Offre de stage de fin d'étude (Bac+5), machine 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)

80, rue du Général Leclerc, 94270 Le Kremlin-Bicêtre

> INTITULE DU STAGE

Développement de méthodes d'apprentissage profond pour le traitement automatique des textes mentionnés sur les certificats de décès et l'application de la classification internationale des maladies

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

Data science, Mathématiques (optimisation convexe et recherche opérationnelle), Informatique

Contexte

La classification automatique des documents médicaux est un domaine scientifique qui connait un intérêt constant depuis de nombreuses années [1]. Le développement des méthodes d'apprentissage profond a permis d'améliorer sensiblement les performances des méthodes de prédiction utilisées dans ce cadre [1]. L'utilisation de méthodes de transfert d'apprentissage appliquées au traitement automatique des langues (NLP) permet d'accroitre encore la performance des modèles couramment utilisés [2]. Le CépiDc de l'Inserm est l'organisme qui en France à la charge de traiter la partie médicale des certificats de décès pour leur enregistrement, selon les recommandations de l'OMS, dans la classification internationale des maladies. Capitalisant sur les millions d'observations annotées par des experts suivant de standards internationaux qu'il détient, le CépiDc s'investi depuis de nombreuses années dans le développement de méthodes de traitement automatique des langues pour automatiser l'enregistrement des causes de décès [1]. Des expériences avec des méthodes d'apprentissage profond ont obtenus des résultats très encourageants [3]. Le CépiDc souhaite poursuivre le développement de méthodes d'apprentissage profond pour continuer d'améliorer ce traitement automatique.

Objectifs

Développer 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 [2,3], étudier l'apport de nouvelles architectures permettant :

- D'exploiter et de documenter le gain de performance des méthodes de transfert d'apprentissage
- De mettre en place une architecture de type encodeur/décodeur pour normaliser les textes et aider à leur codage dans la classification internationale des maladies
- De développer des méthodes d'évaluation permettant de discriminer les situations où le traitement peut être complètement automatisé.

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] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Oct 2018. https://arxiv.org/pdf/1810.04805.pdf
[3] Louis Falissard, Claire Morgand , Sylvie Roussel, Claire Imbaud, Walid Ghosn, Karim Bounebache and Grégoire Rey. Neural translation and automated recognition of ICD10 medical entities from natural language. 2020. arXiv cs.CL 2004.13839

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

☐ **extrême** ☐ **moyen** ☒ **faible**

Connaissances et aptitudes recherchées chez le stagiaire :

Connaissances des outils suivants :

- *Principes de l'apprentissage statistique et applications*
- *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.

Il bénéficiera de l'expertise de Louis Falissard, expert en apprentissage profond actuellement en doctorat au 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).

Gratification : environ 500€ / mois

Durée du stage : 6 mois

> PERSONNE(S) A CONTACTER

Dr. Rémi Flicoteaux (remi.flicoteaux@aphp.fr)

Aude Robert (aude.robert@inserm.fr)