SE2-2-1

Learning to Code: Automatic Occupation and Business Coding via Machine Learning Approaches

Hsin-Min Lu[1], Yu-Hao Lee[1], Chih-Chia Huang[1]
[1] National Taiwan University

We investigate the potential of automatically assigning the occupation and business codes based on the text descriptions (in traditional Chinese) from open-ended survey questions. We formulated the research question as two separate supervised learning tasks: one for occupation coding and the other for business coding. Since the text description is usually short, we propose to incorporate the world knowledge and improve the representation of text data by incorporating distributional word representation trained on 329,589 Chinese Wikipedia documents. We considered the skip-gram model with vector lengths of 80, 100, 150, 200, and 300. The word vectors were incorporated into the feature set for training machine learning models. We considered k-nearest-neighbors (KNN), random forests, and logistic regressions for the two supervised learning tasks. We developed our research testbed by removing occupation codes and business codes that appeared less than five times. Our final testbed consisted of 5,610 records that covered 89 unique business codes and 189 unique occupation codes. Our evaluation using five-fold cross-validation showed that using the word vector trained on Chinese Wikipedia documents improved classification accuracy. Our best model for occupation code prediction achieved a classification accuracy of 0.6113. The best accuracy for business coding is 0.7490. The best accuracy for both models was achieved using the proposed word vector trained using logistic regression.