PhD Thesis Proposal

Informations About the PhD Thesis

Title	**Mining Host-Microbiome Interactions using Natural Language Processing**
Keywords	Relation extraction, biomedical scientific papers, human microbiome, natural language processing, contextualized word embeddings
Supervision	Samuel Chaffron (team ComBi, LS2N), Solen Quiniou (team TALN, LS2N)
Contact	samuel.chaffron@ls2n.fr, solen.quiniou@ls2n.fr
Partners	Caroline Trang-Poisson (CHU de Nantes), Antoine Roquilly (CHU de Nantes)

Abstract

The proposed PhD thesis concerns the extraction of relations between human microbiota species, and their relationships with genes, biomolecules, and diseases, from scientific papers, using natural language processing techniques. We will focus on papers dealing with the gut microbiota and the lung microbiota and we will validate the results obtained through a close collaboration with researchers from the University Hospital of Nantes working on these two microbiota and their role in inflammatory diseases.

Context of the PhD Thesis

Problematics

The human gut microbiota is composed of thousands of commensal bacteria essential to our health and well-being. The progressive acquisition of these communities of bacteria (in interaction) in the newborn and the child influences the development of the immune system and the susceptibility to certain diseases. Indeed, these commensal communities have a fundamental impact on host physiology, mainly by helping digestion and efficient absorption of nutrients, by shaping the immune system, but also by protecting us against the invasion of pathogens.

Today, microbial association networks inferred from environmental sampling (or metagenomic) data have revealed potential interactions between species and between broader taxonomic groups [CRPvM10, FSI+12]. However, the existence and nature of these interactions are difficult to validate and characterize [LMFH+15, GCB+16]. There is thus a real need to develop efficient methods to search the scientific literature (more than 20 million full-text papers) in order to extract known interactions between species of the human microbiota (such as mutualism, commensalism, parasitism or even predation), using natural language processing approaches. A resource listing known and annotated interactions between microorganisms, but also with the host, its genes and known diseases or molecules, would be very useful to validate predicted associations using large-scale sequencing data from clinical studies of the human microbiome (e.g. Human microbiome project, French gut project).

Research Teams Involved

The supervision of the thesis will bring together two teams of the LS2N: the TALN team, which works on natural language processing, in particular on written documents such as scientific papers, and the ComBi team, which focuses on developing graph-based methods and algorithms for analysing and modeling microbial communities and their interaction. Moreover, this subject will be developed in strong interaction with 2 research teams of the CHU of Nantes: the team of Dr. Caroline Trang-Poisson, which conducts research on clinical inflammatory bowel diseases and gastroenterology, within the Institute of Digestive Tract Diseases (IMAD), and the team of Prof. Antoine Roquilly (EA3826), which is interested in clinical and experimental therapies of pulmonary infections at the CHU of Nantes. They will allow us to validate the proposed approaches as well as the obtained results.

PhD Thesis Subject

Scientific Context

Relation extraction [BB07] is a subtask of information extraction in which semantic relationships are extracted from unstructured texts and classified into different categories. Different machine learning approaches can be used: supervised, semi-supervised or unsupervised approaches. On the one hand, the first two types of approaches obtain good results but require annotated data, in more or less important quantities. On the other hand, unsupervised approaches do not use labeled data but the extracted relations can be very large and only a small part of them correspond to interesting relations that must be validated either by experts or by using knowledge bases or ontologies, which can be difficult to achieve.

In 2018, we have carried out a first work to build a corpus from scientific articles coming from the ISTEX platform (this platform gives access to millions of scientific papers, with their meta-data and textual content, and also allows to select the set of papers corresponding to a query) and biological databases built manually, dealing with relationships between marine micro-organisms. This corpus has allowed the development of a first workflow from the labeling of the organisms present in the corpus to the extraction of relationships between organisms, using an approach based on sequential pattern extraction (unsupervised learning).

In recent years, new deep learning models, based on contextualized lexical embeddings, have been proposed, such as the BERT [DCLT19] model or the BioBERT [LYK+19] model, which is an adaptation of the BERT model to biomedical data. These new models obtain very good results for many natural language processing tasks and especially for information extraction tasks. In addition, a new paradigm has recently been proposed for automatic relation extraction: distant supervision [SCM18, BZKA20]. This approach combines the advantages of semi-supervised and unsupervised approaches, providing an alternative method for automatically generating labeled data using a knowledge base. In 2020, we carried out a second work, by reproducing results obtained with the BERT model, on state-of-the-art corpora for relationship extraction. Today, we want to continue to develop this type of model and also to evaluate it on interaction data between micro-organisms associated with living organisms (in particular Human).

Problematics

In the HOLONP project, we will focus on human microbiota, and in particular on gut and lung microbiota. One of the first objectives of the project will be to identify the available data on these two sub-domains. In a first step, we will be able to reuse the unsupervised approach that we have implemented, to apply it to these new data.

The second objective will be to propose new approaches, taking advantage of recent deep learning models, especially those adapted to the biomedical domain such as BioBERT, but also distant supervision models to be able to adapt these models, using existing knowledge bases or thanks to our clinical experts. As the gut microbiota sub-domain has a larger literature than the pulmonary microbiota sub-domain, we will be able to first focus on the latter to propose adapted relationship extraction models. We will then be able to use these new models to adapt them to the sub-domain of the pulmonary microbiota, which is less characterized and for which there is therefore less data available in the literature.

A last objective will be to make available the extracted relations, in the form of knowledge graphs, which will be easily accessible and questionable through a graphical web interface, especially for the researchers of the Nantes University Hospital.

Work Plan

The work plan for the thesis is as follows:

- Study of the state of the art on relation extraction, especially in the biomedical domain and with deep learning approaches based on contextualized word embeddings as well as distant learning approaches with supervision;
- Construction of a corpus of scientific papers of the two fields studied, using the ISTEX platform as well as knowledge bases of these two fields either with the help of existing bases, or thanks to the collaboration with our two medical experts;

- Evaluation of unsupervised approaches already proposed, on these new sub-domains;
- Implementation of new approaches to relation extraction, using supervision distant learning, by adapting models based on deep learning and contextualized word embeddings;
- Evaluation of these new approaches on our two target sub-domains, especially on the aspects of knowledge transfer that can be achieved between the two sub-domains;
- Display of the extracted relations in the form of knowledge graphs.

Application to the PhD Thesis

Profile of the Candidate

- Master (or equivalent) in Computer Science with a Natural Language Processing and/or a Machine Learning specialty;
- Knowledge or experience in one of the following fields: Machine Learning, Natural Language Processing;
- Good skills in Python.

Application modalities

Applications should be sent by email to solen.quiniou@ls2n.fr and samuel.chaffron@ls2n.fr and must include: a detailed CV, a cover letter, your grades in master 1 and master 2 as well as one or two recommendation letters (or email addresses of referees).
Applications must be sent before the **26/04/2021** and must preferably sent before the **12/04/2021**.

References

[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.