PhD Research Project:
Enhancing Scalable Medical Ontologies and Reasoning

Research Fields: Knowledge Representation and Reasoning for Medicine and Health, Medical Text Analysis

Supervisors and Laboratory:

- Pascal VAILLANT (pascal.vaillant@univ-paris13.fr), Associate professor in Computer Science, LIMICS
- Chan LE DUC (chan.leduc@univ-paris13.fr), Full professor in Computer Science, LIMICS

Scientific Context: Intelligent systems need ontologies to represent structured knowledge from different data sources. The ontologies should be expressed in a language equipped with formal and unambiguous semantics such as Description Logics [1]. Such a semantics allows to develop automated procedures, on the one hand, for checking the usability of an ontology (e.g. consistency), and on the other hand, for exploiting an ontology (e.g. entailment, answering queries) [5].

In the last two decades, numerous biomedical ontologies have been designed to provide standardized terminologies used for recording clinical details of patients. For instance, SNOMED-CT is one of the most complete biomedical ontologies and covers various medical domains such as diagnostics, diseases, medications, anatomy, and procedures. A large volume of a recent version of SNOMED containing 359,017 classes and 729,496 subclass axioms, motivates researchers in the field to address the scalability issue. The main use of medical ontologies so far is to offer clinicians a nomenclature for creating medical documents which can be exchanged between different health care providers and researchers. However, the use of medical ontologies as OWL ontologies allowing for powerful reasoning tasks remains very limited. For instance, one can straightforwardly use an OWL reasoner to check whether a class defined in the ontology is subsumed by another one, but it is less obvious to use such a reasoner to discover potentially conflicting portions of knowledge such as AbleToUseMedication, UnableToUseMedication, AllergicTo because of the absence of individuals and negated classes, properties from the ontology. This example shows that enhancing a medical ontology by populating some classes can make it more exploitable in terms of reasoning.

In order to enhance such a medical ontology by populating it, one can use electronic health records (EHR) in natural language since they are a huge data source on diagnostics, medications, medical family history of patients. This knowledge is adapted to be read by a human, but is not easily adapted to tasks such as bulk querying, since the information is not structured. Extracting structured information from EHR is a task that has received a lot of interest from the medical knowledge engineering community [6].

Objectives and Approaches: The research project of this PhD attempts to address the following two main issues:

1. Populating a scalable medical ontology with knowledge from electronic health records (EHR) in natural language [2]. This is the first task: it consists in extracting the knowledge contained in EHR in the form of natural language sentences like "The patient has allergies to erythromycin, which causes a rash" and converting it to a structured format, formally equivalent to assertions such as: AllergicTo (#PatientId, SCTID 30427009). This task involves a phase of *semantic annotation* of the text: mentions of entities of interest are spotted in the text, and then linked to a standard concept identifier in a formal ontology (this is crucial so that the same concept (e.g. 387458008) is recognized under different possible character strings (e.g. "Aspirin", "Acetylsalicylic acid"). Then before generating assertions, a careful second phase should discriminate between mentions of facts that are asserted and mentions of facts that are negated or simply stated hypothetically. Achieving state-of-the-art information extraction requires a mixture of various approaches (traditional NLP methods and machine learning methods).
2. Developing new methods for reasoning on scalable ontologies. This objective consists in proposing a new semantic foundation for ontological languages such that it would lead to reducing computational complexity of reasoning algorithms. For this purpose, we will consider an approach which uses category theory [3] to define the semantics of ontological languages in place of set theory. In this setting, a reasoning algorithm needs to build just relationships between concepts instead of a model composed of individuals for representing concepts. A preliminary investigation of this approach [4] is conducted and shows that the new semantics would allow us to reduce reasoning complexity in space.

Expected Results: First, the successful candidate should begin by carrying out a state of the art on different approaches of text analysis, and methods of reasoning on OWL ontologies. Next, she/he should propose a new method

based on a new idea or a combination of existing approaches for analyzing EHRs. She/he should show that the proposed method improves the state of the art in the research field on medical text analyzing. Regarding the second objective, the successful candidate should propose categorical definitions of the semantics of basic logical constructors and a reasoning procedure based on the new semantics. She/he should implement the algorithms resulting from her/his research and perform experiments on real-world datasets (e.g. EDS, MIMIC) for evaluating her/his results.

It is expected that the new methods and experimental results will be published in good conferences/journals in the fields of NLP, Semantic Web and Knowledge Representation and Reasoning.

Expected Qualifications and Skills:

- Master's degree in Computer Science, Medical Informatics or Applied Mathematics
- Interactions with other researchers
- Good knowledge of formal representations and techniques related to the Semantic Web and text analysis
- Ability to design and program in Python, Java or C++.

Provisional Schedule: This PhD program can be organized into overlapped periods according to different tasks. For instance, research on analysis of texts and reasoning algorithms based on category theory should be performed in parallel all along the program.

- Task 1: The PhD student investigates the state of the art on existing methods and tools for analyzing texts and generating OWL assertions and axioms. She/he should focus on specific EHRs to be able to develop new methods which are more adapted to them
- Task 2: The PhD student continues enhancing some selected medical ontologies by using existing methods and tools for analyzing specific texts, and starts to investigate a possible combination of some of them to get better results. For instance, the PhD can combine deductive methods such as cTakes[1] with inductive methods such as spaCy[2] for analyzing EHRs. From this, the PhD student proposes a new method for populating/enhancing medical ontologies from a larger class of medical texts. She/he should start to implement the method and perform evaluations of the results.
- Task 3: The PhD student performs research on Description Logics and category theory with the aim of proposing a category-theoretical semantics for usual logical constructors. Next, she/he develops a reasoning procedure for the inexpressive description logic $\mathcal{ALC}$ based on this new semantics, and extends it to other logical constructors. As usual, the PhD student should implement the method and perform benchmarks against the existing reasoners.

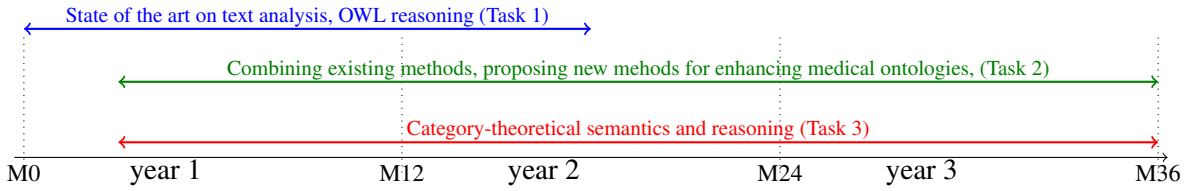

Figure 1: Provisional Schedule

REFERENCES

[1] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, USA, 2nd edition, 2010.

[2] S. Després, C. Duclos, C. Le Duc, and P. Vaillant. Enhancing medical ontologies with knowledge from discharge summaries. Technical report, 2020.

[3] R. Goldblatt. *Topoi : The categorial Analysis of Logic*. Dover Publications, 1984.

[4] C. Le Duc. Category-based reasoning in the description logic $\mathcal{ALC}$. Technical report, 2021.

[1] https://ctakes.apache.org
[2] https://spacy.io/

[5] J. Lhez, C. Le Duc, T. Dong, and M. Lamolle. Decentralized reasoning on a network of aligned ontologies with link keys. In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part I*, pages 418–434, 2019.

[6] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49, 2018.

Avis favorable

Marie-Christine Jaulent
DR Inserm
Directrice LIMICS UMR_S 1142
marie-christine.jaulent@inserm.fr

Pr. Sylvie DESPRES
Directrice adjointe