Application de techniques de machine learning pour coder les professions en PCS 2020

JMS 2022

Tom Seimandi, Théo Leroy, Lucas Malherbe

Recensement et codification de la PCS - I

- La nomenclature des Professions et Catégories Socioprofessionnelles (PCS) sert à la codification du recensement et des enquêtes que l'Insee réalise auprès des ménages.
- En fonction de sa situation, la personne recensée renseigne l'intitulé de sa profession parmi 3 types : la profession salariée, la profession non salariée et la profession antérieure.
- Pour la version actuelle de la nomenclature (PCS 2003), la codification automatique est assurée par l'outil Sicore.
- La nomenclature PCS a été rénovée (PCS 2020) et mettre à jour Sicore demanderait énormément de travail. Expérimentations autour de techniques de machine learning pour coder la profession : des tests sur des données annotées en PCS 2003 concluants nous ont poussé à organiser une campagne d'annotation en PCS 2020.

Recensement et codification de la PCS - II

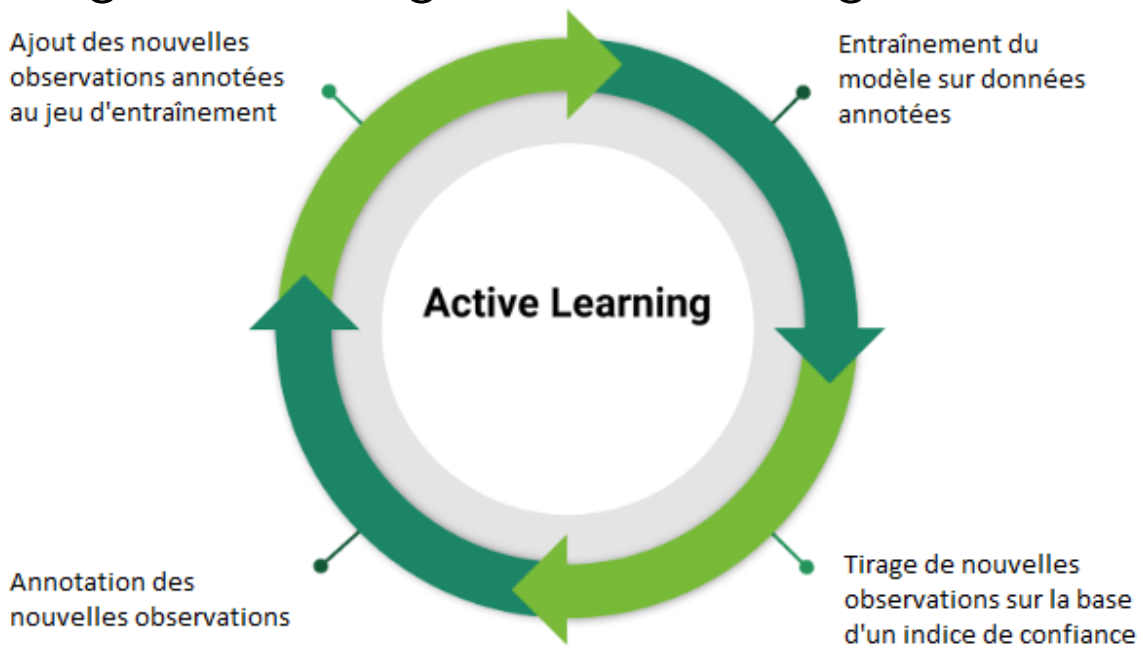
- La nomenclature PCS 2020 est hiérarchique sur 4 niveaux : on compte 6 groupes, 30 catégories socioprofessionnelles, 126 professions regroupées et 316 professions.
- Pour la profession actuelle salariée, outre le libellé de profession, les variables suivantes peuvent aider à coder la PCS :

Variable	**Nombre de modalités**
Situation principale	7
Statut professionnel	4
Position professionnelle	9
Activité de l'établissement	718
Catégorie juridique de l'établ.	177
Effectifs de l'établ.	15

- Pour la PCS 2020, on a une liste de libellés associés à un code de manière univoque. On considère ici les libellés hors liste.

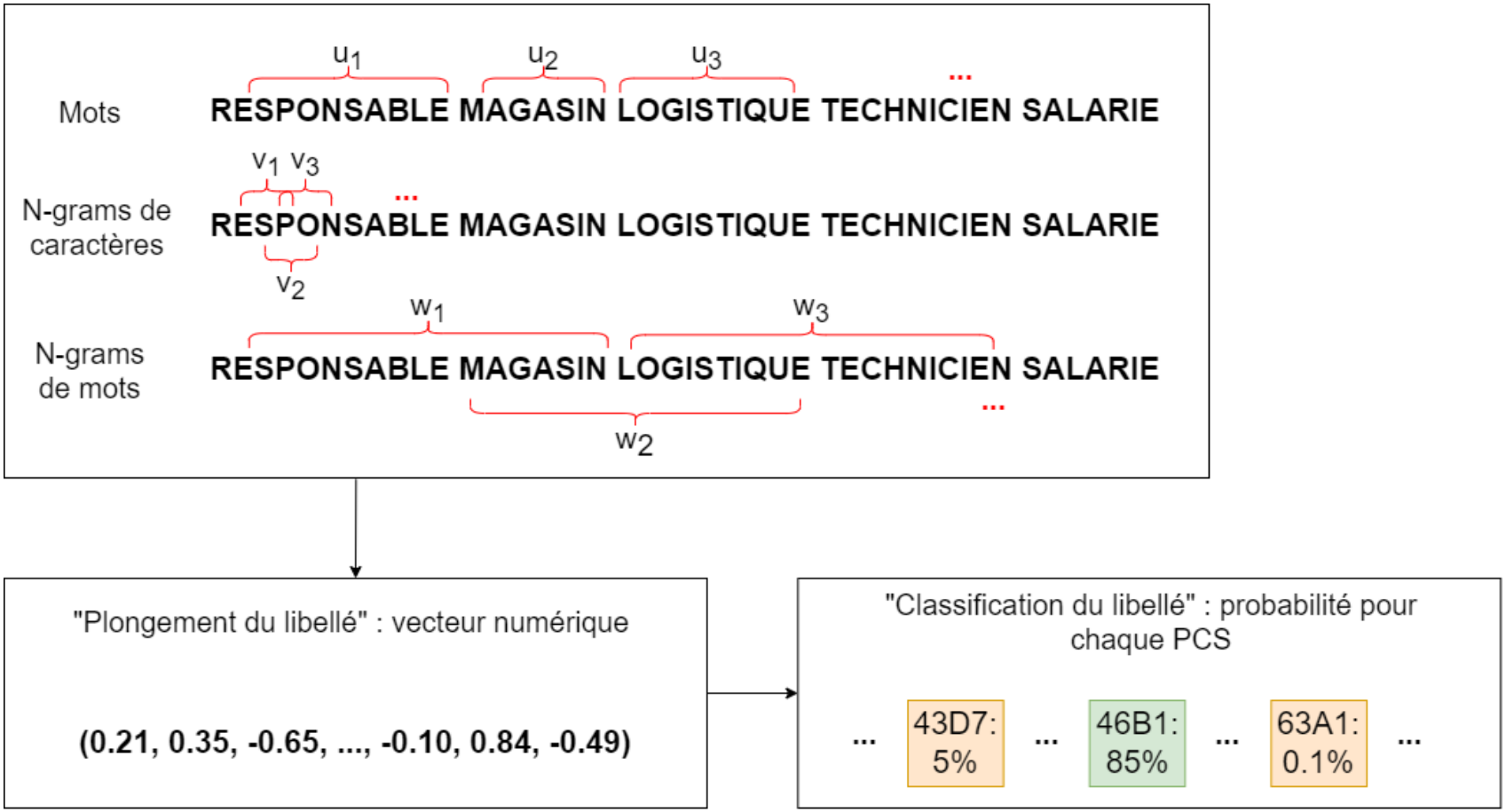
Campagne d'annotation

- Plusieurs options pour la stratégie de tirage de 120 000 bulletins du recensement :
  - Tirage aléatoire avec probabilité d'inclusion proportionnelle à la fréquence d'apparition des caractéristiques individuelles ;
  - Tirage aléatoire à partir de plongements lexicaux ;
  - Tirage avec stratégie d'active learning.



Margin sampling : $x^* = \arg\max_x(\mathbb{P}_\theta(\hat{y}_1|x) - \mathbb{P}_\theta(\hat{y}_2|x))$

Méthode - I


Mots
RESPONSABLE MAGASIN LOGISTIQUE TECHNICIEN SALARIE
N-grams de caractères
RESPONSABLE MAGASIN LOGISTIQUE TECHNICIEN SALARIE
N-grams de mots
RESPONSABLE MAGASIN LOGISTIQUE TECHNICIEN SALARIE
"Plongement du libellé" : vecteur numérique
(0.21, 0.35, -0.65, ..., -0.10, 0.84, -0.49)
"Classification du libellé" : probabilité pour chaque PCS
43D7: 5%
46B1: 85%
63A1: 0.1%

Méthode - II

- Inclusion des variables annexes, 2 possibilités :
  - Concaténation au libellé de profession. Le plongement lexical se fait sur un libellé enrichi (**A**) ;
  - Une matrice de plongement indépendante est entraînée pour chaque variable annexe. La réprésentation vectorielle d'une observation est la concaténation du plongement lexical du libellé et des plongements de chaque variable annexe (**B**).
- En aval de la phase de plongement, 2 possibilités :
  - Schéma *softmax* : si on note w_i le plongement obtenu pour l'observation i, les probabilités prédites pour chaque classe sont égales à $f(Bw_i)$, où f est la fonction softmax et B est la matrice de coefficients du classifieur linéaire (**1**) ;
  - Schéma *one-versus-all* : un classifieur binaire indépendant est entraîné pour chaque classe (**2**).

Méthode - III

- Optimisation par descente de gradient stochastique (ou extension, e.g. Adam) d'une perte *cross-entropy* :

$$L(w_1, ...w_n) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(f(Bw_i)).$$

- Pour évaluer la performance des modèles, on pondère la proportion d'observations bien classées :

$$\text{Acc} = \frac{\sum_{i=1}^{n} \frac{t_i \mathbb{1}(\hat{h}(w_i) = y_i)}{\pi_i}}{\sum_{i=1}^{n} \frac{t_i}{\pi_i}},$$

où t_i est le nombre d'occurrences de l'observation i dans la base de sondage (EAR 2020) et π_i est la probabilité d'inclusion de l'observation i dans l'échantillon de test.

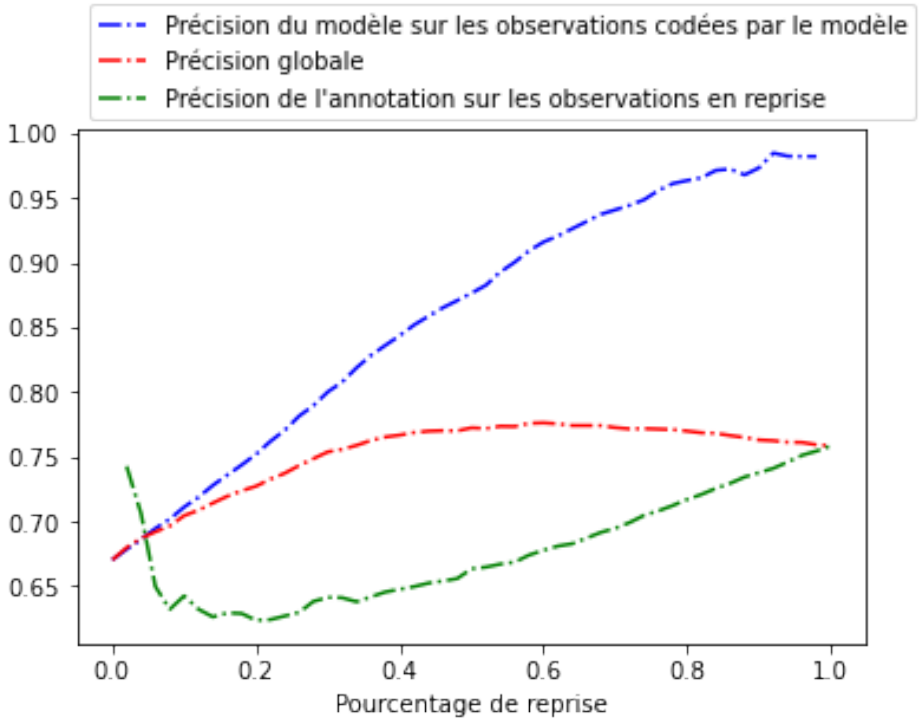
Résultats

- Proportion pondérée d'observations bien classées:
  - Modèle avec variables annexes concaténées à la chaîne de caractère du libellé de profession (**A2**) : **66.8%** ;
  - Modèle avec plongement *indépendant* de chaque variable annexe (**B1**) : **66.0%** ;
- Les performances atteignent les niveaux attendus par la maîtrise d'ouvrage du RP ($\rightarrow$ mise en production) ;
- Optimisation des hyperparamètres grâce à un jeu de validation. Par exemple pour le modèle **A2** :

Hyperparamètre	**Valeur**
Dimension de l'espace de plongement	150
Taille maximale des n-grammes de mots	3
Taille maximale des n-grammes de caractères	4
Taille minimale des n-grammes de caractères	3

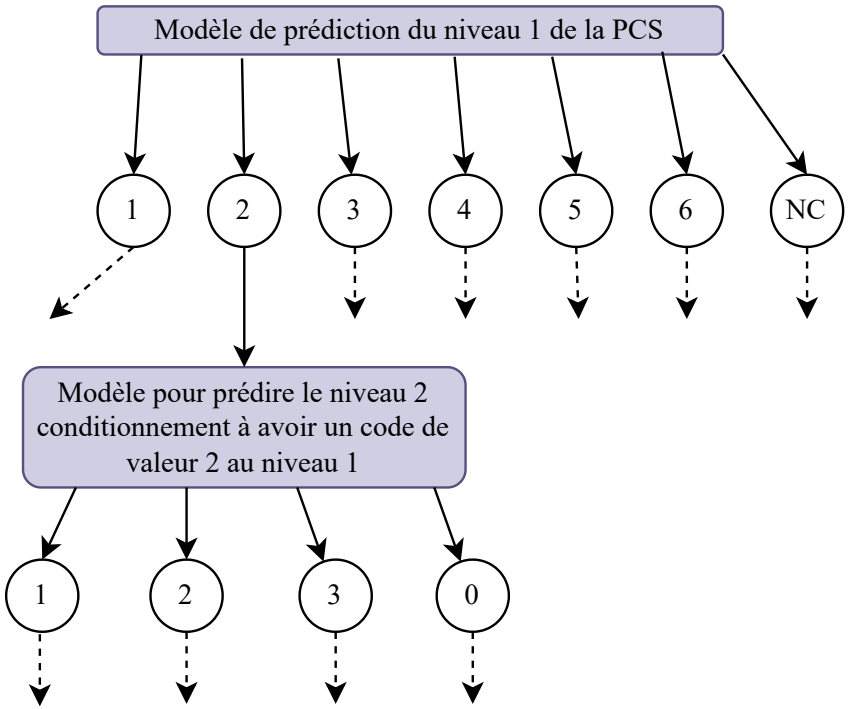
Reprise

- La reprise manuelle n'est pas obligatoire. Envoyer en reprise manuelle les libellés les plus difficiles à coder pour le modèle permet d'augmenter tout de suite la précision de la codification et d'optimiser le gain d'information lorsqu'on le ré-entraîne avec les données nouvellement annotées.



Suite

- Choix du classifieur utilisé après la phase de plongement ;
- Prise en compte de la hiérarchie de la nomenclature :



- Augmentation de la taille du jeu d'apprentissage : mobilisation d'autres sources, annotation de nouvelles données au cours des futures phases de "reprise manuelle".