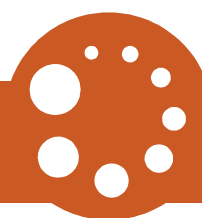




NOGAuto: Automation of the coding of the unities in the Business and Enterprise Register

Summary

EXPERIMENTAL STATISTICS



Neuchâtel, 2019

Published by: Federal Statistical Office (FSO)
Editors: Claude Macchi
Topic: 00 Statistical Basis and Overviews
Original text: French
Translation: FSO language services

Layout concept: Section DIAM
Downloads: www.statistics.ch
Copyright: FSO, Neuchâtel 2019
Reproduction with mention of source authorised
(except for commercial purposes)



NOGAuto pilot project: summary

30 June 2019

Nomenclature is the basic element of statistical production. The coding quality of the observed units has a direct impact on all data production processes, on credibility and on the quality of the statistical results. This is even more important in the context of registers and administrative data, which are the starting point for many different statistics and the basis of survey frameworks and data analyses.

To constantly improve the coding quality of units in the Business and Enterprise Register (BER) and to reduce the burden on businesses in terms of their obligation to supply information to the Federal Statistical Office (FSO), the Business Registers Data Section (URD) at the FSO launched a project – as part of the FSO's data innovation strategy programme – aimed at automating the allocation of economic activity codes to businesses.

The coding procedures are currently fairly standardised. Coders analyse and interpret information on the activities of the businesses, such as data supplied by the businesses themselves, information from surveys, and descriptions from the company register and various administrative data. On this basis, they define keywords which are compared with a list of keywords and concepts linked to each of the classification positions and to their explanatory notes, and in this way the coders decide which code to allocate to the business in question.

Using innovative methods, the URD Section plans to develop a machine learning system to automate the manual coding procedure. This system ("NOGAuto") performs the steps of reading and interpreting the descriptions of economic activities (preparation phase) and automatically links the business to a classification code (modelling phase). This involves learning based on an existing data set that has already been assigned a code manually. The scope of the system can subsequently be expanded by searching directly online for other sources of information on the business to be classified.

The results obtained in the initial phase of the project are highly promising. Although resources were very limited and the infrastructure was suboptimal, a prediction rate of 89% was achieved by the system in tests carried out on a selection of codes.

It is also important to emphasise that the effort invested in realising NOGAuto is not only limited to classification of the economic activities of businesses, but that a slight adaptation of the system would allow it to be used to assign codes to professions, diseases, causes of death, products, etc.

Complementary analysis methods used:

- for the **preparation** phase, i.e. the phase comprising reading, comprehension and text mining, we implemented natural language processing (NLP) technologies, for example based on the naïve Bayes methods for compact language detection;
- for the **modelling** phase culminating in the allocation of a NOGA code to the business to be classified, the system used various architectures of artificial neural networks and random forests.