Universidade de Coimbra
Faculdade de Ciências e Tecnologia
Departamento de Engenharia Informática

Semantic and Natural Language Technologies
2024/2025

Masters in Informatics Engineering
Masters in Data Science and Engineering

Project

Professor:
Hugo Gonçalo Oliveira
hroliv@dei.uc.pt

Attention: plagiarism will not be tolerated!

1 Goals

The main goal of the Semantic and Natural Language Technologies project is **to tackle a challenge** in the domain of the topics covered by the course. Its high-level requirements are that the challenge involves both:

- **Semantic Web Technologies**
- **Natural Language Processing**

As long as their utilisation is well-supported, any computational tools can be used, as well as data sources, which can be on any domain.

2 Stages

The project should be developed by groups of **two** students, must go through the following **stages**:

1. **Research & Problem Identification:** definiting the challenge to tackle, supported on scientific literature. Two deliverables are expected, both of which can be later integrated in the **final report**:
    * **Literature Review:** a short report, with no more than **three pages in a 2-column format**[1], reviewing one or more recent papers on the topics of the course. At least one paper should be less than 5 years old and related to the project. Among others, the review must highlight the connections with the topics covered by the course.
    * **Project Proposal:** a single-page document describing the problem to tackle, setting the **goals** of the project and **dates of intermediate checkpoints**. The proposal must further enumerate, clearly, the **data** and **tools** to be used, proposed approach(es), as well as connections with the reviewed literature (e.g., initial inspiration, what will be made different) as well as with the topics covered by the course.
2. **Project Execution:** comprising the development of the project (design, implementation, evaluation), always **meeting the checkpoints** set and their discussion with the Professor.

[1] Formatted according to the IJCAI guidelines, `https://www.ijcai.org/authors_kit`

3. **Final Report:** all the work developed must be described in a **final document**, in the form of a scientific paper, with no more than **eight 2-column pages**[2]. The following structure is suggested:
   (a) Introduction (including motivation, problem description, goal and approach; may be adapted from the Project Proposal);
   (b) Related Work (for which the Literature Review can be a starting point);
   (c) Data & Approach (an overview of the data used and of the approach adopted for tackling the problem);
   (d) Implementation (including data exploitation, applied algorithms and tools used);
   (e) Experimentation (examples of good and bad results, quantitative evaluation when possible, and their discussion);
   (f) Conclusions (stressing the main achievements and main difficulties faced) and Future Work (what could still be done);
   (g) Bibliographic References.
4. **Final Presentation:** in the end of the semester, there will be 10-minute (tentative) presentations of the projects developed, followed by a brief discussion between the Professor and colleagues (if present).

All documents can be **written either in Portuguese or English**.

3 Evaluation Criteria

Overall, the project is worth **65%** of the final grade of the Semantic and Natural Language Technologies course (**13/20**), split in the following:

- Research & Problem (**3/20**, incl. Literature Review, Project Proposal);
- Checkpoints (**3/20**, **at least two**, assessed in Project Support classes);
- Final (**7/20**, incl. Final Report, Presentation).

Evaluation will focus on, but not be limited to, the following aspects:

- Meeting high-level **requirements**.
- Meeting **checkpoints** and **goals** set in the Project Proposal.
- Clarity of the produced **documentation**.
- Proper application of **course-related concepts**.

[2] Ideally, seven pages for content plus one for bibliographic references.

- **Novelty** of the proposed solution.
- **Complexity** of the developed work.
- **Quality** and **meaningfulness** of results / discussion / conclusions.
- **Final presentation** and **discussion**.

Further remarks:

- **Missing a deadline** will result in a **penalty**. This includes the **submission** of the Literature Review, Project Proposal, Final Report, and **checkpoints**.
- The **Final Presentation is mandatory**.
- Even though the project is to be developed in a group, the final grade may be **different for elements of the same group**, depending on individual performance.

4 Deadlines

The project has **two** main (hard) deadlines, with results submitted to *Inforestudante*:

- [**18**th **October 2024**] Literature Review & Project Proposal
- [**6**th **December 2024**] Final Report

Two dates must be set by the group in the Project Proposal for checkpoints, where progress will be presented to and discussed with the Professor. These can be selected among the following dates,

- October 25, 29
- November 5, 12, 19, 26
- December 3

Both members of the group **must** be present in all checkpoints.

Oral presentations should coincide with the final classes, but additional slots may be necessary to accommodate for all groups:

- 9th **December 2024**, 4PM–6PM
- 10th **December 2024**, 2PM

Attention: plagiarism will not be tolerated!

A Suggested Topics

As long as the main requirements are met, the group is free to choose the topic and goals for their project, which should, nevertheless, be discussed with and validated by the Professor. Here are some generic suggestions of possible topics:

- Knowledge Graph Enrichment
  - Select any existing Knowledge Graph
  - Enrich it with entities and / or triples extracted from textual documents [1] or from LLMs
  - Analyse the quality of the extracted knowledge
- Knowledge Graph Extraction or Ontology Learning from LLM
  - Explore prompt engineering for triple extraction [2]
  - Represent extracted knowledge in RDF
  - Evaluate the results
  - May tackle a related public challenge on Knowledge Extraction and LLMs (e.g., LLMs4OL [3], LM-KBC [4])
- Knowledge Injection in Neural Model / LLM [5]
  - Select an NLP task for which an annotated dataset is available (e.g., Text Classification)
  - Explore a LLM for performing the task (e.g., zero-shot few-shot, fine-tuning)
  - Select one or more Knowledge Graphs for providing additional features to inject in the process
  - Analyse the impact of knowledge injection
- Conversion of any dataset to RDF [6]
  - Select any lexicon or dataset
  - Convert this dataset to a RDF-based format (i.e., define classes, properties, consider reusing an existing vocabulary, etc.)
  - Explore the advantages / disadvantages of this format (e.g., query the dataset with SPARQL, consider linking it to other datasets in the Linked Open Data Cloud [7])
  - Use the resulting dataset in some experiment (e.g., classification, question answering)
- Knowledge Graph Embedding [8]
  - Select an NLP task for which an annotated dataset is available and a knowledge graph can be useful (e.g., Question Answering)

- Embed the selected knowledge graph
- Analyse the advantages of using the graph directly or its embeddings when performing the task.

The work may also combine features from more than one suggestion.

References

[1] Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, 56(11):1–39, 2024.

[2] Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*, 2022.

[3] Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. Llms4ol: Large language models for ontology learning. In *International Semantic Web Conference*, pages 408–427. Springer, 2023.

[4] Sneha Singhania, Tuan-Phong Nguyen, and Simon Razniewski. Lm-kbc: Knowledge base construction from pre-trained language models. *the Semantic Web Challenge on Knowledge Base Construction from Pre-trained Language Models*, 2022.

[5] Ariana Martino, Michael Iannelli, and Coleen Truong. Knowledge injection to counter large language model (llm) hallucination. In *European Semantic Web Conference*, pages 182–185. Springer, 2023.

[6] Manuel Fiorelli and Armando Stellato. Lifting tabular data to RDF: A survey. In *Metadata and Semantic Research: 14th International Conference, MTSR 2020, Madrid, Spain, December 2–4, 2020, Revised Selected Papers 14*, pages 85–96. Springer, 2021.

[7] Florian Bauer and Martin Kaltenböck. Linked open data: The essentials. *Edition mono/monochrom, Vienna*, 710(21), 2011.

[8] Xiou Ge, Yun Cheng Wang, Bin Wang, C-C Jay Kuo, et al. Knowledge graph embedding: An overview. *APSIPA Transactions on Signal and Information Processing*, 13(1), 2024.