Postdoctorat au Loria : Extraction à partir de textes de relations pharmacogénomiques fines, en considérant les connaissances du domaine

–

Postdoctoral position: Extracting pharmacogenomic relationships from text, considering domain knowledge

Adrien Coulet, Chedy Raïssi, Yannick Toussaint
Equipe-projet Orpailleur, laboratoire LORIA-Inria, Nancy, France

Abstract

En Français: L'objectif de ce postdoctorat est de proposer de nouvelles méthodes d'extraction de connaissances à partir de texte, et de les appliquer à un cas d'utilisation réel en pharmacogénomique. Plus précisément, il s'agit (1) d'extraire des relations binaires, voir n-aires, typées, à partir d'articles scientifiques biomédicaux ; (2) d'étudier comment les bases de connaissances du domaine (sur les médicaments ou les maladies par exemple) peuvent être prises en considération lors de l'extraction pour améliorer ou guider ses résultats ; puis (3) de proposer une comparaison des unités de connaissances extraites avec les bases de données de référence en pharmacogénomiques, comme PharmGKB afin de proposer leur enrichissement. Nous proposons d'utiliser des méthodes de traitement du langage naturel qui s'appuient sur des approches d'apprentissage profond, comme les réseaux de neurones récurrents car celles-ci permettent d'extraire des relations complexes comme celles existantes en pharmacogénomique. Nous proposons également d'utiliser un encodage du contexte des entités qui permettent de considérer à la fois leur environnement lexical et des éléments structuraux définis dans des bases de connaissances (par exemple leurs parents, instances).

In English: The aim of the postdoc project is to propose novel methods for knowledge extraction from text, and to apply them to a real-world use case in pharmacogenomics (PGx). More precisely we aim at (1) extracting fine-grained and typed relationships about PGx from the biomedical literature; (2) studying how this extraction can take advantage of considering domain knowledge that is available in knowledge bases about drugs and diseases; (3) comparing these knowledge units with data from PharmGKB, the reference database for PGx and propose enrichments of how drug response phenotypes are documented. To achieve these three goals, we propose to use natural language processing methods that rely on deep learning approaches such as recursive neural networks. Considering the machine learning process, we would like to study embedding approaches that combine classical word embedding, where the textual context of entities is encoded, conjointly with structured embedding of knowledge bases, which proposes to encode elements of knowledge bases such as subsumption relationships or instantiation.

Durée : 12 mois, renouvelable
Début souhaité : Septembre 2016
Lieu : Loria – Inria Nancy Grand Est, Nancy (`http://www.loria.fr/`)
Equipe : Orpailleur (`http://www.inria.fr/equipes/orpailleur`)
Contacts :

Adrien Coulet, `prenom-point-nom-au-loria-point-fr`, 03 54 95 86 38
Chedy Raïssi, `prenom-point-nom-au-loria`, 03 83 59 20 79
Yannick Toussaint, `prenom-point-nom-au-loria`, 03 83 59 20 91

Contexte du postdoctorat

Ce postdoctorat s'inscrit dans le cadre du projet ANR (Agence Nationale pour la Recherche) intitulé PractiKPharma[1] (Practice-based evidences for actioning Knowledge in Pharmacogenomics) qui a débuté en 2016 [1]. PractiKPharma s'intéresse à un domaine biomédical particulier : la pharmacogénomique, qui étudie l'impact des facteurs génétiques sur la réponse aux médicaments. Dans ce domaine, les connaissances de l'état de l'art sont nombreuses et sont disponibles soit dans la littérature, soit dans des bases de données spécialisées, comme PharmGKB[2]. Ces connaissances ont classiquement la forme d'une relation ternaire entre un *gène*, un *médicament* et un *effet indésirable*, représentant le fait que les patients qui possèdent une version particulière du gène réagissent de façon inattendue au médicament en subissant un effet indésirable.

Parmi les connaissances de l'état de l'art, certaines sont bien connues et validées, mais d'autres n'ont été observées que sur de petits nombres de patients et leur statut reste incertain [2]. L'objectif du projet PractiKPharma est de fouiller les données de dossiers patients électroniques afin de confirmer, ou modérer, les connaissances de l'état de l'art qui ne sont pas complètement établies.

Au sein de ce projet, le postdoctorant proposera des méthodes d'extraction de connaissances à partir de texte et les mettra en oeuvre pour établir de façon plus précise et structurée les connaissances de l'état de l'art en pharmacogénomique. Ces connaissances devront être comparée aux données contenues dans la base PharmGKB, puis plus largement à des connaissances extraites de dossiers patients. Il semble pertinent poru faciliter la comparaison d'utiliser les ontologies de référence du domaine. Le context du projet impliquera la nécessité d'échanger avec un doctorant du Loria et les chercheurs de consortium PractiKPharma (LIRMM, Montpellier ; CHU Saint Etienne ; HEGP, Univ. Paris Descartes). Des échanges sont également prévus avec l'équipe de Russ Altman à l'Université Stanford qui développe la base PharmGKB .

References

[1] Adrien Coulet and Malika Smaïl-Tabbone. Mining electronic health records to validate knowledge in pharmacogenomics. *ERCIM News*, 2016(104), 2016.

[2] John P.A. Ioannidis. To replicate or not to replicate: The case of pharmacogenetic studies. *Circulation: Cardiovascular Genetics*, 6:413–8, 2013.

[1] Site web du projet PractiKPharma : `http://practikpharma.loria.fr/`
[2] Site web de PharmGKB : `http://www.pharmgkb.org`