Proposition de sujet de thèse

Informations sur la thèse

Titre	**Mining Host-Microbiome Interactions using Natural Language Processing**
Mots-clés	Extraction de relations, articles scientifiques biomédicaux, microbiome humain, traitement automatique des langues, plongements de mots contextualisés
Encadrement	Samuel Chaffron (équipe ComBi, LS2N) et Solen Quiniou (équipe TALN, LS2N)
Contact	samuel.chaffron@ls2n.fr, solen.quiniou@ls2n.fr
Partenaires	Caroline Trang-Poisson (CHU de Nantes), Antoine Roquilly (CHU de Nantes)

Résumé du sujet

Le sujet de thèse proposé porte sur l'extraction de relations entre espèces de microbiotes humains, et leurs relations avec des gènes, biomolécules, et maladies, à partir de la littérature scientifique, en utilisant des techniques de traitement automatique des langues. Nous nous intéresserons plus particulièrement aux articles traitant du microbiote intestinal et du microbiote pulmonaire et nous validerons les résultats obtenus grâce à une collaboration étroite avec des chercheurs du CHU de Nantes travaillant sur ces deux microbiotes et leur rôle dans les maladies inflammatoires.

Contexte de la thèse

Présentation de la problématique

Le microbiote intestinal humain est composé de milliers de bactéries commensales essentielles à notre santé et bien-être. L'acquisition progressive de ces communautés de bactéries (en interaction) chez le nouveau-né et l'enfant influence le développement du système immunitaire et la susceptibilité à certaines maladies. En effet, ces communautés commensales ont un impact fondamental sur la physiologie de l'hôte, principalement en aidant à la digestion et une absorption efficace des nutriments, en façonnant le système immunitaire, mais aussi en nous protégeant contre l'invasion d'agents pathogènes.
Aujourd'hui, les réseaux d'associations microbiennes inférés à partir de données de séquençage environnemental (ou métagénomique) ont révélé des interactions potentielles entre espèces et entre groupes taxonomiques plus larges [CRPvM10, FSI+12]. Cependant, l'existence et la nature de ces interactions sont difficiles à valider et à caractériser [LMFH+15, GCB+16]. Il existe ainsi un réel besoin de développer des méthodes efficaces pour fouiller la littérature scientifique (plus de 20 millions d'articles complets) afin d'extraire des interactions connues entre espèces du microbiote humain (telles que le mutualisme, le commensalisme, le parasitisme ou encore la prédation), en utilisant des approches de traitement automatique des langues. Une ressource recensant les interactions connues et annotées entre micro-organismes, mais aussi avec l'hôte, ses gènes et des maladies ou des molécules connues, serait très utile pour valider les associations prédites en utilisant les données de séquençage générées en masse dans les études cliniques du microbiome humain (e.g. Human microbiome project, French gut project).

Équipes impliquées

L'encadrement de la thèse regroupera deux équipes du LS2N : l'équipe TALN qui travaille sur le traitement automatique des langues, en particulier sur des documents écrits tels que les articles scientifiques, et l'équipe ComBi qui s'intéresse en particulier à la modélisation des communautés microbiennes et de leur interaction sous forme de graphes. De plus, ce sujet se développera en interaction forte avec 2 équipes de recherche du CHU de Nantes : l'équipe du Dr. Caroline Trang-Poisson, qui mène des recherches sur les maladies inflammatoires cliniques intestinales et en gastroentérologie, au sein de l'Institut des maladies de l'appareil digestif (IMAD), et l'équipe du Prof. Antoine Roquilly (EA3826), qui s'intéresse aux thérapies cliniques et expérimentales des infections pulmonaires au CHU de Nantes. Ils nous permettront de valider les approches proposées ainsi que les résultats obtenus.

Sujet de thèse

Contexte scientifique

L'extraction de relations [BB07] est une sous-tâche de l'extraction d'information dans laquelle des relations sémantiques sont extraites, à partir de textes non structurés, puis classées en différentes catégories. Différentes approches à base d'apprentissage automatique peuvent être utilisées : des approches supervisées, semi-supervisées ou encore non supervisées. D'un côté, les deux premiers types d'approches obtiennent de bons résultats mais nécessitent des données annotées, en quantité plus ou moins importante. D'un autre côté, les approches non supervisées n'utilisent pas de données étiquetées mais les relations extraites peuvent être très nombreuses et seule une petite partie d'entre elles correspondent à des relations intéressantes qui doivent être validées soit par des experts, soit en utilisant des bases de connaissance ou des ontologies, ce qui peut être difficile à réaliser.
En 2018, nous avons réalisé un premier travail pour constituer un corpus à partir d'articles scientifiques provenant de la plateforme ISTEX [1] et de bases de données biologiques constituées manuellement, portant sur des relations entre micro-organismes marins. Ce corpus a permis d'élaborer une première chaîne de traitement allant de l'étiquetage des mentions d'organismes présents dans celui-ci jusqu'à l'extraction de relations entre mentions d'organismes, en utilisant une approche s'appuyant sur de l'extraction de motifs séquentiels (apprentissage non supervisé).
Ces dernières années, de nouveaux modèles d'apprentissage profonds, s'appuyant sur des plongements lexicaux contextualisés, ont été proposés, tels que le modèle BERT [DCLT19] ou encore le modèle BioBERT [LYK+19], qui est une adaptation du modèle BERT aux données biomédicales. Ces nouveaux modèles obtiennent de très bons résultats pour de nombreuses tâches de traitement automatique des langues et notamment des tâches d'extraction d'information. De plus, un nouveau paradigme a récemment été proposé pour l'extraction automatique de relations : la supervision distante [SCM18, BZKA20] (*distant supervision*). Cette approche combine les avantages des approches semi-supervisées et non supervisées, en fournissant une méthode alternative pour générer, de manière automatique, des données étiquetées en utilisant une base de connaissances. En 2020, nous avons réalisé un deuxième travail, en reproduisant des résultats obtenus avec le modèle BERT, sur des corpus de l'état de l'art pour l'extraction de relations. Aujourd'hui, nous souhaitons continuer à développer ce type de modèle et aussi l'évaluer sur des données d'interactions entre micro-organismes associés aux organismes vivants (en particulier l'Homme).

Problématique

Dans le projet HOLONP, nous nous intéresserons aux microbiotes humains, et en particulier aux microbiotes intestinal et pulmonaire. Un des premiers objectifs du projet sera ainsi d'identifier les données disponibles sur ces deux sous-domaines. Dans un premier temps, nous pourrons ainsi reprendre l'approche non supervisée que nous avons mise en place, pour l'appliquer à ces nouvelles données.
Le deuxième objectif consistera à proposer de nouvelles approches, en tirant partie des modèles récents à base d'apprentissage profond, notamment ceux adaptés au domaine biomédical tels que BioBERT, mais aussi des modèles de supervision distante pour pouvoir adapter ces modèles, en utilisant des bases de connaissances existantes ou fournies par nos experts cliniciens. Comme le sous-domaine du microbiote intestinal comporte une bibliographie plus importante que celui du microbiote pulmonaire, nous pourrons tout d'abord nous focaliser sur celui-ci pour proposer des modèles d'extraction de relations adaptés. Nous pourrons ensuite nous reposer sur ces nouveaux modèles, pour les adapter au sous-domaine du microbiote pulmonaire, qui est bien moins caractérisé et pour lequel il existe donc beaucoup moins de données disponibles dans la littérature.
Un dernier objectif sera la mise à disposition des relations extraites, sous la forme de graphes de connaissances, qui pourront facilement être accessibles et questionnables au travers d'une interface web graphique, notamment pour les chercheurs du CHU de Nantes.

1. Cette plateforme donne accès à des millions d'articles scientifiques, avec leurs méta-données et leur contenu textuel, et permet également de sélectionner l'ensemble des articles correspondant à une requête.

Plan de travail

Le planning prévu pour le travail est le suivant :

- Étude de l'état de l'art sur l'extraction de relations, notamment dans le domaine biomédical et avec des approches d'apprentissage profond s'appuyant sur des plongements de mots contextualisés ainsi que des approches d'apprentissage distant avec supervision ;
- Constitution des corpus d'articles scientifiques des deux sous-domaines étudiés, en utilisant la plate-forme ISTEX ainsi que des bases de connaissances de ces deux sous-domaines soit à l'aide de bases existantes, soit grâce à la collaboration avec nos deux experts médecins ;
- Évaluation des approches non supervisées déjà proposées, sur ces nouveaux sous-domaines ;
- Mise en place de nouvelles approches d'extraction de relations, à l'aide d'apprentissage distant avec supervision, en adaptant des modèles à base d'apprentissage profond et de plongements de mots contextualisés ;
- Évaluation de ces nouvelles approches sur nos deux sous-domaines cibles, notamment sur les aspects de transfert de connaissances qui pourront être réalisés entre les deux sous-domaines ;
- Mise à disposition des relations extraites sous la forme de graphes de connaissances.

Candidature à la thèse

Compétences souhaitées du candidat

- Master (ou équivalent) en informatique avec une spécialisation en traitement automatique des langues et/ou en apprentissage automatique ;
- Compétences théoriques et/ou expérience dans au moins l'un des domaines suivants : apprentissage automatique, traitement automatique des langues ;
- Bonne niveau en programmation Python.

Modalités de candidature

Les candidatures doivent être envoyées par email à solen.quiniou@ls2n.fr et samuel.chaffron@ls2n.fr et doivent inclure : un CV détaillé, une lettre de motivation, vos notes de master 1 et master 2 ainsi qu'une ou deux lettres de recommandation (ou les adresses mails de personnes qui peuvent être contactées pour référence).
Les candidatures peuvent être envoyées jusqu'au **26/04/2021** et seront préférablement envoyées avant le **12/04/2021**.

Références

[BB07] Nguyen Bach and Sameer Badaskar, *A review of relation extraction*, 2007.

[BZKA20] Nada Boudjellal, Huaping Zhang, Asif Khan, and Arshad Ahmad, *Biomedical relation extraction using distant supervision*, Scientific Programming **2020** (2020), 1–9.

[CRPvM10] S. Chaffron, H. Rehrauer, J. Pernthaler, and C. von Mering, *A global network of coexisting microbes from environmental and whome-genome sequence data*, Genome Research **20** (2010), 947–959.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *BERT : Pre-training of deep bidirectional transformers for language understanding*, Proceedings of NAACL, 2019.

[FSI+12] K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower, *Microbial co-occurrence relationships in the human microbiome*, PLoS Computational Biology **8** (2012), no. 7.

[GCB+16] L. Guidi, S. Chaffron, L. Bittner, D. Eveillard, A. Larhlimi, S. Roux, Y. Darzi, S. Audic, L. Berline, J. R. Brum, L. P. Coelho, J. C. I. Espinoza, S. Malviya, S. Sunagawa, C. Dimier, S. Kandels-Lewis, M. Picheral, J. Poulain, S. Searson, Tara Oceans Consortium Coordinators, L. Stemmann, F. Not, P. Hingamp, S. Speich, M. Follows, L. Karp-Boss, E. Boss, H. Ogata, S. Pesant, J. Weissenbach, P. Wincker, S. G. Acinas, P. Bork,

C. de Vargas, D. Iudicone, M. B. Sullivan, J. Raes, E Karenti, C. Bowler, and G. Gorsky, *Plankton networks driving carbon export in the oligotrophic ocean*, Nature **532** (2016), 465–470.

[LMFH+15] G. Lima-Mendez, K. Faust, N. Henry, J. Decelle, S. Colin, F. Carcillo, S. Chaffron, J. C. Ignacio-Espinosa, S. Roux, F. Vincent, L. Bittner, Y. Darzi, J. Wang, S. Audic, L. Berline, G. Bontempi, A. M. Cabello, L. Coppola, F. M. Cornejo-Castillo, F. d'Ovidio, L. De Meester, I. Ferrera, M.-J. Garet-Delmas, L. Guidi, E. Lara, S. Pesant, M. Royo-Llonch, G. Salazar, P. Sánchez, M. Sebastian, C. Souffreau, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, G. Gorsky, F. Not, H. Ogata, S. Speich, L. Stemmann, J. Weissenbach, P. Wincker, S. G. Acinas, S. Sunagawa, P. Bork, M. B. Sullivan, E. Karsenti, C. Bowler, C. de Vargas, and J. Raes, *Determinants of community structure in the global plankton interactome*, Science **348** (2015).

[LYK+19] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, *BioBERT : a pre-trained biomedical language representation model for biomedical text mining*, Bioinformatics **36** (2019), no. 4, 1234–1240.

[SCM18] Alisa Smirnova and Philippe Cudré-Mauroux, *Relation extraction using distant supervision : A survey*, ACM Computing Surveys **51** (2018), no. 5, 1–35.