Labor Space: A spatial representation of the labor market

Keywords: Labor Space, Labor Market, Word Embedding, Computational Economics

Extended Abstract

The labor market can be considered a hierarchical ecosystem of various economic units — skills, jobs, and industries. Firms headhunt, and we hunt-gather some profitable skills and knowledge to survive the capitalized world. Hence, a true understanding of the labor market requires an ecological, holistic perspective considering the interrelationships between industry, occupation, and skill.

However, existing studies have focused on either a single unit (skill, occupation, or industry) separately or a relationship between two units. For example, Alabdulkareem et al. (2018) [1] have visualized a skill space where cognitive and physical skills have polarized to explain wage inequality. Bana et al. (2020) [2] have constructed an occupation space to characterize occupational change over time. While these studies have merits in themselves, analyzing only a singular labor market unit prevents us from understanding the complex evolution of the labor market system driven by multi-unit interactions.

How can we ecologically overview the different levels of labor units? Industries, occupations, and skills appear to be structurally separate. However, they are linked to each other, especially when they are affected by an external economic factor. For instance, the advancement in information technology influences industry, occupation, and skill all together but separately. On the industry level, the innovation in IT impacts on the information industry, while on the occupational dimension, it affects to hire more programmers. On the skill level, it drives workers to learn a programming language such as C or Python. Hence, we can imagine the labor market as a high-dimensional space of multiple types of entities, where each entity is categorized into multiple units — industry, occupation, and skill, in the current example — while they are distant to each other based on conceptual similarity.

Here, we create a high-dimensional embedding space of heterogeneous units in the labor market, called *Labor Space*, using the word embedding approach [5]. In particular, we use the prevalent pre-trained word embedding model Bidirectional Encoder Representations from Transformers (BERT) to map labor unit labels to vectors of real numbers, creating a vector space. While most existing pre-trained language models use a unidirectional approach to interpret semantic relationships within word sequences, BERT jointly conditions on both left and right context in all layers, resulting in more powerful results on natural language processing tasks [3].

Our Labor Space consists of 308 North America Industry Classification System (NAICS) 4-digit code descriptions, 1,016 occupations, and 307 skills descriptions. Fig. 1 presents a brief overview of how we construct Labor Space. The input data consists of titles and descriptions of industries, occupations, and skills. The embedding algorithm represents the components based on their descriptions, and then labels their titles to avoid lexical similarity between labor words in the vector space. Since BERT was trained with 2.5 billion Wikipedia and 0.8 billion BookCorpus data, the resulting embedding space reflects a general conceptual relation, notion, and idea. To capture the latent structure of the labor market, we fine-tuned our model using descriptions from the NAICS, ONET, and ESCO databases to generate the embedded Labor Space.

Labor Space provides an ecological landscape that connects components of the labor market through semantic relations and reveals multiple consistent spatial patterns. As shown in Fig 2, the whole space is clustered into three sub-spaces: industry, occupation, and skill. Since the projection operation in the continuous vector representation of words measured the shared association between two vectors [4], we project all the vectors onto the industrial structure dimension created by vector subtraction (service-biased industry - production-biased industry). The color tint corresponds to the projection values. This embedded Labor Space has the potential to support a wide range of applications, including labor market analysis, skills mapping, and workforce planning.

It also enables us to identify the entities strongly connected to other levels of units, for instance, industry-friendly skills or skill-friendly occupations in each industrial structure. One possible application of the Labor Space would be measuring the effect of recession by occupation type. We would expect a greater impact of the recession on industry-friendly occupations rather than skill-friendly occupations. The assumption could be validated by comparing the annual median wage differential between industry-friendly and skill-friendly occupations in the recession period. Through further analyses, we expect that Labor Space allows us to trace and predict the evolutionary processes of industries, occupations, and skills all together.

References

[1] A. Alabdulkareem, M. R. Frank, L. Sun, B. AlShebli, C. Hidalgo, and I. Rahwan. Unpacking the polarization of workplace skills. *Science advances*, 4(7):eaao6030, 2018.

[2] S. Bana, E. Brynjolfsson, D. Rock, and S. Steffen. job2vec: Learning a representation of jobs. 2020.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] A. C. Kozlowski, M. Taddy, and J. A. Evans. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949, 2019.

[5] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

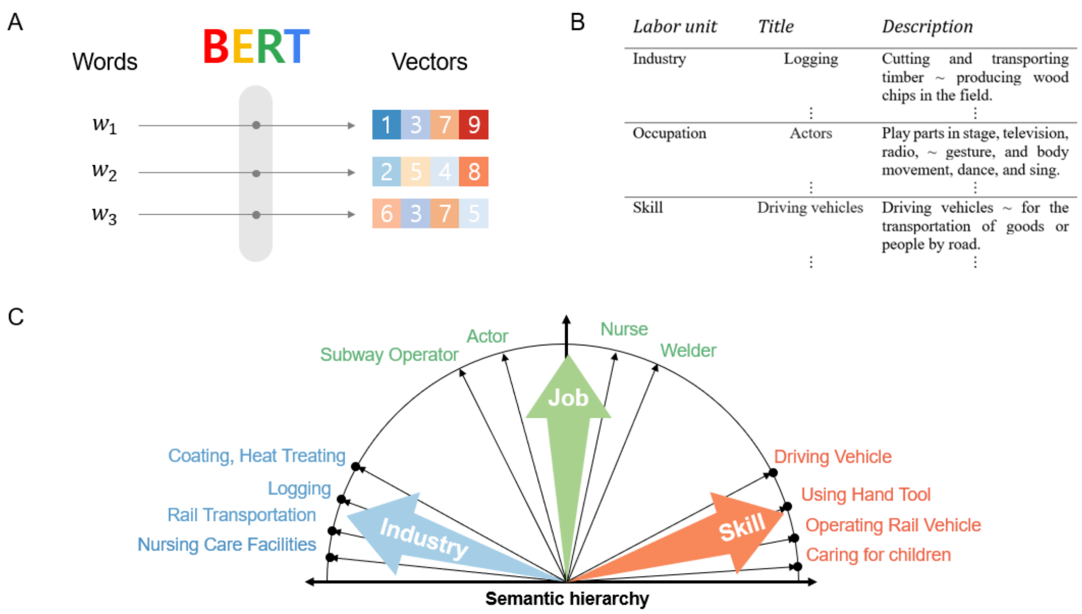

Figure 1: **A-C : Constructing the Labor Space.** (A) Google's BERT is a pre-trained large language model that represents words as vectors. Each element of these vectors is associated with the semantic relationships between words. (B) To apply BERT to the labor market, we fine-tuned it using descriptions from existing databases such as NAICS, ONET, and ESCO. We extracted the vectors representing these descriptions and labeled them with the corresponding titles. (C) This allowed us to create a vector representation of the heterogeneous units of the labor market. By embedding these units in a semantic hierarchy, we can see the direction in which they move in the embedding space. This approach provides a powerful tool for understanding the relationships between different labor titles and descriptions in the labor market.

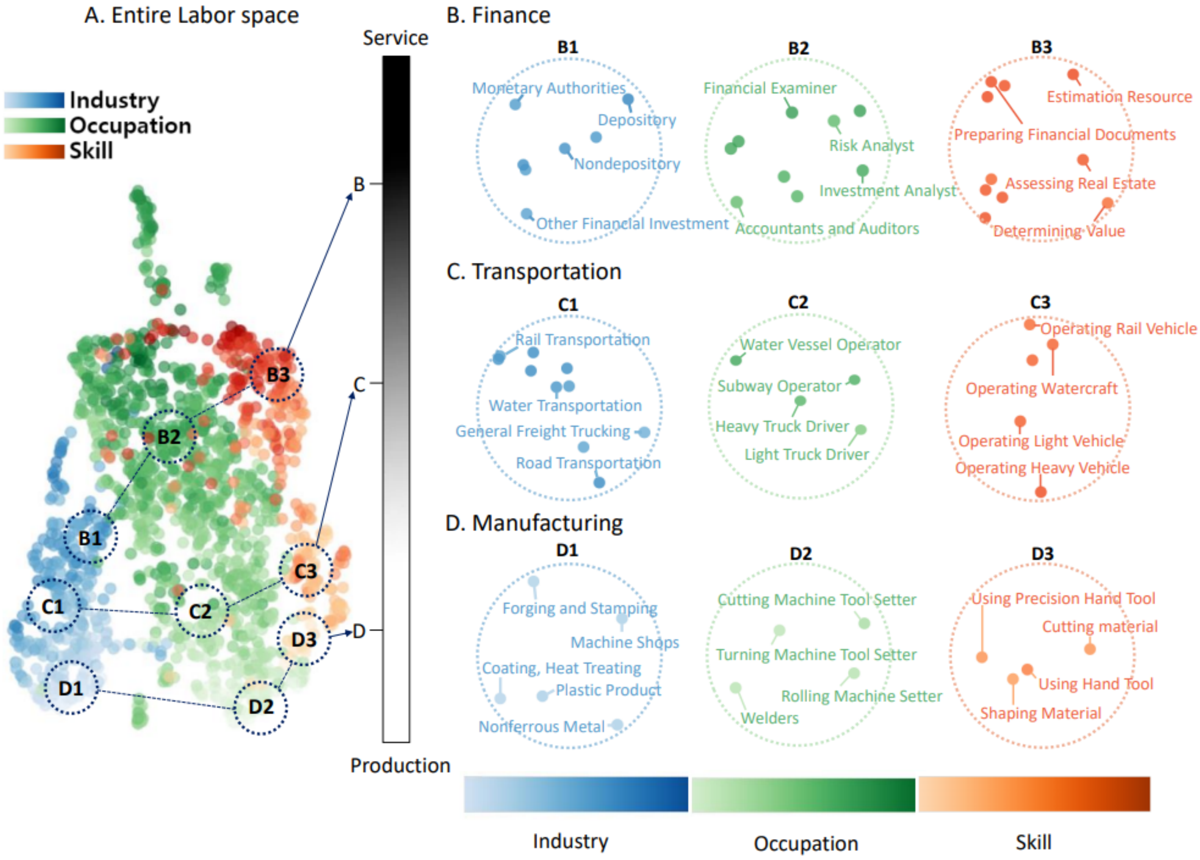

Figure 2: **A-D: The labor space and its labor units.** The entire labor space (A) and industry-occupation-skill elements of exemplary sectors - finance (B), transportation (C), and manufacturing (D). For visualization, we reduce the original vector of the labor elements (dimension of 768) to a 2-dimensional vector space using UMAP. On the horizontal axis, we observe the hierarchical change of labor class. On the vertical axis, we discovered the industrial structure sorted from production to services.