort.
- Upcoming: full benchmarking, validation, and packaging.

Thank You!