Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Bundesamt für Statistik BFS

NOGAuto: Automatisierung der Kodierung der Einheiten im Betriebs- und Unternehmensregister

Zusammenfassung

EXPERIMENTAL STATISTICS

Neuchâtel, 2019

Herausgeber: Bundesamt für Statistik (BFS)
Redaktion: Claude Macchi
Themenbereich: 00 Statistische Grundlagen
Originaltext: Französisch
Übersetzung: Sprachdienste BFS

Layoutkonzept: Sektion DIAM
Download: www.statistik.ch
Copyright: BFS, Neuchâtel 2019
Wiedergabe unter Angabe der Quelle
für nichtkommerzielle Nutzung gestattet

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI

Bundesamt für Statistik BFS
Abteilung Register

Pilotprojekt NOGAuto: Zusammenfassung

30. Juni 2019

Nomenklaturen sind Schlüsselelemente der Statistikproduktion. Die Qualität der Kodierung der beobachteten Einheiten wirkt sich direkt auf den gesamten Datenproduktionsprozess, auf die Glaubwürdigkeit und auf die Qualität der statistischen Ergebnisse aus. Dies gilt insbesondere im Zusammenhang mit den Register- und Administrativdaten, auf denen zahlreiche Statistiken, Stichprobenrahmen und Analysen basieren.

Zur fortlaufenden Verbesserung der Kodierung der im Betriebs- und Unternehmensregister (BUR) erfassten Einheiten sowie zur Entlastung der Unternehmen, die gegenüber dem Bundesamt für Statistik (BFS) auskunftspflichtig sind, hat die Sektion Unternehmensregisterdaten (URD) des BFS im Rahmen der Dateninnovationsstrategie ein Projekt gestartet, mit dem die Zuteilung der Codes der wirtschaftlichen Tätigkeit an die Unternehmen automatisiert werden soll.

Die Kodierverfahren sind heute weitgehend standardisiert. Die für die Kodierung zuständigen Mitarbeitenden analysieren und interpretieren die Informationen über die wirtschaftliche Tätigkeit der Unternehmen. Dabei handelt es sich um Daten, die von den Unternehmen selbst vorgelegt werden, Informationen aus Erhebungen, Beschreibungen in Firmenregistern oder verschiedene Administrativdaten. Auf dieser Grundlage definieren sie Schlüsselwörter, die mit einer Liste von Schlüsselwörtern und Konzepten verglichen werden, die sich auf jede der Positionen der Klassifizierung beziehen und übernehmen den für das Unternehmen zutreffenden Code.

Die Sektion URD möchte das manuelle Kodierverfahren nun mit einem innovativen, auf maschinellem Lernen basierenden System automatisieren. Dieses System (NOGAuto) soll das Lesen und Interpretieren der wirtschaftlichen Tätigkeiten übernehmen (Vorbereitungsphase) und den Unternehmen die entsprechenden NOGA-Codes automatisch zuteilen (Modellierungsphase). Es lernt von bereits vorhandenen, manuell kodierten Daten. In einem weiteren Schritt soll das System so erweitert werden, dass es im Internet direkt nach weiteren Informationsquellen über das zu kodierende Unternehmen suchen kann.

Die Ergebnisse der ersten Projektphase sind äusserst vielversprechend. Trotz sehr begrenzter Ressourcen und einer noch verbesserungswürdigen Infrastruktur erreichte das System bei den Tests mit einer Auswahl von Codes eine Vorhersagegenauigkeit von 89%.

Wichtig ist an dieser Stelle der Hinweis, dass der Aufwand für die Umsetzung von NOGAuto nicht allein auf die Kodierung der wirtschaftlichen Tätigkeiten von Unternehmen abzielt. Mit geringfügigen Anpassungen könnte das System beispielsweise auch für die Kodierung von Berufen, Krankheiten, Todesursachen oder Produkten verwendet werden.

Angewandte komplementäre Analysemethoden:

- In der **Vorbereitungsphase**, d. h. für das Lesen, Verstehen und Suchen von Texten (Text-Mining), wurden Technologien zur Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) angewandt, die u. a. auf Naive-Bayes-Methoden zur «Compact Language Detection» basieren.
- In der **Modellierungsphase**, bei der den Unternehmen ein NOGA-Code zugeteilt wird, verwendet das System unterschiedliche Architekturen von künstlichen neuronalen Netzen (Artificial Neural Networks) und «Random Forests».