

Apprendre à classer des textes hospitaliers rédigés en anglais selon la classification CIM-9 avec une approche par budget

Leonardo Moros^{*,***}, Jérôme Azé^{*}, Sandra Bringay^{*,**}, Pascal Poncelet^{*}
Maximilien Servajean^{*,**}, Caroline Dunoyer^{***,****}

^{*} LIRMM UMR 5506, Université de Montpellier, CNRS, Montpellier, France
prenom.nom@lirmm.fr

^{**} Groupe AMIS, Université Paul-Valéry, Montpellier, France
prenom.nom@univ-montp3.fr

^{***} Département d'Information Médicale, CHU Montpellier, Montpellier, France
prenom.nom@chu-montpellier.fr

^{****} IDESP, UMR UA11, INSERM - Université de Montpellier, Montpellier, France
prenom.nom@umontpellier.fr

Résumé. Le codage médical vise à annoter les comptes rendus médicaux selon les diagnostics et les traitements. Cette tâche est liée à la facturation clinique. Dernièrement, le codage automatique est devenu un domaine de recherche actif pour lequel de nombreux modèles, utilisant des architectures basées sur des réseaux neuronaux, ont été proposés. La plupart de ces approches sont validées sur le jeu de données MIMIC-III. Dans cet article, nous revenons sur la qualité de ce jeu et proposons un nouveau découpage. Puis, nous expérimentons un classificateur basé sur une approche par budget dont l'objectif est de faciliter ce codage en adaptant le nombre de codes à plusieurs combinaisons de contraintes.

1 Introduction

Les professionnels de santé documentent minutieusement chaque rencontre avec les patients dans des dossiers via des documents structurés et semi-structurés, contenant des informations sur les traitements, les procédures et les diagnostics. Afin d'obtenir des financements, les établissements de santé doivent associer aux séjours des patients des codes de facturation, issus de la Classification Internationale des Maladies (CIM). De nombreux travaux ont proposé des systèmes (semi-)automatiques pour ce codage. Des chercheurs ont notamment exploré les méthodes d'apprentissage profond. Les réseaux neuronaux convolutifs (CNN) et récurrents (RNN) avec des mécanismes d'attention (Xie et Xing, 2018; Mullenbach et al., 2018; Vu et al., 2020) correspondent à l'état de l'art actuel. Dans ce travail, nous proposons un classifieur intégrant une approche par budget inspirée par les travaux de Lorieul et al. (2021) pour palier aux limites de ces derniers. Par ailleurs, la majorité des articles évaluent les approches sur la base MIMIC-III. Or, les découpages créés par Mullenbach et al. (2018) posent des problèmes de stratification, ce qui rend difficile l'évaluation des approches. Nous évaluons notre nouvelle architecture en la comparant à LAAT (Vu et al., 2020) sur un nouveau découpage du jeu de données MIMIC-III, garantissant la stratification des labels dans les échantillons.