

Détection de maladies rares par analyse et regroupement sémantique de descriptions cliniques via réseaux de neurones artificiels

Inria / HCL

Présentation :

Une maladie est considérée comme rare lorsqu'elle touche moins d'une personne sur 2.000. À ce jour, il y a entre 7.000 et 10.000 maladies identifiées comme rare dans le monde. En France, ces maladies rares concernent près 3 millions de personnes, soit 4.5% de la population, ce qui représente un enjeu majeur de santé publique [1]. La principale problématique de ces maladies est qu'elles sont souvent liées à errance diagnostique, pouvant aller jusqu'à plus de 5 ans d'attentes, pour un quart des malades, avant d'avoir un diagnostic confirmé [2]. Ces errances diagnostiques pouvant entraîner des complications plus ou moins grave en fonction des maladies, puisque nous traité convenablement, il est donc essentiel d'essayer de les détecter au plus tôt. En d'autres termes, à partir des descriptions cliniques pouvant être faites par des médecins généraliste.

Les récentes avancées en traitement automatique du langage (TAL), basées sur des réseaux de neurones artificiels de type *Transformers* [3], offre des perspectives intéressantes en termes d'analyse de textes. Cependant, de par leur nature « rare », peu de données sont disponible pour chaque maladie. Un travail de classification classique est donc compliqué à mettre en place. Une autre approche possible est celle de la détection d'anomalie [4], afin de repérer des descriptions cliniques « atypiques » caractéristiques des maladies rares.

L'objectif de ce stage est donc d'explorer la faisabilité de cette approche par détection d'anomalie pour la détection prévention de maladie à partir de descriptions cliniques. Cette exploration se concentrera sur les maladies rénales rares, pour lesquelles plus de données cliniques sont disponibles, et se déroulera en plusieurs étapes. Dans un premier temps, les descriptions cliniques de maladie rares issues de la classification Orphanet [5] seront analysés sémantiquement à l'aide de réseaux de neurones de type *Transformer* entraînés sur du français ou des textes médicaux, tels que CamemBERT [6] ou DrBERT [7], et regroupés via des algorithmes de *clustering* [8, 9]. Dans un second temps, ces regroupements seront comparés avec les arborescences proposées par Orphanet et par le registre européen des maladies rénales rares (ERKReg) [10], afin de discuter des points de convergences et de divergences de ces différents regroupements. Dans un troisième temps, des descriptions cliniques de maladies non-rares seront aussi analysées pour mettre en place une méthode permettant la détection de descriptions cliniques « atypiques » via analyse sémantique. Enfin, dans un quatrième temps, des descriptions cliniques réelles seront analysées via la méthode proposée afin d'en évaluer les performances.

Pour le développement d'une solution la plus adaptée à la pratique, il est indispensable de travailler en étroite collaboration avec des médecins spécialisés tout au long de la construction de l'outil [11]. Une emphase sera aussi mise sur l'explicabilité des modèles proposés et de leurs résultats [12].

Volet technique :

- Entretiens avec des médecins de centres maladies rares
- Proposition d'une méthode basée sur des réseaux de neurones pour l'analyse sémantique
- Évaluation de la méthode sur des données synthétiques
- Évaluation des performances de la méthode sur des données cliniques réelles

Localisation :

- Inria, site de l'Hopital Neuro, 56 Bld Pinel, Lyon.

Encadrants :

- Thomas GUYET - AIstroSight (INRIA)
- Antoine Richard - CICLY (HCL)

Références :

- [1] Ministère du Travail, de la Santé, des Solidarités et des Familles (Déc. 2022) – « Rapport d'activité 2021 filières de santé Maladies Rares » - https://sante.gouv.fr/IMG/pdf/rapport_d_activite_des_filiere_de_sante_maladies_rares_2021.pdf
- [2] Alliance Maladies Rares – « Rapport de l'enquête Erradiag » - <https://alliance-maladies-rares.org/erradiag/>
- [3] Anthony Gillioz *et al.* (2020) - « Overview of the Transformer-based Models for NLP Tasks. » - <https://doi.org/10.15439/2020F20>
- [4] Varun Chandola *et al.* (2009) – « Anomaly Detection: A Survey » - <https://doi.org/10.1145/1541880.1541882>
- [5] Orphanet - <https://www.orpha.net/fr>
- [6] Louis Martin *et al.* (2020) - « CamemBERT: a Tasty French Language Model » <https://doi.org/10.18653/v1/2020.acl-main.645>
- [7] Yanis Labrak *et al.* (2023) - « DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains » <https://doi.org/10.48550/arXiv.2304.00958>
- [8] Rajvardhan Patil *et al.* (2023) - « A Survey of Text Representation and Embedding Techniques in NLP » - <https://doi.org/10.1109/ACCESS.2023.3266377>
- [9] Alina Petukhova *et al.* (2025) - « Text clustering with large language model embeddings » - <https://doi.org/10.1016/j.ijcce.2024.11.004>
- [10] Giulia Bassanese *et al.* (2021) - « The European Rare Kidney Disease Registry (ERKReg): objectives, design and initial results » - <https://doi.org/10.1186/s13023-021-01872-8>
- [11] Katharine E. Henry (2022), « Human-machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system » - <https://doi.org/10.1038/s41746-022-00597-7>
- [12] Alejandro Barredo Arrieta *et al.* (2020), « Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI » - <https://doi.org/10.1016/j.inffus.2019.12.012>