Sujet de thèse :
Enrichissement et Requêtage pour un Réseau d'Ontologies Médicales

Domaines de recherche : Informatique médicale, Web Sémantique, Raisonnement, Analyse de textes

Encadrants et laboratoire :

- Chan LE DUC (chan.leduc@univ-paris13.fr), Professeur des universités en Informatique, LIMICS
- Sylvie DESPRES (sylvie.despres@univ-paris13.fr), Professeur des universités en Informatique, LIMICS

Projet de recherche : Un système intelligent est souvent fondé sur des ontologies modélisant des connaissances issues de différentes sources de données. Ces ontologies doivent être formalisées dans un langage muni d'une sémantique formelle telle que les logiques de description [1]. Une telle sémantique permet de développer des procédures automatisées, d'une part, pour la vérification de l'utilisabilité d'une ontologie (e.g. la cohérence), et d'autre part, pour l'exploitation d'une ontologie (e.g. déduction, réponse à des requêtes) [5]. Pour qu'une ontologie ou un réseau d'ontologies alignées soit exploitable par des catégories d'utilisateurs plus spécifiques (médecins, patients, etc.), il est nécessaire de (i) peupler les ontologies, i.e. instancier des classes et propriétés présentes dans les ontologies [2] ; (ii) de fournir aux utilisateurs des patrons facilitant la composition des requêtes pour interroger les ontologies. En effet, la composition d'une requête formelle intéressante en SPARQL ou DL nécessite non seulement une bonne maîtrise de la logique formelle mais aussi une connaissance solide relative aux ontologies concernées. Par exemple, un médecin devant un patient ayant une angine à Streptocoque mais ne tolérant pas l'Amoxicilline, pourrait souhaiter trouver des médicaments ayant la même indication que l'Amoxicilline sans ses effets indésirables. Cependant, il ne serait pas en mesure de composer une requête en logique de description telle que :

$$\mathsf{Antibiotique} \sqcap \forall \mathsf{indication}.\exists \mathsf{indication}^{-}.\{\mathsf{Amoxicilline}\} \sqcap \forall \mathsf{effetIndesir}.\forall \mathsf{effetIndesir}^{-}.\neg\{\mathsf{Amoxicilline}\} \quad (1)$$

Les services de santé publique (AP-HP) mettent à la disposition des chercheurs du LIMICS l'accès à l'EDS (https://eds.aphp.fr/) qui inclut plus de 40 millions de compte-rendus médicaux (CRM) en texte brut, et un grand nombre de bases de données sur les patients, médicaments et maladies. Une utilisation envisageable des CRM est de servir à enrichir/peupler les ontologies provenant des bases de données à l'EDS ou d'ailleurs.

Le travail de cette thèse consiste en deux volets. Le premier volet vise à élaborer une méthode hybride d'analyse de textes fondée à la fois sur une technique inductive (e.g. apprentissage statistique) et un mécanisme de raisonnement déductif (e.g. raisonnement sur des ontologies). A l'issue de cette étape, les CRM seraient analysés rendant ainsi possible l'enrichissment/peuplement des ontologies médicales. L'outil d'analyse de textes SPACY (https://spacy.io/) est l'un des premiers travaux [3, 4] qui s'inscrivent dans cette direction. Le second volet porte sur la conception de patrons permettant de faciliter la composition des requêtes DL. Cela consiste à proposer les interfaces de requêtes qui correspondent à différents niveaux et intérêts des utilisateurs. Par exemple, un médecin pourrait avoir besoin de composer des requêtes qui sont susceptibles d'être plus complexes que celles d'un pharmacien.

Résultats attendus : En premier lieu, le candidat devra faire un état de l'art sur l'analyse de textes par l'apprentissage statistique et les méthodes de raisonement pour les ontologies en OWL. Notamment, il considérera en détail la construction de modèles d'apprentissage efficaces pour l'analyse des CRM. Puis, le candidat étudiera les systèmes à base d'ontologies et/ou de règles permettant de réduire la taille de jeux de données d'entraînement exigés par la méthode d'apprentissage statistique. A l'issue de ces études, le candidat proposera une méthode hybribe, entre inductif et déductif, permettant d'enrichir/peupler des ontologies à partir de données en texte brut. Afin de rendre les ontologies peuplées exploitables, le candidat devra concevoir et développer un ensemble de patrons de requêtes sur les ontologies obtenues. Il est attendu du candidat qu'il implémente les algorithmes obtenus à partir des résultats de sa recherche. De plus, le candidat devra procéder à des expérimentations sur des données réelles pour évaluer ses résultats de recherche.

Compétence attendues :

- Master en informatique ou informatique médicale
- Interactions avec d'autres chercheurs
- Bonne connaissance des représentations formelles et des techniques liées au Web Sémantique et l'analyse de textes
- Capacité à concevoir et programmer en Python et Java

RÉFÉRENCES

[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. Desprès. Construction d'une ontologie modulaire. application au domaine de la cuisine numérique. *Revue d'Intelligence Artificielle*, 30(5) :509–532, 2016.

[3] T. Gherasim, M. Harzallah, G. Berio, and P. Kuntz. Analyse comparative de méthodologies et d'outils de construction automatique d'ontologies à partir de ressources textuelles. In A. Khenchaf and P. Poncelet, editors, *Extraction et gestion des connaissances (EGC'2011), Actes, 25 au 29 janvier 2011, Brest, France*, volume RNTI-E-20 of *Revue des Nouvelles Technologies de l'Information*, pages 377–388. Hermann-Éditions, 2011.

[4] I. Lerner, N. Paris, and X. Tannier. Terminologies augmented recurrent neural network model for clinical named entity recognition. *J. Biomed. Informatics*, 102 :103356, 2020.

[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.

Marie-Christine Jaulent
DR INSERM
Directrice LIMICS - UMRS1142

Marie-Christine Jaulent
PhD, DR Inserm, Directrice du Limics
INSERM U1142, LIMICS
15, rue de l'École de Médecine
75006 Paris, France
+33(0)6 84 38 31 25
marie-christine.jaulent@crc.jussieu.fr