Offre de stage de fin d'étude (Bac+5), deep learning/TAL

>LIEU DU STAGE

INSERM (Institut National de la Santé et de la Recherche Médicale)

CépiDc (Centre d'épidémiologie sur les causes médicales de décès)

46, rue Albert 75013 Paris/ Hôpital Paul-Brousse. 12, avenue Paul Vaillant-Couturier. 94804 Villejuif.

>INTITULE DU STAGE

Apprentissage profond pour le traitement automatique des textes mentionnés sur les certificats de décès : application à la prédiction de codes de la classification internationale des maladies et de textes facilement interprétables

>DOMAINE(S) COUVERT(S) PAR LE STAGE

Datascience, NLP/TAL, deep learning, LLM

Contexte La classification automatique des documents médicaux est un domaine scientifique qui connaît un intérêt constant depuis de nombreuses années. Le CépiDc de l'Inserm a la charge de traiter la partie médicale des certificats de décès pour leur enregistrement, selon les recommandations de l'Organisation mondiale de la santé (OMS), dans la classification internationale des maladies (CIM). Capitalisant sur les millions d'observations annotées par des experts au fil des années suivant de standards internationaux, le CépiDc investit dans le développement de méthodes de traitement automatique des langues pour automatiser l'enregistrement des causes de décès [1]. Un système expert basé sur le reconnaissance de mots-clés d'un dictionnaire et des règles de décision permet de traiter automatiquement les deux tiers des données. En outre, les données des années récentes sont en partie prédites par des méthodes d'apprentissage profond entraînées sur les données passées [2,3,4]. Précisément, des modèles seq-to-seq transformers [4,5] et des modèles long-term short- term memory bidirectionnels (BiLSTM, [6]) ont pour tâche de prédire les codes de la CIM version 10 correspondant aux textes rédigés par les médecins constatant les décès et d'identifier la cause initiale du décès, codée dans cette même nomenclature. Aujourd'hui, l'OMS promeut une nouvelle classification, la CIM 11, qui rendra caduque les travaux d'apprentissage précédents, à moins que ce soit des textes facilement interprétables qui soient prédits plutôt que les seuls codes de la nomenclature. L'objectif est de s'appuyer sur les avancées récentes du CépiDc et les données déjà présentes pour tester l'apport des architectures à l'état de l'art des *large language models* sur cette problématique [7,8].

Objectifs Développer et tester l'apport d'architecture LLM permettant de prédire à la fois textes et codes des méthodes de traitement automatique des textes des certificats de décès permettant d'automatiser leur enregistrement par le CépiDc. Plus spécifiquement, à partir d'architectures, définies comme baseline, de réseaux de neurones ayant montré leur efficacité pour la classification multilabel de textes, étudier l'apport de nouvelles architectures permettant :
• De normaliser les textes et aider à leur codage dans la classification internationale des maladies en prévision du changement de nomenclature (CIM11)
• De développer des méthodes d'évaluation permettant de discriminer les situations où le traitement peut être complètement automatisé et d'évaluer la performance.

Bibliographie

[1] Aurélie Névéol, et al. CLEF eHealth 2018 Multilingual Information Extraction task overview: ICD10 coding of death certificates in French, Hungarian and Italianax http://ceur-ws.org/Vol-2125/invited_paper_18.pdf

[2] Zambetta E, Razakamanana D, Robert A, Clanché F, Martin D, Hebbache Z, et al. Codage des causes de décès 2018 2019 en CIM10 - Approche combinant deep learning, système expert et codage manuel ciblé. *Mimeo*. 2023.

[3] Clanché, Razakamanana, Coudin, Robert, "Les statistiques provisoires sur les causes de décès en 2018 et 2019, une nouvelle méthode de codage faisant appel à l'intelligence artificielle", Drees Méthode n°8

[4] Falissard, Louis, Morgand, Claire,Ghosn, Walid, Imbaud, Claire, Bounebache, Karim and Rey, Grégoire. (2020). Neural translation and automated recognition of ICD-10 medical entities from natural language: Algorithm Development and Validation (Preprint). JMIR Medical Informatics. https://pubmed.ncbi.nlm.nih.gov/35404262/

[5] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz and Polosukhin, Illia. "Attention is all you need." Paper presented at the meeting of the Advances in Neural Information Processing Systems, 2017. https://arxiv.org/abs/1706.03762?context=cs

[6] Graves, A. and Schmidhuber, J., "Framewise phoneme classification with bidirectional LSTM networks," Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., Montreal, QC, Canada, 2005, pp. 2047-2052 vol. 4, doi: 10.1109/IJCNN.2005.1556215.

[7] Raffel, C., Shazeer, N.M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P.J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv, abs/1910.10683

[8] Zhao, W. X. et al: A Survey of Large Language Models, 2023, 2303.18223, arXiv. https://arxiv.org/abs/2303.18223

Le cas échéant, degré prévisible de confidentialité du rapport de stage : faible

>CONNAISSANCES ET APTITUDES RECHERCHEES

Connaissances des outils suivants :
- Apprentissage statistique et applications, apprentissage profond
- Méthodes de traitement automatique des langues,

Aptitudes :
- Logiciels : Python, openCV, Tensorflow,
- Aisance en programmation
- Manipulation de bases de données volumineuses
- Traitement sur données médicales confidentielles
- Anglais lu et écrit courant

>ENVIRONNEMENT DE LA MISSION

Intitulé, activité, compétences statistiques de l'unité d'accueil et du maître de stage :
CépiDc (Centre d'épidémiologie sur les causes médicales de décès). Les missions du CépiDc sont :
- la production de la base nationale des causes médicales de décès,
- la diffusion de cette base pour des objectifs de recherche et de santé publique,
- la production d'analyses statistiques et de recherche sur cette base de données.

Cette dernière mission a donné lieu à l'application des méthodologies statistiques adaptées pour de nombreuses publications dans des revues scientifiques internationales.

Le stage sera co-encadré par :
- Remi Flicoteaux, médecin DIM à l'AP-HP et directeur médical du CépiDc spécialisé en méthode de traitement automatique des langues et machine learning,
- Aude Robert, ingénieur au CépiDc spécialisé en traitement automatique des langues.

Seront aussi associés Daniel Razakamanana datascientist, Elisa Zambetta data engineer et Elise Coudin directrice du CépiDc.

Ressources mises à la disposition du stagiaire :
Données nationales d'enregistrement des causes de décès (plus de 3 millions d'enregistrements annotés)
Plateforme de calcul du CépiDc (sur base de 3 GPU).
Interactions quotidiennes avec l'équipe automatisation /datascience du CépiDc.

Gratification : environ 500€ / mois
Durée du stage : 6 mois (négociable), date de début négociable.

>PERSONNES A CONTACTER

Dr. Rémi Flicoteaux (remi.flicoteaux@aphp.fr)
Aude Robert (aude.robert@inserm.fr)