LIRMM SIFR project

Internship position
Master in computer science
2018-2019

Title:

Machine-learning based automatic assignment of Semantic Types to biomedical concepts

Information :

Supervisor:	Clement Jonquet (LIRMM, University of Montpellier) – jonquet@lirmm.fr
Profile:	Computer science or (bio)informatics master students
Context:	Project SIFR (Semantic Indexing of French Biomedical Data Resources). Collaboration with LGI2P (EMA): Andon Tchechmedjiev)
Where:	University of Montpellier, Laboratory of Informatics, Robotics, & Microelectronics of Montpellier (LIRMM)
When:	2nd semester 2018-2019

Keywords:

Semantic Web, biomedical ontologies, knowledge representation, machine learning, classification.

Technologies:

BioPortal, UMLS Metathesaurus and Semantic Network, Semantic Web technologies (RDF, OWL, SKOS), Machine learning or classification framework (TensorFlow, Weka, etc.)

Abstract:

A key aspect in addressing semantic interoperability for life sciences is the use of terminologies and ontologies as a common denominator to structure biomedical data and make them interoperable. Ontologies formalize the knowledge of a domain by means of concepts, relations and rules that apply to that domain [1]. Stanford University has invested lot of efforts in developing terminology/ontology-based tools and services to assist health professionals and users in their search for electronic information available on the Web and in the use of ontologies. The group has developed a Web-based portal, the *NCBO BioPortal* (http://bioportal.bioontology.org) [2] that offers a variety of services to search or index biomedical data as well as searching, exploring, annotating and visualizing the available standards ontologies. In parallel, we develop the SIFR BioPortal (http://bioportal.lirmm.fr), a similar resource but dedicated to French [3]. These two platforms exploit terminologies extracted from the UMLS Metathesaurus [4] and value the Semantic Types [5] that are assigned to the concepts of these terminologies. This typing, done manually by experts, allows to manipulate through all the resources of the UMLS only certain types (virus, tissue, chemical, etc.) of concepts. The Semantic Network offers 133 Semantic Types and they have been grouped also within 15 Semantic Groups (anatomy, objects, procedures, etc.) [6]. However, this "coarse" typing is only available for UMLS resources [7].

The internship aims to develop an automatic classification component to assign concepts of any terminology or biomedical ontology one or more Semantic Types. To do this, we will adopt a supervised machine learning approach that will use already tagged resources as training and test corpus. We will identify the features (e.g., labels, label-patterns, source ontology, hierarchy, etc.) that will help to classify new concepts and start first by assigning semantic groups, then types. We will first focus on French resources (in the SIFR BioPortal) and then generalize on a larger scale to resources in English (or other language, in the NCBO BioPortal). The internship will result in a web application prototype that eventually will be incorporated into the NCBO technology.

Intern Tasks:

The intern tasks will consist of:

- Reviewing the papers describing the context of the project
- Select a machine learning framework and identify the classification features
- Extract the relevant training/test data from UMLS
- Implement a methodology to automatically classify concepts from a new ontology (Semantic Groups first, then Semantic Types)
- Evaluate the results using cross validation
- Enrich the existing ontologies and terminologies in the SIFR BioPortal and involve their developpers for validation
- Write a publication about the project and its outcomes

References:

1. Gruber, T.R.: A translation approach to portable ontologies. Knowl. Acquis. 5, 199–220 (1993).
2. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N.B., Jonquet, C., Rubin, D.L., Storey, M.-A., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 37, 170–173 (2009).
3. Jonquet, C., Annane, A., Bouarech, K., Emonet, V., Melzi, S.: SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique. In: 16th Journées Francophones d'Informatique Médicale, JFIM'16. p. 16. , Genève, Suisse (2016).
4. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32, 267–270 (2004).
5. McCray, A.T.: An Upper-Level Ontology for the Biomedical Domain. Comp. Funct. Genomics. 4, 80–84 (2003).
6. McCray, A.T., Burgun, A., Bodenreider, O.: Aggregating UMLS semantic types for reducing conceptual complexity. Stud. Health Technol. Inform. 84, 216 (2001).
7. Tchechmedjiev, A., Jonquet, C.: Enrichment of French Biomedical Ontologies with UMLS Concepts and Semantic Types for Biomedical Named Entity Recognition Though Ontological Semantic Annotation. In: Workshop on Language, Ontology, Terminology and Knowledge Structures, LOTKS'17. , Montpellier, France (2017).

UMLS Semantic Network: https://semanticnetwork.nlm.nih.gov/

Current Semantic Types: https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

Expected profile:

- Computer science or (bio)informatics master degree students.
- Experience with machine learning tools and motivation to learn more.
- Experience with semantic Web technologies will be appreciated but is not mandatory.
- Good English oral and writing skills. Good knowledge of French or motivation to learn is desirable.
- Excellent writing skills as reports, documentations, and technical notes will always be necessary.
- Autonomy and initiative, take on technical decisions within the project and justify choices.
- Friendly person to join a small research team in Montpellier.

LIRMM SIFR project

Internship position
Master in computer science
2018-2019

Application:

For more information about this position, please contact Clement Jonquet (jonquet@lirmm.fr) and Andon Tchechmedjiev (andon.tchechmedjiev@mines-ales.fr). To apply, please send an email including links to (Please NO ATTACHED DOCUMENTS) the following:

- a motivation letter describing an explanation of YOUR interest for the intern;

- a curriculum vitae describing your experience and the matches with the expected profile;

- names and contact details of referees.

Date are flexibles over the 2018-2019 scholar year.