Coding occupations and coding products: two NLP applications for Official Statistics

Elise Coudin
Head of SSP Lab
DMCSI, INSEE

Mesurer pour comprendre

Two **examples** of NLP/supervised ML model projects for **coding**, currently ongoing

- Why two ? Many common points: data (short descriptions, few additional features),classifiers
- Still **experimental** projects, not yet questions on integration into coding system or production
- Where we are **now**
- And what we **plan** to do next year as the current context is pushing us to accelerate…

Coding scan data products for turnover indexes

Context

- Since 2020, the French CPI has used scan data from large food retailers

- CPI strategy - rely on a purchased barcode repository (IRI worldwide which classifies barcodes into families of product). Use the link between barcode and families of products together with a correspondence table between families of products and the nomenclature dictionary.

- Other short-term economic indicators could benefit from scan data (high frequency) - especially **turnover indexes in value**

- However, using the IRI referential/family of products, we can code only 78% of the total amount of turnover (NA2008, 129 categories) → Need to enlarge the scope

- **Project**:
  - Use links between IRI repository/families of product/NACE categories to construct a labelled sample of product short descriptions (such as they appear on receipts)
  - Train and assess performance of supervised machine learning algorithms
  - Predict and assess performance for non-indexed products

Approach: (1) construct a labelled dataset, (2) pretreatments (3) use the supervised module of fastText (Joulin et al., 2017, Bojanowski et al., 2016)

- two-layer neural network that uses n-grams of characters and words as tokens and encode product description as an average of its tokens' embeddings.
- the use of n-grams makes it robust to abbreviations and typos.
- rapid to train - 20 minutes on 5 CPUs for 5 millions of products, embeddings: 100, learning rate:0.1, Nb epochs:70, loss function one-vs-all.

Results

- Trained on 80% of the sample (17 millions of products, 1h10 on 4 CPU))
- Global precision of 94% - F1 scores varying .79 - .99 cross categories (except for some rare categories<3000 items)
- Even better when weighted by turnover amounts (what we are interested in): **global precision of 97%**
- Use the difference of prediction probabilities for the top-2 most likely categories to quantify the confidence in a product classification, and use it to monitor manual work/labelling campaign

Next step

- Labelling campaign using a web application to evaluate performance out of sample -- planned in 2021Q1, and maybe enrich the train sample.

Coding occupations of paper census responses in a new nomenclature

Context

- In 2019/2020, experts built a new dictionary of the French occupation nomenclature - Professions et catégories socio-professionnelles, namely PCS 2020 dictionary (316 categories, upgrade of the PCS 2003): cleaning, new occupations, transversal groups of occupations (sustainable, digital…)

- These experts also promoted the use of auto-completed response tools - proposing a list of around 5 000 precise occupations (x sectors/ public-private/position..). Only responses not in the list would be sent to manual coding, for those in the list (index) the coding is known/ direct/ unambiguous.

- Sounds good for surveys (no paper, small samples): LFS test in 2020
- However, raises challenges for the Census (annual survey, about 12% of population)
    - -yearly 1.3 million of paper questionnaires, with only 30% occupations that belong to the list.
    - -the rule-based automatic coding system (SICORE) that used to code in PCS2003 will not be maintained/ heavy to adapt to the new dictionary.

Project:

- Experiment supervised machine learning algorithm performance
- Estimate the minimal size of a labelled dataset needed for initiating the algo
- Constraints: reach at least 80% of good predictions, propose less than 12% of the questionnaires to manual re-coding.

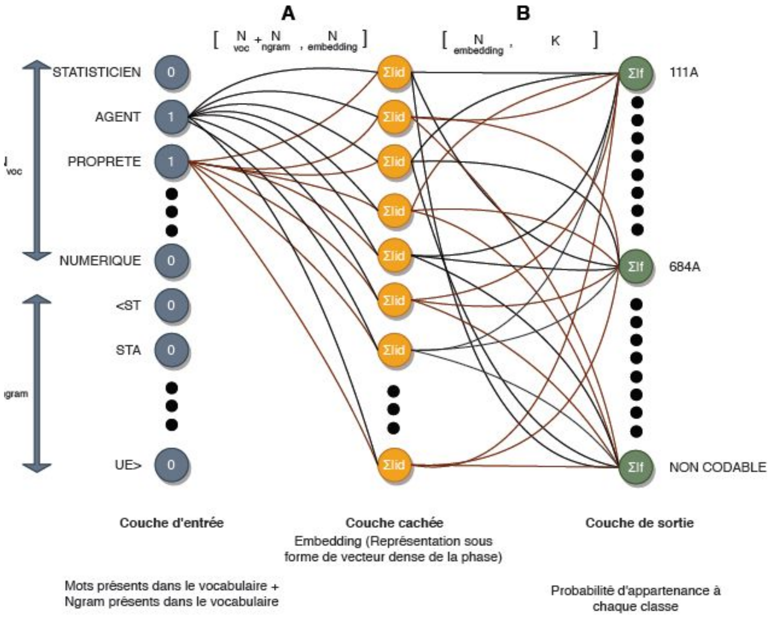
Approach:

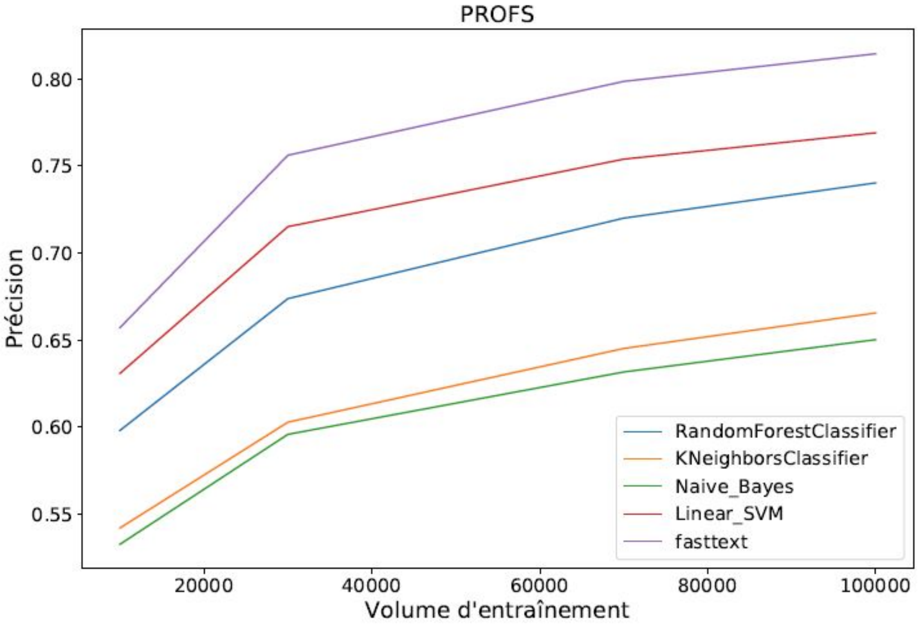
- test and simulate using the 2015-2019 data coded with the old dictionary (PCS2003) supervised ML models to determine a minimal size of a training set
- Compare models = classifiers x hyperparameters x selected features
- More precisely, 3 models to train - prediction of current occupations for employees/self-employed and previous occupation for retirees and non-employed.

Results

- No matter the classifier, the selected features are the same : occupation declared + annex variables (public/private/both, professional position, sectors, firm size) used for coding (manually or automatically)

- FastText classifier overcomes largely other classifiers (SVM with TF-IDF embeddings, RF….)
  Presence of n-grams enable automatic spelling corrections


A
[N_{voc} + N_{ngram} , $N_{embedding}$]
B
[$N_{embedding}$, K]
STATISTICIEN
AGENT
PROPRETE
NUMERIQUE
<ST
STA
UE>
N_{voc}
N_{gram}
0
1
1
0
0
0
0
Σlid
Σlf
111A
684A
NON CODABLE
Couche d'entrée
Couche cachée
Embedding (Représentation sous forme de vecteur dense de la phase)
Couche de sortie
Mots présents dans le vocabulaire + Ngram présents dans le vocabulaire
Probabilité d'appartenance à chaque classe


PROFS
Précision
Volume d'entraînement
0.55
0.60
0.65
0.70
0.75
0.80
20000
40000
60000
80000
100000
RandomForestClassifier
KNeighborsClassifier
Naive_Bayes
Linear_SVM
fasttext

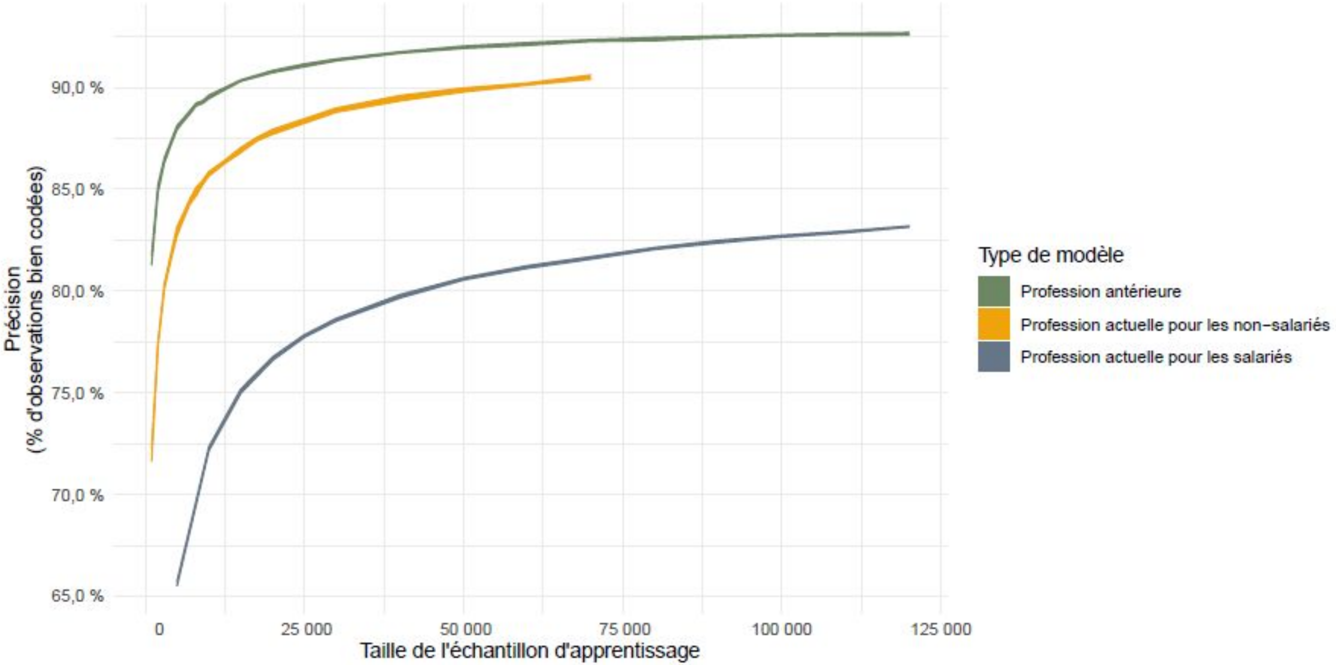
Results (..)

- No need for repeated data in the train sample but need for one occurence of the most frequent cases

- Minimal size

Labels in the Occupation list
 (index)

+



Results (..)

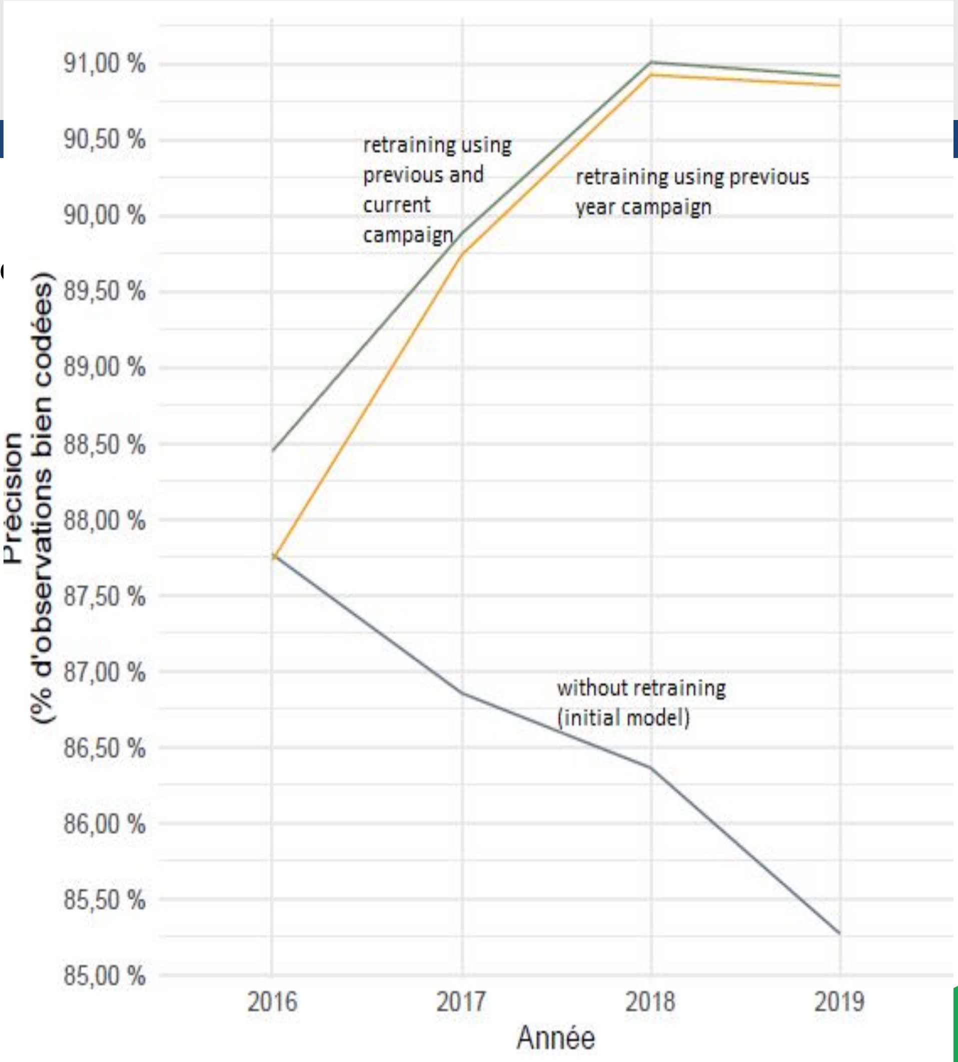
- Minimal sizes (including test samples): 100 000 (
+ 12500 (self-employed) + 7500 (previous occupation) + index -> accuracy >83%

- Accuracy will get better with future re-coding campaigns (around 300 000 Census slips coded per year)

- Use a confidence index: p1(x)-p2(x) to choose which observations it is optimal to re-code manually, and to re-train the model with.

- Importance to retrain yearly.

Then, things accelerated ….



Next steps

Due to the Covid-2019 crisis and current lockdown in France, the 2021 Census survey is postponed to 2022 → Census teams will be available during S1 2021.

Great opportunity to conduct a large one-shot labelling campaign in PCS2020

Proposition of labelling twice + trade-off 120 000 questionnaires

That is what we are working on now !

So quite interested in the ONS proposal for ML-group next year program on *Workstream 4: How to get good training data, how to keep it up to date, when to relearn a model, what does 'good' mean, how to measure that?*

Thank you ! Questions?

insee.fr

Mesurer pour comprendre