구인공고에 나타난 숙련수요 : IT 직종을 중심으로*

장지연 · 심지환**

본고에서는 구인공고 텍스트를 입력하면, 이 구인공고가 요구하는 숙련을 'ICT 숙련분류체계'의 숙련코드로 제시하는 딥러닝 모델을 소개하였다. 본고에서 소개한 딥러닝 모델은 하나의 구인공고는 여러 가지 숙련을 요구할 수 있으므로 복수의 숙련 소분류 코드를 제시하도록 하여, 2017~2022년 '사람인' 사이트에 게시된 구인공고와 2021~2022년 '잡코리아'에 게시된 구인공고 약 133만 건에 적용하였다. 이렇게 포착된 숙련수요를 시기별, 업종별로 시각화하고, AI 관련 숙련에 대한 수요도 제시하였다.

Ⅰ. 머리말

기업의 숙련수요가 가시적으로 드러나는 장(場)은 온라인 구인공고이다. 어떤 일자리에서 어떤 능력을 가진 사람을 구한다는 내용이 여기에 드러난다. 오늘날 구인공고는 대부분 웹사이트에 게시된다. 취업알선이 오프라인에서만 이루어지는 경우는 매우 적으며, 특히 IT 일자리의 경우, 이 규모는 더욱 작을 것으로 예상된다.

온라인 취업알선 사이트에 올라오는 구인공고(OJPs : Online Job Postings)를 분석하면 숙련수요를 알 수 있지만, 주로 자연어 텍스트로 되어 있기 때문에 분석이 쉽지 않다. 온라인에서의 대다수 구인구직 관련 데이터는 기존의 연산 및 통계 프로그램이 분석하던 정형 데이터가 아닌, 텍스트나 이미지 등을 포함하는 비정형데이터이다.

오늘날 컴퓨팅 성능 향상과 딥러닝 기법의 발전으로 비정형데이터(텍스트, 이미지, 음성 자료

* 이 글은 장지연 외(2023), 『정보통신기술직의 숙련수요 : 구인공고 텍스트 분석을 통한 시론』의 제3장을 요약 · 정리한 것이다.

** 장지연=한국노동연구원 선임연구위원(jchang@kli.re.kr),
심지환=국민대학교 데이터사이언스학과 박사과정(sim2080@gmail.com).

등)를 분석할 수 있는 환경이 조성되었다. 자연어 처리(Natural Language Processing) 기법이 빠르게 발전하고 있으며, 최근에는 자연어 처리에도 딥러닝을 활용한 연구가 주류를 형성하고 있다. 비정형데이터 분석기법을 활용하면 기존의 계량경제학적 방식으로는 파악하기 힘들었던 새로운 사실들을 발견하거나 혹은 새로운 트렌드나 정보를 보다 신속하게 알아낼 수 있다.

해외에서도 텍스트 형태로 되어 있는 노동시장 관련 데이터를 수집하고 분석하려는 노력이 체계적으로 이루어지고 있으며, 그중에서도 가장 활발하게 연구가 이루어지는 데이터는 온라인 구인공고 데이터이다. 미국의 민간기업인 Lightcast(구 Burning Glass Technology)는 미국뿐만 아니라 전 세계 영어권 국가에서 온라인 구인공고를 수집하여, 연구, 기업 컨설팅, 정책 컨설팅에 활용하고 있다. 이 자료를 활용한 다양한 논문이 제출되고 있으며, OECD나 ILO 같은 국제기구에서도 활용하고 있다.

본 연구에서는 우리나라에서 대표적인 온라인 구인공고 사이트인 사람인(www.saramin.co.kr)과 잡코리아(www.jobkorea.co.kr)에 게시된 구인공고를 분석한다. 사람인 사이트에서는 2016년 말부터 2022년 말까지에 해당하는 데이터를 수집하였고, 잡코리아에서는 2021년과 2022년의 데이터를 수집하였다.

분석을 위해서는 다양한 자연어처리 기법을 활용하였다. 자연어처리 기법이란, 각 단어의 형태소를 구분한 후에 이를 의미를 반영하는 수치들의 벡터로 변형하여 컴퓨터가 구별할 수 있도록 만드는 것을 의미한다. 또한, 본 연구의 분석에는 딥러닝(Deep Learning) 기법을 활용하였다. 딥러닝은 인공지능(AI)의 하위개념이다. 분류(classification)에서 탁월한 성과를 보여 오다가 최근에는 생성(generation)에서 더욱 훌륭한 성능을 보여준다.

Ⅱ. 숙련분류모델 개발과 적용

1. XLM-RoBERTa를 이용한 숙련분류모델

XLM-RoBERTa는 기존의 RoBERTa 모델을 기반으로 다양한 언어를 지원하는 대규모 전이학습(transfer learning) 모델이다.[1] 100개의 언어로 구성된 약 2.5TB 규모의 데이터로부터 사전훈련된(pretrained) 모델로서, 텍스트 분류, 시퀀스 레이블링, 질의응답과 같은 NLU(Natural Language Understanding) 태스크에서 우수한 성능을 보인다(Conneau et al., 2019). XLM-RoBERTa는 기

1) XLM은 Cross-lingual Language Model을 의미한다.

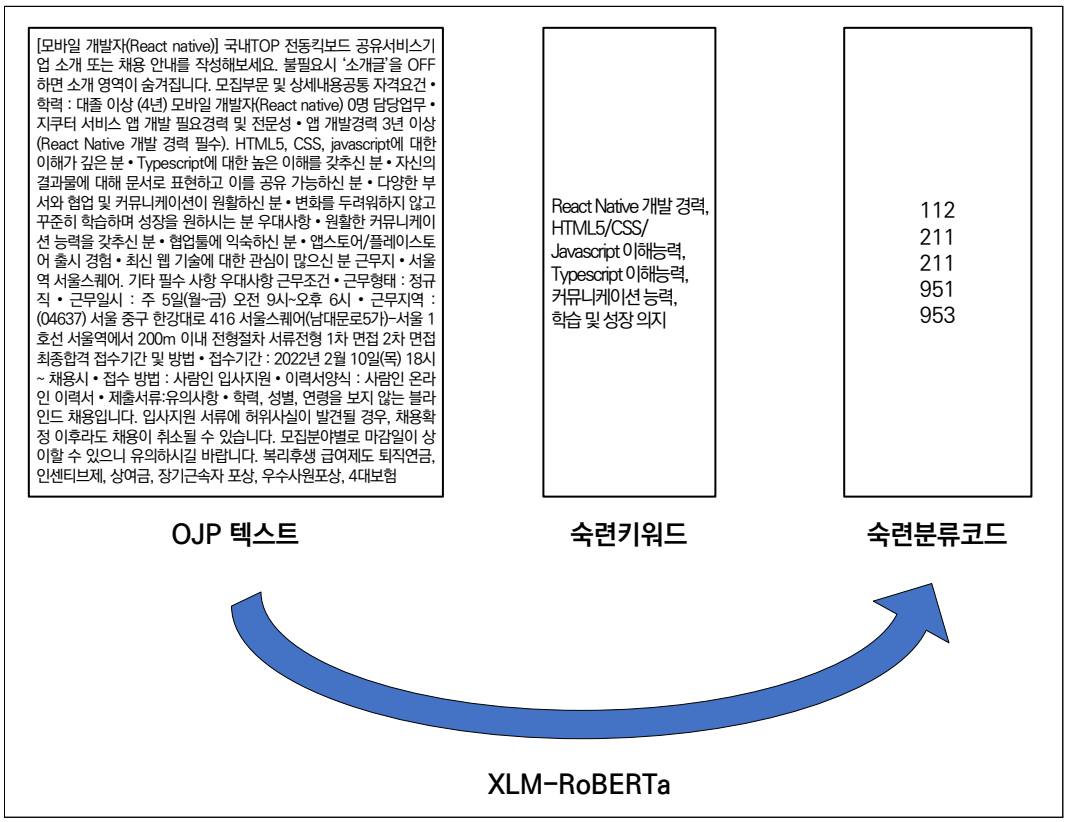
존의 BERT와 같이 MLM(Masked Language Model) 방식으로 훈련되지만, 더 다양하고 방대한 양의 텍스트 데이터로 사전학습되었기 때문에 다국어표현의 분류 작업에 효과적으로 적용될 수 있을 것이다. 특히, 국내 IT 관련 온라인 구인공고 특성상 프로그래밍 언어(e.g. PYTHON, JAVA 등) 또는 응용소프트웨어 명(e.g. FRAMEWORK, JAMOVI 등) 등이 요구 숙련으로서 포함되어 있는 경우가 많다. 즉, 국문 및 영문이 혼재되어 있는 온라인 구인공고의 특징을 고려해 보았을 때, XLM-RoBERTa가 효과적으로 숙련 단어를 분류하는 작업에도 적용될 수 있을 것으로 기대한다.

XLM-RoBERTa는 텍스트를 입력받았을 때, 텍스트를 구성하는 단어의 의미 혹은 문장 간의 문맥을 파악할 수 있는 언어모델이다. 그리고 이러한 언어모델(BERT, GPT, XLM 등)들은 여러 응용작업들에 활용될 수 있는데, 여기서 응용작업들을 Downstream Task라 한다. 가령, 본 연구와 같이 입력된 OJPs에서 숙련을 분류하는 Downstream Task를 정의할 때, OJPs가 가지는 문맥적 의미를 토대로 숙련을 추출할 수 있는 분류기(classifier)만을 미세조정(fine-tuning)해 주는 작업을 수행하면 된다. 단, XLM-RoBERTa 모델이 입력 문서를 적절하게 문맥적 의미를 내포하고 있는 문서벡터로 변환시키기 위해서는 사전학습에 사용되었던 단어 사전을 바탕으로 입력 문서가 가지는 각각의 단어들을 숫자로 변환해 주는 작업이 필요하다.

본 연구에서는 위와 같은 작업을 수행하기 위해 HuggingFace에서 제공하는 Python 라이브러리인 transformers를 활용하여 사전학습 모델과 해당 사전학습 모델에 사용되었던 토크나이저를 활용한다. 먼저, XLM-RoBERTa의 토크나이저를 활용하여 문자로 구성된 OJPs를 숫자로 구성된 시퀀스로 변환한다. 변환된 숫자 시퀀스는 XLM-RoBERTa의 사전학습에 활용되었던 단어들이 가졌던 숫자를 그대로 상속받도록 토크나이저가 설계되어 있다. 예를 들어, 사전학습에 사용되었던 문서 중 PYTHON이라는 단어가 번호 10을 가졌었다면, OJPs에 포함된 PYTHON이라는 단어도 동일하게 번호 10을 가지게 된다. 그리고 이와 같이 토크나이저를 통해 변환된 숫자 시퀀스는 XLM-RoBERTa의 입력값으로 구성되며, 출력값은 입력 문서에 대한 문서벡터가 된다. 그리고 이 문서벡터는 분류기의 입력값으로 주어지고, 출력값으로 숙련분류코드가 되도록 하는 구조로 설계되어 있다.

ICT 숙련분류체계를 만들기 위해 사용했던 약 16만 건의 구인공고 데이터와 각 구인공고에 매칭된 숙련분류코드는 숙련분류모델을 만드는 데 훈련데이터로 다시 한 번 사용된다. 어떤 구인공고에 어떤 숙련코드가 매칭되었는지 쌍(pair)으로 묶으면 숙련코드 추출모델의 훈련데이터로 사용할 수 있다. 예전이라면 사람이 주석달기(annotating)를 통해서 훈련데이터를 만들었을 것을 OpenAI가 제공하는 GPT를 이용하여 손쉽게 만들었다고 보면 된다.

[그림 1] OJP 텍스트애레 숙련분류코드 부여하기



2. 적 용

XLM-RoBERTa를 기반으로 만든 숙련분류모델을 사람인과 잡코리아 ICT 일자리 구인공고 텍스트에 적용하였다. 이렇게 해서 숙련분류코드를 부여한 구인공고 건수는 약 133만 건이다(표 1 참조). 이들 구인공고에 나타난 숙련수요는 앞서 언급한 분류코드를 기준으로 정량화할 수 있다.

〈표 1〉 분석에 사용된 구인공고 데이터 규모

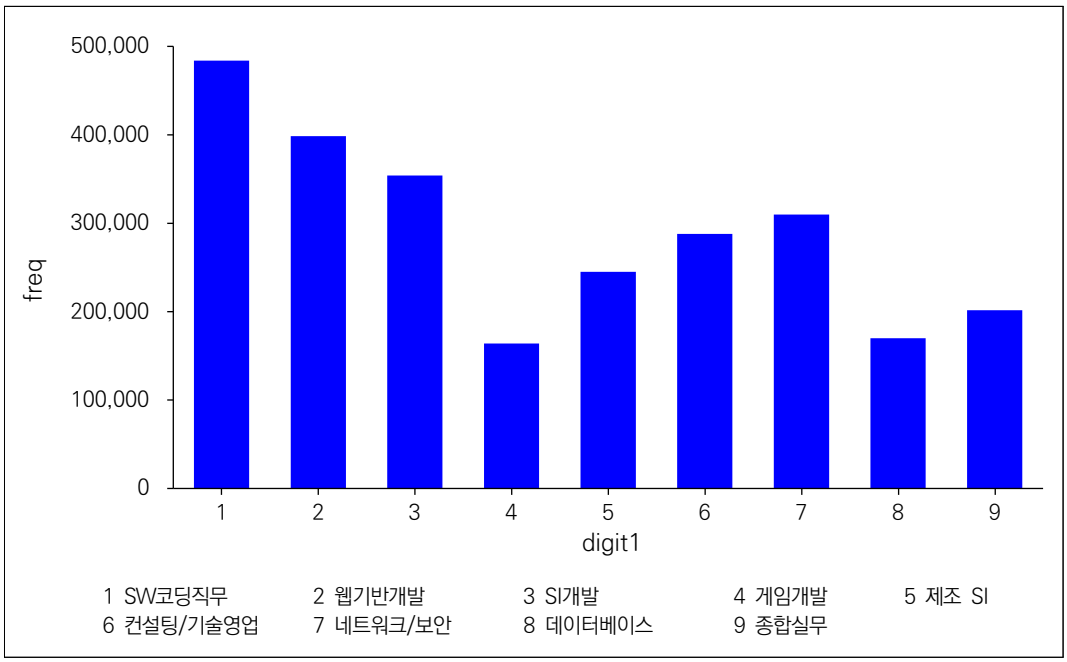
사람인 2016~2021년	673,683건
사람인 2022년	176,945건
잡코리아 2021년	252,430건
잡코리아 2022년	229,707건
전체	1,332,765건

Ⅲ. ICT 숙련수요

1. ICT 숙련수요 전체

이 장에서는 사람인과 잡코리아 ICT 일자리 구인공고에 나타난 숙련수요를 시각화하였다. [그림 2]는 숙련 대분류별 숙련수요다. 범용 소프트웨어 코딩직무와 웹기반개발, 시스템통합(SI) 개발 순으로 수요가 많았고, 컨설팅/기술영업의 수요도 많았다.

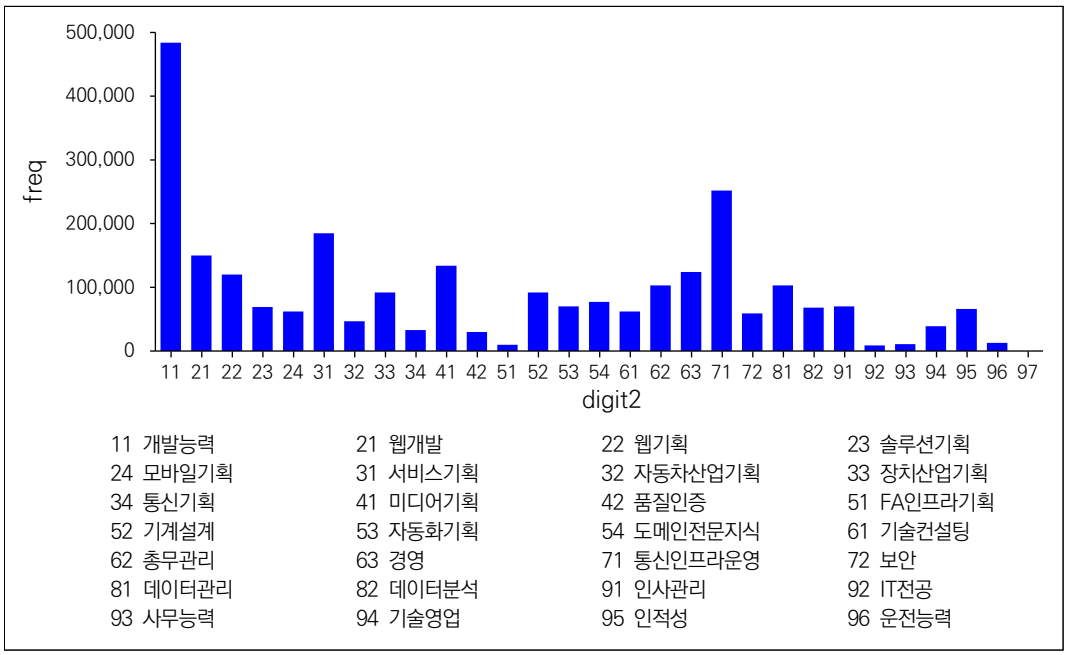
[그림 2] 숙련 대분류별 수요



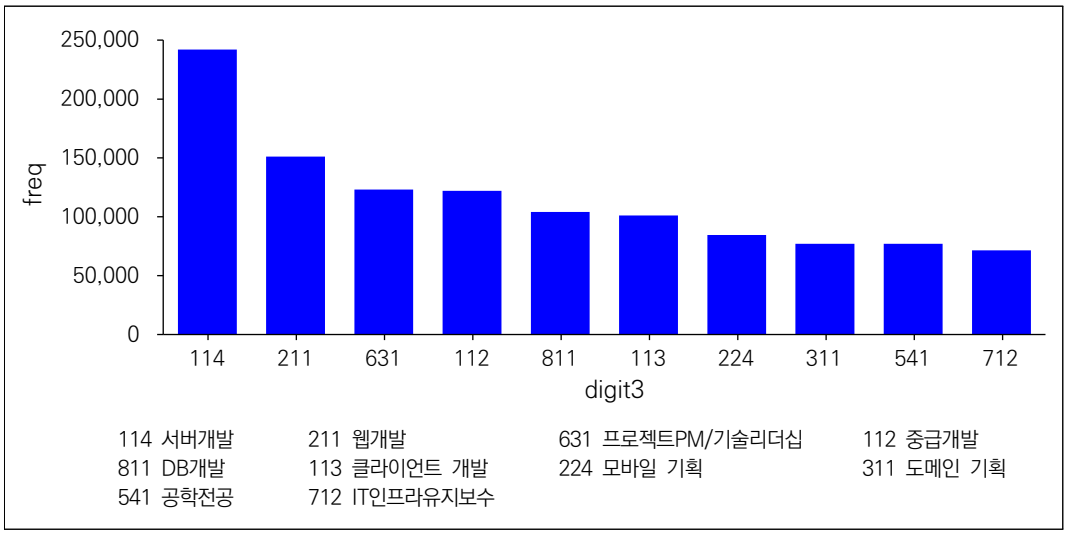
[그림 3]은 숙련 중분류별 수요다. 일반적인 개발능력(11)을 요구하는 경우가 압도적으로 많고, 통신인프라운영(71), 서비스기획(31) 능력을 요구하는 일자리가 많은 것으로 나타났다.

[그림 4]는 숙련 소분류별 수요 순으로 상위 10개를 나타낸 것이다. JAVA 등 일반적인 개발능력을 요구하는 114, 112, 113이 모두 포함되어 있고, 웹개발(211)과 모바일 개발을 위한 기획(224)이 포함되어 있다. 프로젝트를 전반적으로 이끌어갈 수 있는 다방면의 능력을 요구한다는 경우(631)도 높게 나타났다.

[그림 3] 숙련 중분류별 수요

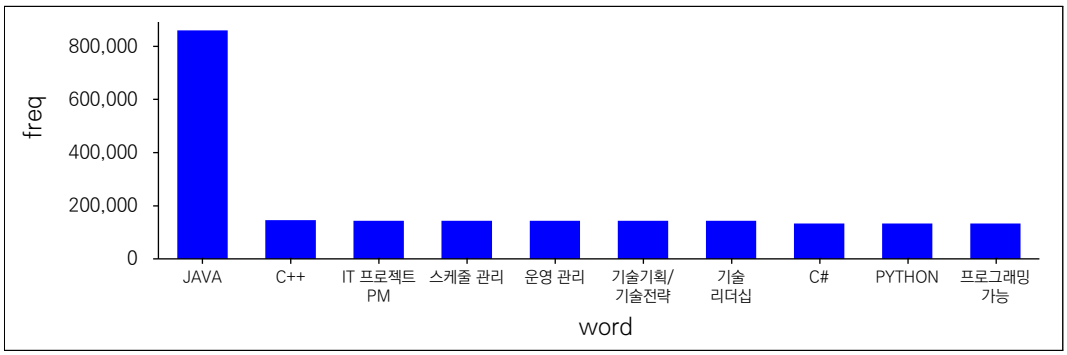


[그림 4] 숙련 소분류별 수요 순위 Top 10



[그림 5]는 자주 등장하는 숙련 키워드 10개를 나열한 것인데, JAVA가 압도적으로 많이 등장하고, 그 밖에 자주 등장하는 키워드는 C++, IT프로젝트 PM, 스케줄 관리, 운영관리, 기술기획, 기술리더십, C#, Python, 프로그래밍 가능 등이었다.

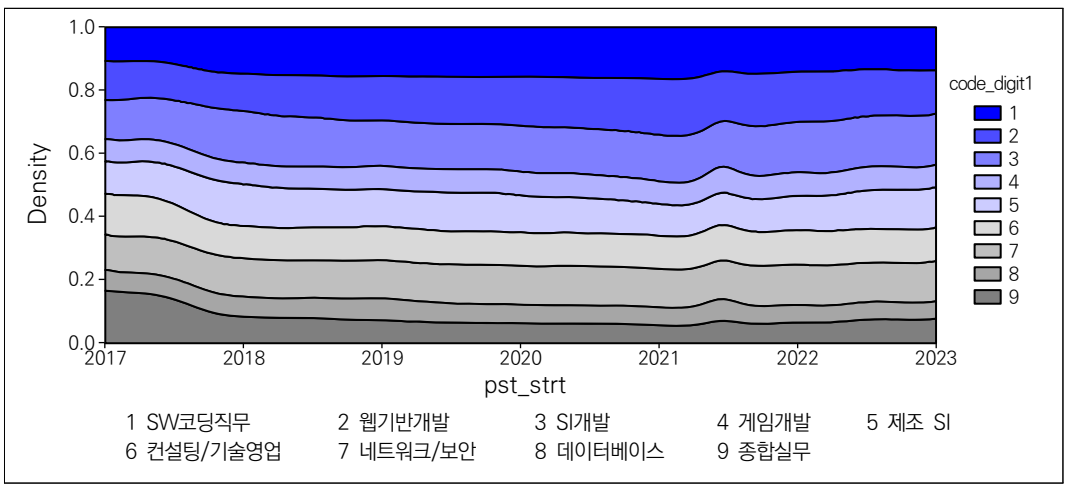
[그림 5] 숙련 키워드별 수요 순위 Top 10



2. 연도별 숙련수요 변화

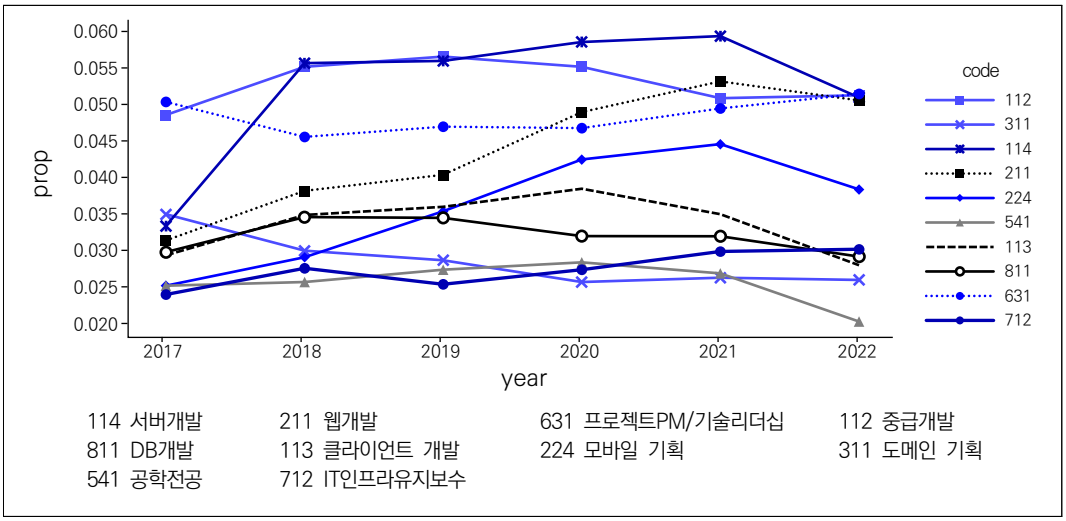
'사람인' 구인공고는 2017년부터 2022년 말까지 자료가 수집되었기 때문에 6년간 숙련수요의 변화를 살펴볼 수 있다. 먼저 [그림 6]은 대분류별 수요변화인데, 전체를 1로 두고 상대적인 수요변화를 본 것이다.[2] 2017년에 비해 그 이후 연도에는 종합실무(분류코드9)가 줄어들었다. 분류코드9에는 협업능력, 커뮤니케이션능력, 외국어능력과 인사업무 능력 등 '소프트스킬'이라고 할 수 있는 하위범주들이 포함되어 있다. 이에 비해 '하드스킬'이라고 할 수 있는 대분류 1, 2, 3 수요는 증가하는 편이다.

[그림 6] 숙련 대분류별 수요 변화(2017~2022)



2) 시계열 분석에서 절대수치가 아니라 상대적인 비중을 살펴본 것은, 최근으로 올수록 '사람인'에 올라오는 구인공고 수가 많아지는 추세를 보이기 때문이다.

[그림 7] 숙련 소분류별 수요 비중 변화

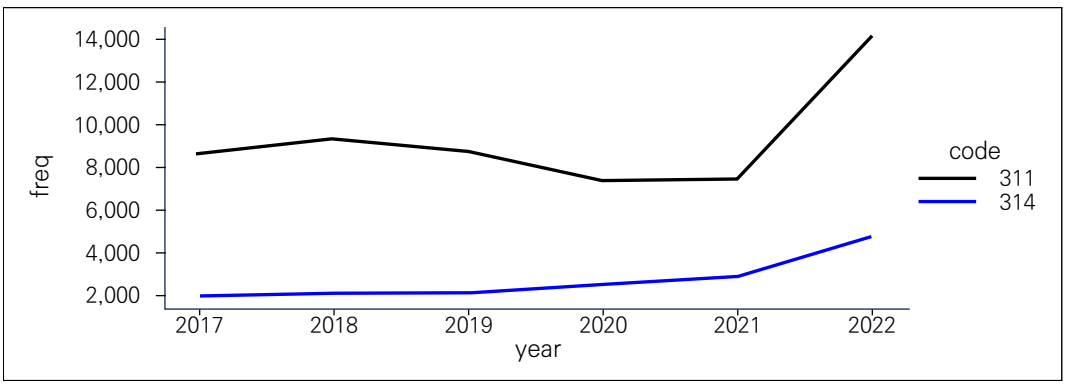


[그림 7]은 앞에서 상대적으로 수요가 많은 것으로 나타났던 숙련 소분류별 숙련코드 10개의 수요를 시계열로 살펴본 것이다. 이 기간 동안 중급개발(112)의 수요가 꾸준히 증가한 것으로 나타났다. 서버개발(114)과 DB개발(811)도 증가 추세라고 볼 수 있다. 웹개발(211)과 프로젝트 PM/기술리더십(631)은 지속적으로 높은 수준을 유지하고 있다.

3. AI 관련 숙련수요

최근 관심이 높아지고 있는 AI 관련 숙련수요는 소분류 311 자연어처리와 314 AI 기획으로 포착된다. [그림 8]에 따르면 AI 숙련수요는 2022년에 큰 폭으로 증가하였다.

[그림 8] AI 관련 숙련수요



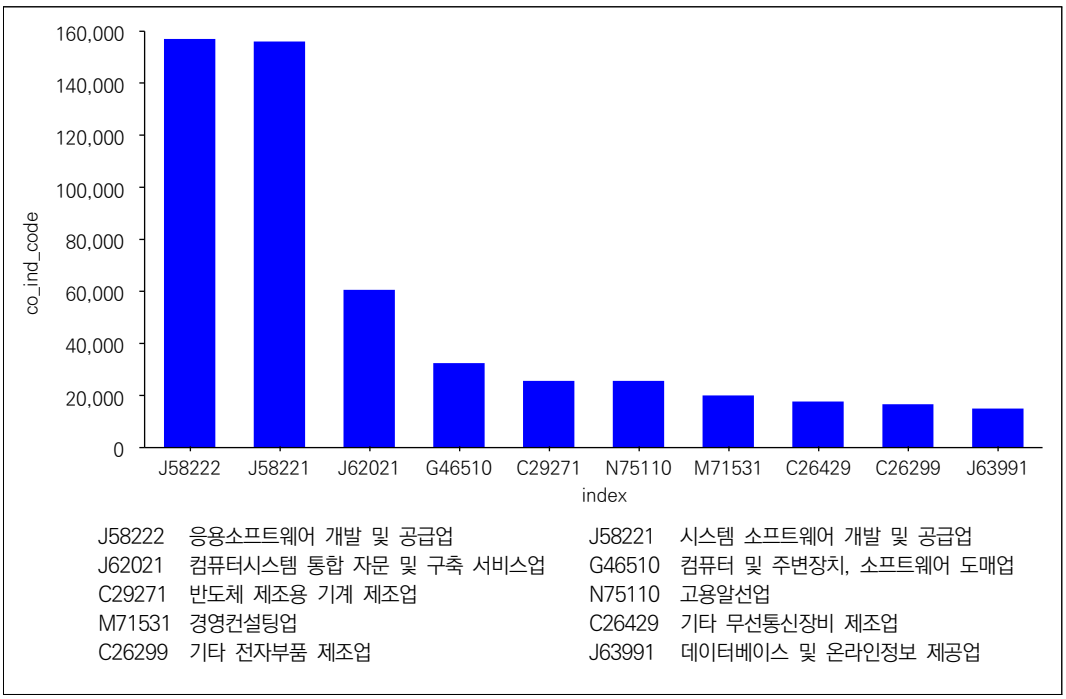
IV. 업종별 ICT 숙련수요

ICT 숙련수요가 높은 업종은 어떤 업종인지 알아보기 위해서는 구인공고를 낸 기업을 식별해내고 그 기업이 속한 업종을 알아내야 한다. 이를 위해 우리는 구인공고에 나타난 기업명과 주소지를 가지고 한국기업데이터(KED)에서 그 기업을 찾아내는 작업을 수행했다. 실제로 기업을 식별하여 사업자등록번호를 알아낸 것은 전체 구인공고 데이터의 절반 정도에 불과했다. 이 데이터를 활용하여 업종별 숙련수요를 살펴본 결과는 다음과 같다.

[그림 9]에 따르면 숙련수요가 가장 흔하게 포착된 업종(표준산업분류 제10차)은 응용소프트웨어 개발 및 공급업, 시스템 소프트웨어 개발 및 공급업, 컴퓨터시스템 통합 자문 및 구축 서비스업, 컴퓨터 및 주변장치, 소프트웨어 도매업, 반도체 제조용 기계 제조업 순으로 나타났다.

〈표 2〉는 정보통신업과 제조업에서 수요가 높은 ICT 숙련을 순서대로 10개씩 소개한 것이다. 순위는 조금씩 달라지지만, 세분류별로 살펴본 숙련수요는 두 업종 간에 큰 차이가 없었고, 대체로 [그림 4]에서 소개된 숙련소분류코드와 비슷했다.

[그림 9] ICT 숙련수요가 높은 업종 Top 10



〈표 2〉 정보통신업과 제조업에서 수요가 높은 숙련 Top 10

숙련소분류 빈도 순위	J 정보통신업	C 제조업
1	114	114
2	211	112
3	112	211
4	631	631
5	811	541
6	113	811
7	224	113
8	311	224
9	712	311
10	541	712

V. 맺음말

본 연구는 온라인 구인공고에 나타난 숙련수요를 정량화하여 살펴보았다는 데 의의가 있다. 특정 업종 또는 직종에서 어떤 숙련을 요구하는지, 그 규모를 가늠해 볼 수 있다는 의미이다. 산업계에서 사용하는 테크놀로지의 발전이 빠르게 이루어지고 있기 때문에, 노동시장에 새로 진입하고자 하는 청년 구직자뿐 아니라 재직 중인 근로자도 숙련수요의 변화에 발 빠르게 대처할 필요가 있다. 본 연구에서 사용한 방법의 장점은 일단 숙련분류체계가 만들어지고 나면, 데이터의 수집과 정제, 분석과정이 오래 걸리지 않는다는 점이다.

숙련수요 정량화는 크게 두 가지 분야에서 활용도가 높을 것으로 예상된다. 첫째, 정부의 교육·훈련 정책과 고용서비스에서 실용적으로 활용될 수 있다. 기업의 숙련수요가 곧바로 정부의 교육·훈련 프로그램에 반영되고 고용서비스에 적용되는 체계가 만들어지기를 기대한다. 둘째, 노동시장 연구에서 다양하게 활용될 수 있다. 예컨대, 본 연구진은 ICT 숙련수요를 기업의 ICT 투자의 대리변수로 보고 기업의 성과를 설명하는 연구에 사용하였다. 본 연구에서 사용한 텍스트분석기법을 숙련수요뿐 아니라 숙련공급분석에도 적용하여 수요·공급의 매칭 수준을 분석할 수도 있다.

본 연구에서 사용한 데이터와 분석방법은 연구목적에 비추어 볼 때 꼭 필요할 뿐 아니라 앞으로도 확대되어야 할 분석기법임에 틀림없으나, 현재 시점에서 한계도 분명하다. 첫째, 데이터의 커버리지 문제이다.[3] 사람인과 잡코리아에 게시된 구인공고가 IT 일자리 구인공고 전체를

반영한다고 볼 수 없다. 최근에 새롭게 등장한 취업알선 앱이 많고, 특히 단기 일자리나 프리랜서 일자리 공고는 이런 신규 앱에 게시되는 비율이 높을 것으로 짐작된다. Lightcast 데이터는 오히려 구인공고의 중복을 우려하면서 동일한 공고를 제거하는 작업을 한다고 하는데, 우리는 아직 커버리지를 우려하는 수준에 있다. 추후 연구에서는 다른 온라인 구인공고 사이트를 추가하여 커버지리를 향상시킬 필요가 있다.

둘째, 딥러닝을 이용한 분류모델은 정확도가 아직 높은 수준에 다다르지 못했다. 시간을 더 투입해서 다양한 분류모델을 테스트해 보고 데이터의 정제과정도 달리 시도해 볼 여지가 많음에도 불구하고, 본 보고서 작성 시점에서 최선이라 여겨지는 모델을 적용할 수밖에 없었다. 후속 연구를 통해서 분류성능을 향상시켜 가고자 한다.

셋째, 분류모델의 정확도를 더 높인다고 할지라도, 이런 분석기법은 현상을 기술(description)하는 데 그친다는 점을 분명히 할 필요가 있다. 통계적인 추론 작업과는 전혀 다르다. 분류모델을 적용하여 개별 구인공고가 요구하는 숙련을 찾아내는 작업이 기여할 수 있는 역할은 설명변수를 만들어내는 데 있다고 볼 수 있다(Feature Extraction). KLI

[참고문헌]

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, …, and V. Stoyanov(2019), Unsupervised Cross-lingual Representation Learning at Scale. arXiv preprint arXiv:1911.02116.

3) 표본을 추출하여 분석하는 것이 아니기 때문에 대표성이라는 용어를 사용하지 않고 커버리지라는 용어를 사용하였다.