Journées de méthodologie statistique de l'Insee 2022

Application de techniques de *machine learning* pour coder les professions dans la nomenclature des professions et catégories socio-professionnelles 2020

Théo LEROY (*), Lucas MALHERBE (*), Tom SEIMANDI (*)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

theo.leroy@insee.fr lucas.malherbe@insee.fr tom.seimandi@insee.fr

Mots-clés. (6 maximum) : PCS, fastText, classification hiérarchique

Domaines. Codification automatique, NLP, réseaux de neurones

Résumé

La nomenclature des professions et catégories socioprofessionnelles (PCS) a subi une rénovation en 2020. Celle-ci s'accompagne de la promotion d'un outil d'autocomplétion des libellés de profession dans une liste de libellés enrichis permettant le codage direct dans un poste de la nomenclature. Jusqu'ici, la codification automatique des professions issues de l'enquête annuelle de recensement était assurée par un moteur de règles déterministes appelé Sicore. Le passage par la liste de libellés enrichis rend caduque cet environnement dès lors que l'outil d'autocomplétion est utilisé. Il n'est pas prévu de développer un moteur de règles complet similaire à Sicore pour les libellés en dehors de la liste de libellés enrichis en PCS 2020. Or, l'outil d'autocomplétion ne sera disponible que pour des collectes informatisées et par ailleurs il comprend la possibilité de répondre « hors liste ».

Pour être en mesure de coder les bulletins papier comme les réponses informatisées « hors liste », un algorithme de codification automatique en PCS 2020 de ces bulletins doit être créé. L'objectif est de maintenir le taux de codifications correctes tout comme le taux d'envoi en reprise manuelle à des niveaux similaires à l'existant.

Il s'agit de proposer, ici, plusieurs solutions à partir de modèles de classification supervisée. Des premiers travaux d'utilisation de techniques de *machine learning* sur la PCS 2003 se sont révélés suffisamment prometteurs pour poursuivre dans cette direction. Suite au report de l'enquête de recensement de 2021, une base de données a été annotée dans la nouvelle nomenclature PCS 2020 lors d'une campagne de labellisation. Cette vaste campagne d'annotation d'environ 120 000 bulletins s'est tenue en double codage avec arbitrage et offre une opportunité unique d'entraîner directement des modèles de *machine learning* à prédire dans cette nouvelle nomenclature, sans recourir à l'implémentation de règles déterministes comme le faisait Sicore.

Les meilleurs modèles identifiés lors des tests menés sur la PCS 2003 ont été répliqués sur ces données pour coder en PCS 2020. Ils font usage du classifieur *fastText*, un réseau de neurones à une couche cachée. Cependant, un travail d'adaptation et de réestimation des modèles est nécessaire pour plusieurs raisons. D'abord, les nomenclatures PCS 2003 et PCS 2020 présentent des différences non négligeables. Ensuite, les variables mobilisables pour prédire un poste de la nomenclature ne sont pas exactement les mêmes pour les deux nomenclatures.

Les premiers résultats évalués sur un échantillon de test annoté aussi et indépendant de l'échantillon d'apprentissage suggère une précision de codage de 67 % pour des libellés hors champ de l'index des métiers aux libellés enrichis utilisés par l'outil d'autocomplétion. Dans le but d'améliorer les performances des modèles, plusieurs axes sont explorés. Un premier vecteur d'amélioration est la prise en compte de la hiérarchie dans la nomenclature en agrégeant des modèles sur plusieurs niveaux de la nomenclature (combinaison linaire des probabilités obtenues à chaque niveau, arbre de modèles conditionnels). Une autre piste est d'augmenter la taille de l'échantillon d'apprentissage avec des informations provenant d'autres enquêtes. Cet article présentera le résultat de ces approches.

La nomenclature PCS 2020 sera intégrée au recensement de la population à partir de la collecte 2024.

Bibliographie

[1] Alexis Eidelman, Olivier Chardon. La rénovation de la nomenclature socioprofessionnelle (2018-2019). 2019.

[2] Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov. Bag of tricks for efficient text classification. arXiv preprint arXiv :1607.01759, 2016.

[3] Théo Leroy, Tristan Loisel. Machine Learning approaches for coding occupations into the new national occupational classification, NTTS 2021

[4] Gweon, Hyukjun, Schonlau, Matthias, Kaczmirek, Lars, Blohm, Michael and Steiner, Stefan. "Three Methods for Occupation Coding Based on Statistical Learning" Journal of Official Statistics, vol.33, no.1, 2017, pp.101-122