Projects/Thesis proposals for the LDKR course

Introduction

As part of the LDKR exam, students are asked to carry out a project (possibly leading to a master thesis) in which the modeling part is done using a logic language. The choice of the most appropriate language to use should be done according to the expressiveness required by the problem at hand. Each project is assigned to groups of *no more than 2 students*.

There are 3 possible modalities:

- *short* (only exam and short project)
- *medium* (exam + research topics, or exam + thesis)
- *long* (exam + research topics + thesis)

Each group will be assigned to an advisor who will assist the students and evaluate the project/thesis.

Projects/Thesis proposals

1. Convert GeoWordNet in RDF (short) (Bisu, Enzo) ALREADY ASSIGNED

GeoWordNet[1] is a multilingual geo-spatial open source ontology built from the full integration of WordNet, GeoNames and the Italian part of MultiWordNet. It is available as a dump of an SQL database. The goal is to (a) formalize the problem and (b) convert GeoWordNet in RDF format.

2. Convert GeoWordNet in WordNet format (short) (A. Autayeu) ALREADY ASSIGNED

GeoWordNet is a multilingual geo-spatial open source ontology built from the full integration of WordNet, GeoNames and the Italian part of MultiWordNet. It is available as a dump of an SQL database. The goal is to (a) formalize the problem and (b) convert GeoWordNet in WordNet format.

3. Formalization of Wikipedia (short, medium, long) (Enzo)
ONLY MEDIUM AND LONG ARE AVAILABLE

Wikipedia is a collection of pages each informally describing a concept (e.g. university) or an entity (e.g. University of Trento). The goal of the thesis is (a) to collect some hundreds of pages and distinguish those *denoting a particular entity type*, for instance locations, organizations and people; (b) look at the Wikipedia info-boxes (usually represented as a table on the right of the page) to determine the set of typical attributes (an the words used to express them) associated to the entities according to the entity type; (c) formalize such entities by defining the necessary entity types, the relations, the attributes, the entities and the concepts. Each group will focus on a specific entity type.

4. Formalization of YAGO (short, medium, long according to the quantity of facts) (Enzo)
ONLY MEDIUM AND LONG ARE AVAILABLE

YAGO is a collection of facts each describing a relation between two objects, each of them representing a concept (e.g. university) or an entity (e.g. University of Trento). The goal of the thesis is (a) to analyze some thousands of facts provided by the advisor and formalize them in the corresponding

[1] http://semanticmatching.org/background-knowledge-datasets.html

TBox and ABox using DL (final format to be agreed). As part of the analysis, the corresponding entity types have to be identified (see previous thesis).

5. Providing support for the IJCAI social network (long) (Fausto)

The IJCAI social network is the group of people, the set of tools and the background knowledge built around the IJCAI conference on Artificial Intelligence (AI). Starting from the papers available online from the websites of the previous editions of the conference, the goal is to enrich the background knowledge with a dictionary of terms useful to support the activities of the people in the social network, e.g. to support indexing and search of the material and understand the AI terminology.
This must be installed as a service of both IJCAI and Entitypedia[2]. In fact, it will be used to enrich the background knowledge of any site

6. DBLP importing (medium, long) (Enzo)

DBLP (Digital Bibliography & Library Project) is the famous computer science bibliography website listing more than 1.3 million articles on computer science and corresponding metadata, including in particular *authors* and *topics*. Related to the previous thesis, the goal is to analyze, formalize and import the content of DBLP into a relational database (the schema has to be agreed) in which the formal relationships between the single entities are made more explicit and formal.

7. Use DERA to formalize a specific domain (medium, long) (Bisu)

DEPA (Domain-Entity-Property-Action) is a framework used in libraries to organize the knowledge of a domain. For instance, in Medicine (D) the main topics are the diseases (P) which affects the body parts (E) and the actions to prevent/cure them (A). This framework is only informally described, for instance in [Bhattacharyya, G. "POPSI: its fundamentals and procedure based on a general theory of subject indexing languages", Library Science with a Slant to Documentation, Vol. 16 No. 1, March, pp. 1-34, 1979]. DERA (Domain-Entity-Relation-Attribute) is a *semantic framework* developed at University of Trento evolving DEPA. The goal is to use DERA to formalize a specific domain.

8. Element Level matchers (medium, long) (Aliaksandr Autayeu)

Given any two graph-like structures, e.g. classifications, database or XML schemas and ontologies, matching is an operator which identifies those nodes in the two structures which semantically correspond to one another. Element level (EL) matchers act at the level of single elements. They have different characteristics according to the kind of objects to compare, e.g. strings, numbers, formal objects like concepts and entities. The performances of a matcher heavily depend on the way the EL matchers are combined together. The goal is to investigate if it is possible to arrive at an automatic configuration that maximizes the accuracy of the matcher and how this is influenced by the quality and quantity of the background knowledge available.

9. Time gazetteers, time representation and reasoning (long) (Fausto)

A gazetteer contains entities of a given kind and corresponding definitions. The goal is to develop a time gazetteer, i.e. containing time entities such as Christmas or Ramadan, including a formal definition and a set of services to manage such entities and in general time intervals and moments.

[2] http://entitypedia.org

Think for instance to the agenda problem in which one could organize two events at the same date/moment.

10. Generate cross word from Entitypedia (long) (Fausto, Aliaksandr)

Starting from the definitions of concepts and entities in Entitypedia, generate crosswords so that people can learn about a give topic. Each crossword is parameterized by:

1. Dimension (8x8, 9x9, YxY)
2. Difficulty
3. Topic (a specific class or etype)

Furthermore:

1. The user can ask for help
2. A user can suggest that there is a mistake and suggest correction
3. In case of suggested correction, it gets submitted to other users for final verification. If approved it gets inserted in Entitypedia

11. Semantic Tag Cloud Semantic labeling for web pages (long) (Aliaksandr, Fausto)

The work is as follows:

1. Take in input a set of pages
2. Extract words, entities and concepts from the text
3. Map known words to Entitypedia
4. Generate tag cloud with most used concepts with links to the pages where they occur
5. It must be possible to navigate Tag cloud at different level of specificity
6. Need an expert graphic designer to have hints about how to visualize tag cloud
7. It must be run as a service of Entitypedia

Workplan

The following workplan is supposed to be followed in general:

	Description
STEP 1	State of the art assessment when applicable
STEP 2	Formalization of the problem in logic using protégé
STEP 3	Implementation using the Java programming language and PostgreSQL DBMS if needed