Patent and Labor Market Outcomes Analysis Based on Natural Language Model

Hyeongjun Bang, Do Won Kwak, Kahyun Lee

Abstract

This report aims to analyze the economic effects and labor market outcomes of patents using two approaches: traditional econometric methods and natural language processing. Patents are one of the outcomes of research and development (R&D) activities, serving as an indirect indicator of R&D or productivity.

In this study, rather than measuring the performance or investment level of R&D activities by patents, we used the R&D investment costs of companies and the "R&D Activity Index" (R&D AI) developed by Korea Rating & Data (KoDATA). The analysis showed that R&D activities measured in terms of R&D investment costs exhibited a positive correlation with productivity measured by per capita sales. However, the R&D AI has a negative correlation. These findings could stem from differences in measurement methods despite the high correlation between the two indicators. Additionally, the R&D AI reflects the long-term sustainability of R&D activities, while R&D investment costs represent short-term expenditures. The increase in R&D expenditure with improved financial performance may be perceived as a direct correlation between R&D investment costs and productivity.

However, an employment indicator measured by the number of employees showed a positive correlation with both the R&D AI and R&D investment costs. This suggests that because R&D professionals have different educational and skill requirements compared to other types of white-collar workers, increasing new hires may be necessary to expand R&D activities. There may be a reverse causality in which the need to expand business and maintain market share drives the expansion of R&D activities. Which of these pathways is greater remains to be fully elucidated.

Moreover, both the R&D Activity Index and R&D investment costs, below a certain threshold, evince negligible associations with indicators such as productivity, sales, or employment. This suggests that R&D activities need to reach at a certain scale or higher in order to yield outcomes.

In Chapter 4, we evaluated Korea's R&D support policy, the National R&D Support Program, from a similar perspective. The results showed an increase in patent applications or registrations among companies participating in the program. However, considering the pre-existing increasing trend in

patents before participation, the actual increase in patent registrations was not observed. This could be due to a time lag effect between R&D activities and patent registrations, or due to selection bias, where companies with a strong interest in R&D actively participate in the program.

Participation in National R&D Support Program was found to have a negative correlation with financial performance such as net profit or sales, likely due to the cost nature of R&D activities in the early stages of investment. However, there was a clear positive effect on employment in companies participating in the program, possibly because of the substantial support for R&D activities leading to long-term productivity growth and subsequent increases in company scale and employment.

Patents encompass diverse textual information, such as abstracts and descriptions. We analyzed the novelty of patents, their value measured by the number of citations, and the characteristics related to employment using natural language processing. The results showed that measuring the novelty of Korean patents using a Korean-based natural language model yielded significantly better results than multilingual models trained primarily on English. This suggests a need for establishing Korean-based natural language models for patent examination or analysis to better capture the distinct linguistic features of Korean.

Accurate value of each patent is unobservable, and estimating it poses a challenge. Consequently, previous literature uses the number of citations to measures the value of a patent, where more citation numbers indicate higher patent value. We examined patents registered by the same corporation over the past 20 years, categorized by the number of citations, and observed the linguistic characteristics to assess the number of citations of patent registered by a corporation as an indicator of productivity. While we found substantive linguistic variances across industries, disparities predicated on the number of citations were not observed. Nonetheless, patents cited over 1,000 times exhibited differences in natural language characteristics, indicating the potential for obtaining more accurate conclusions in future research

Finally, to identify employment effects due to patents based on natural language features, we examined differences in patents between periods of significant employment increase or decrease in machinery and equipment industries. While clear differences were observed between periods, distinct differences in natural language features of patents regarding employment changes were not apparent.

These findings provide several policy implications:

Primarily, achieving actual productivity growth from R&D activities requires investments beyond a certain scale. This policy implication does not necessarily support R&D investment exclusively for large corporations. Supporting large-scale R&D activities for small and medium-sized enterprises (SMEs) can facilitate subsequent independent R&D activities, thereby promoting scaling-up of new SMEs. This also suggests that concentrating on a few large-scale projects may be more efficient than conducting various R&D-related projects.

Additionally, as the discovered cyclical relationship between employment and R&D activities

suggests, R&D activities necessary create high-quality jobs. Since transitioning traditional production or administrative workers to R&D specialists is challenging, companies intending to sustainably expand R&D activities will inevitably have to hire R&D specialists. R&D positions generally require higher educational qualifications and offer decent wages and working conditions, and transitioning non-R&D related blue-collar and white-collar workers to R&D specialists is challenging. Therefore, policies supporting R&D activities can be implemented as one of the strategies to supply high-quality jobs in the labor market.

Last, Korean-based natural language processing (NLP) models have distinct advantages over multilingual-based NLP models for analyzing documents in Korean. Therefore, developing Korean-based NLP models for patent examination, R&D performance evaluation, and patent and R&D analysis is necessary to help the work of examiners and companies, as well as to encourage Korean-based NLP research.