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 du numérique de l'Information et de la Communication
Spécialité : *Informatique*

Par
Adrien BAZOGE

TALMed : Traitement Automatique de la Langue Médicale

En vue de la soutenance de Thèse à Nantes, le 16 janvier 2024
Unité de recherche : UMR6004 – Laboratoire des Sciences et du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Laure SOULIER Maitresse de conférences, Sorbonne Université
Didier SCHWAB Professeur des Universités, Université Grenobles Alpes

Composition du Jury :

Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutée sur la couverture de thèse

Président :	Prénom NOM	Fonction et établissement d'exercice *(à préciser après la soutenance)*
Examinateurs :	Laure SOULIER	Maitresse de conférences, Sorbonne Université
	Didier SCHWAB	Professeur des Universités, Université Grenobles Alpes
	Gayo DIALLO	Professeur des Universités, Université de Bordeaux
Dir. de thèse :	Emmanuel MORIN	Professeur des Universités, Nantes Université
Co-dir. de thèse :	Béatrice DAILLE	Professeur des Universités, Nantes Université
	Pierre-Antoine GOURRAUD	Professeur des Universités-Praticien Hospitalier, Nantes Université

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

Titre : TALMed : Traitement Automatique de la Langue Médicale

Mot clés : TAL clinique, modèle de langue pré-entrainés, entrepôts de données de santé

Résumé : La collecte massive de données de santé a permis l'émergence d'usages secondaires, notamment la recherche et l'évaluation de la qualité des soins. Pour une utilisation optimale, ces données doivent être harmonisées et stockées dans des entrepôts de données de santé (EDS), souvent sous forme textuelle. Le traitement automatique des langues (TAL) est alors nécessaire pour en extraire des informations à grande échelle. Les méthodes actuelles de TAL s'appuient principalement sur des modèles de langue basés sur l'architecture Transformer, qui nécessitent d'être adaptés au domaine médical pour tirer profit des performances. Cette thèse explore deux thématiques : l'adaptation de ces modèles au contexte médical français et leur application en recherche clinique. Premièrement, nous menons plusieurs études d'adaptation au domaine médical de différents modèles pré-entraînés existants. Ces études ont pour but d'évaluer l'impact de différents paramètres pour l'adaptation des modèles, comme la nature des données ou la stratégie de pré-entraînement. Enfin, l'utilisation de ces modèles est étudiée dans deux projets de recherche clinique. Le projet GAVROCHE examine la relation entre la variabilité glycémique et la mortalité chez les patients atteints d'insuffisance cardiaque aiguë. Le second projet vise à extraire des déterminants sociaux de santé à partir des comptes rendus cliniques. Ces cas montrent le potentiel du TAL pour extraire des informations cliniques cruciales.

Title: Medical Natural Language Processing

Keywords: clinical NLP, pretrained language models, clinical data warehouse

Abstract: The massive collection of health data has allowed the emergence of secondary uses, including research and the evaluation of the quality of care. For optimal use, these data need to be harmonized and stored in health data warehouses (HDWs), often in textual form. Natural Language Processing (NLP) is then required to extract information on a large scale. Current NLP methods mainly rely on language models based on the Transformer architecture, which need to be adapted to the medical field to benefit from their performance. This thesis explores two themes: the adaptation of these models to the French medical context and their application in clinical research. First, we conduct several studies on the adaptation of various pre-existing pretrained models to the medical field. The aim of these studies is to evaluate the impact of different parameters for model adaptation, such as the nature of the data or the pre-training strategy. Finally, the use of these models is studied in two clinical research projects. The GAVROCHE project examines the relationship between glycemic variability and mortality in patients with acute heart failure. The second project aims to extract social health determinants from clinical reports. These cases demonstrate the potential of NLP to extract crucial clinical information.