COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Oumaima EL KHETTARI

Mining Host-Microbiome Interactions using Natural Language Processing

Thèse présentée et soutenue à Nantes, le 24 février 2025
Unité de recherche : UMR6004 – Laboratoire des Sciences et du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Claire NÉDELLEC — Directrice de recherche, Université Paris-Saclay
Natalia GRABAR — Chargée de recherche, Université de Lille

Composition du Jury :

Président :	Prénom NOM	Fonction et établissement d'exercice *(à préciser après la soutenance)*
Examinateurs :	Claire NÉDELLEC	Directrice de recherche, INRAe et Université Paris-Saclay
	Natalia GRABAR	Chargée de recherche, CNRS et Université de Lille
	Pascale SEBILLOT	Professeure des Universités, INSA Rennes
	Jose MORENO	Maître de conférences, Université Paul Sabatier
	Emmanuel MORIN	Professeur des Universités, Nantes Université
Dir. de thèse :	Damien EVEILLARD	Professeur des Universités, Nantes Université
Enc. de thèse :	Solen QUINIOU	Maîtresse de conférences, Nantes Université
	Samuel CHAFFRON	Chargé de recherche, CNRS et Nantes Université

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

Titre : Fouille de texte pour l'extraction des interactions hôte-microbiome

Mot clés : Extraction de relations biomédicales, Construction de corpus, Interactions hôte-microbiome, Extraction d'informations en contexte de faibles ressources

Résumé : La compréhension des interactions entre l'hôte et le microbiome est cruciale pour la recherche biomédicale. Cette thèse explore l'extraction de relations biomédicales, en se concentrant sur les interactions entre espèces et maladies dans le microbiome et sur les déséquilibres de processus dans la littérature liée à la COVID-19, en appliquant des techniques de TAL aux domaines à faibles ressources. Une première contribution est un corpus de phrases annotées pour les interactions binaires entre espèces. Ensuite, le corpus MicrobioRel, dédié à la modélisation des interactions du microbiome, a été créé par annotation itérative et sélection d'articles via les termes MeSH. Cette thèse évalue des approches discriminatives et génératives pour l'extraction de relations, montrant l'efficacité des modèles encodeurs tout en explorant le potentiel des méthodes basées sur le résumé avec des grands modèles de langues. Des expériences sur la COVID-19 avec le corpus HOIP soulignent aussi le potentiel de l'alignement avec des ontologies pour détecter des processus implicites. Enfin, des analyses graphiques mettent en lumière des entités clés et leurs connexions, offrant de nouvelles perspectives sur les interactions du microbiome.

Title: Mining Host-Microbiome Interactions using Natural Language Processing

Keywords: Biomedical Relation Extraction, Corpus Construction, Host-Microbiome Interactions, Low Resource Information Extraction

Abstract: Understanding host-microbiome interactions is critical to advancing biomedical research. This thesis investigates biomedical relation extraction, focusing on species and disease interactions in the microbiome and process imbalance detection in COVID-19 literature, using NLP techniques to address key challenges in low-resource settings. The first contribution is a manually annotated sentence-level corpus for binary species interactions. A key outcome is the development of the MicrobioRel corpus, a dataset for modeling microbiome interactions, created through iterative annotation refinement and article selection leveraging MeSH terms. This thesis evaluates both classification and generative approaches for relation extraction, demonstrating the effectiveness of encoder-only models while exploring the potential of summarization-based methods with instruction-tuned LLMs for practical low-resource solutions. In addition, experiments in the COVID-19 domain using the HOIP dataset highlight the potential of aligning LLM outputs with ontology-based matching to detect implicitly mentioned processes, albeit with increased noise. Graph-based analyses further reveal key entities and their connectivity, offering new perspectives on microbiome interactions.