기획논문 「정부학연구」 제30권 제1호(2024): 35~88
DOI https://doi.org/10.19067/jgs.2024.30.1.35

원고투고일: 2024.03.15.
게재확정일: 2024.04.17.

언어모델과 공공부문 행정혁신: 개념, 접근법 및 고려 사항*

박정원**
이규민***
전대성****
최재웅*****
이창용******

정부의 비정형 데이터 제공 및 관리 체계가 확립되면서 언어모델을 활용한 공공부문 행정혁신 연구가 주목받고 있다. 하지만 언어모델을 활용할 필요가 있는 실효성 있는 문제의 발굴과 언어모델의 실질적 적용 방법에 대한 논의는 매우 미비한 실정이다. 본 연구는 언어모델에 대한 이해를 높이고 공공부문에서의 활용을 촉진하기 위한 목적으로 언어모델의 개념 및 접근법과 공공부문에서의 활용 사례를 소개하고, 언어모델 활용을 위한 주요 고려 사항을 제시한다. 먼저, 언어모델의 개념 및 접근법과 초거대 언어모델을 중심으로 최근 연구 동향을 설명한다. 다음으로, 과학기술 분야에 초점을 맞춰 공공부문에서의 언어모델 활용 사례를 소개한다. 마지막으로, 언어모델을 활용할 때 고려해야 하는 사항을 품질 제고 및 활용 영역 확장 관점에서 제시한다. 본 연구가 언어모델을 활용한 공공부문 행정혁신 연구에만 국한되는 것이 아니라, 행정학 및 정책학 분야에서 활용하는 텍스트 데이터 분석 방법론의 고도화와 다각화에 기여할 수 있기를 기대한다.

주제어: 언어모델, 공공부문, 행정혁신

* 모든 저자들은 본 논문의 작성에 1저자로 공동 기여하였음

** 가천대학교에서 산업경영공학 학사학위를 취득하고, 현재 고려대학교 행정학과 석사과정에 재학 중이다. 관심 연구 분야는 자연어 데이터 품질 평가, 언어모델 기반 자연어 처리 응용, 멀티모달 학습을 통한 생성형 AI 응용이다. 최근 연구로는 "사전학습단계 초거대 인공지능 학습용 데이터의 품질 요소 및 검증방안 고찰"(2023), "자연어 분야 인공지능 학습용 데이터의 구축 추세 및 평가 방안 고찰"(2023) 등이 있다(E-mail: jungwon642@korea.ac.kr)

*** 울산과학기술원에서 경영과학 박사학위(2023년, 논문명: Development of decision support systems for technology and business opportunity analysis)를 취득하였으며, 울산과학기술원 융합경영대학원에서 대우교수를 역임한 후, 현재 고려대학교 통계학과 기초연구실 박사 후 연구원으로 재직 중이다. 주요 연구 분야는 비즈니스 분석, 특허 분석 기반 기술경영, 기계학습 및 심층학습 응용이다. 최근 연구로는 “Inventor-licensee matchmaking for university technology licensing: A fastText approach”(2023), “A convolutional neural network model for SOH estimation of Li-ion batteries with physical interpretability”(2023), “Towards expert–machine collaborations for technology valuation: An interpretable machine learning approach”(2022) 등이 있다(E-mail: optimizt@korea.ac.kr)

**** 울산과학기술원에서 기계공학 학사학위를 취득하고, 현재 울산과학기술대학원 경영과학부 박사과정에 재학 중이다. 주요 연구 분야는 AI 기반 자연어 처리 모델 응용, 데이터 기반 기술경영, 정량적 과학기술 평가이다. 최근 연구로는 “A doc2vec and local outlier factor approach to measuring the novelty of patents”(2022), “Measuring the novelty of scientific publications: A fastText and local outlier factor approach”(2023) 등이 있다(E-mail: jun65800@unist.ac.kr)

***** 건국대학교에서 기계학습 기반 특허 분석 방법론에 대한 주제로 산업공학 박사학위(2022년)를 취득하였으며, 한국과학기술연구원 계산과학연구센터에서 박사 후 연구원으로 재직 중이다. 주요 연구 분야로는 자연어처리, 기계학습 응용, 언어모델, 지식그래프, 논문/특허 분석이다. 최근 연구로는 “Accelerating materials language processing with large language models”(2024), “Quantitative Topic Analysis of Materials Science Literature Using Natural Language Processing”(2024), “Exploring a technology ecology for technology opportunity analysis: A link prediction approach based on heterogeneous knowledge graphs”(2023) 등이 있다(E-mail: jwchoi95@kist.re.kr).

****** 교신저자, 서울대학교에서 산업공학 박사학위(2011년, 논문명: Development and application of methods for technology intelligence using patent information)를 취득하였으며, 울산과학기술원에서 조교수/부교수, 서강대학교에서 부교수/교수를 역임한 후, 현재 고려대학교 행정학과 교수로 재직 중이다. 주요 연구 분야는 공공부문의 디지털 전환, 기계학습 및 심층학습 응용, 과학기술정책이다. 최근 연구로는 “Inventor-licensee matchmaking for university technology licensing: A fastText approach”(2023), “Towards expert-machine collaborations for technology valuation: An interpretable machine learning approach”(2022), “A review of data analytics in technological forecasting”(2021) 등이 있다(E-mail: changyonglee@korea.ac.kr)

Ⅰ. 서론

「공공데이터 개방 2.0」 정책 시행에 따라 정부의 비정형 데이터 제공 및 관리 체계가 확립되면서 언어모델(language model)을 활용한 공공부문 행정혁신 연구가 주목받고 있다.[1] 언어모델은 자연어를 기계가 이해할 수 있는 형태로 수치화하는 것을 목적으로 하며, 기계 번역(machine translation), 개체명 인식(named entity recognition), 텍스트 생성(text generation) 등 다양한 자연어 처리(natural language processing) 응용 분야에서 핵심적인 역할을 수행한다. 특히 BERT(bidirectional encoder representations from transformers)와 GPT(generative pre-trained transformer) 같이 트랜스포머(transformer) 구조를 활용한 딥러닝 기반의 언어모델은 방대한 양의 텍스트 데이터로부터 언어의 패턴을 학습함으로써 인간의 언어를 컴퓨터가 이해하고 생성할 수 있는 능력을 획기적으로 향상시켰다. 최근 해외 정부와 기업은 다양한 유형의 언어모델을 개발하여 오픈 소스의 형태로 개방하고 있으며, 많은 기업들이 이러한 파운데이션 모델을 미세 조정(fine tuning)하여 도메인에 특화된 지식이 내재화되도록 맞춤화함으로써 의료, 법률, 금융 등 다양한 산업 분야에서의 혁신을 가져오고 있다.

공공부문에서의 언어모델 활용은 정책 보고서, 민원 문서, 법률 문서 등 사용하는 데이터의 유형과 품질에 따라 그 시사점이 달라질 수 있으나, 일반적으로 내부 업무의 효율성 개선과 더불어 데이터 기반 인사이트 제공, 시민 참여 증진, 공공 서비스 품질 향상 등에 기여할 수 있는 것으로 알려져 있다(이창용, 2023). 이에 국내 공공기관은 자체적으로 보유한 데이터를 토대로 미세 조정이 이루어진 언어모델을 활용하여 도메인에 맞춤화된 공공 서비스를 개발하는 데 역량을 집중하고 있다. 예를 들어, 한국행정연구원은 갈등관리 관련 연구보고서, 갈등 사례, 갈등 관련 통계 등을 종합적으로 활용하여 갈등 특화 지식 그래프를 생성하고, 이를 토대로 갈등관리 과정에서 의사결정을 지원할 수 있는 언어모델을 개발하고 있다. 사회보장정보원은 개인 맞춤형 사회보장 제공과 복지제도의 효과 증대를 위해 내담자의 욕구와 위기도를 바탕으로 적합한 자원을 연계하는 언어모델 기반 추천시스템을 개발하고 있다. 특허청은 특허 심사 업무 혁신을 위해 특허행정과 관련된 국·영문 특허공보, 통지서, 특허 분류 정보, 기계

1) 미국의 142개 주요 연방 기관 중 45%의 기관이 업무 수행을 위해 인공지능을 시범적으로 활용한 경험이 있으며, 전체 157건의 사례 중 약 45%의 사례에서 텍스트 데이터를 활용한 것으로 조사되었다(Engstrom et al., 2020). 국내의 경우, 2018년-2020년 사이 조달청 나라장터를 통해 발주된 인공지능 도입 사업 89건 중 약 40%의 사업에서 텍스트 데이터를 활용한 것으로 조사되었다(행정안전부, 2021).

번역·독해 정보, 상담사례집 등 7종의 정보를 학습한 자체 언어모델을 구축하고 있다. 한국지능정보사회진흥원은 일상어뿐만 아니라 전문용어와 법률 용어 등 공공기관에서 자주 사용하는 용어와 개조식 어법을 집중적으로 학습시키고 정보 유출 방지 등 보안을 강화함으로써 공공부문에 특화된 한국형 언어모델을 구축하고 있다(한국지능정보사회진흥원, 2023).

이러한 변화에도 불구하고 국내 행정학 및 정책학 학계에서의 언어모델에 대한 논의는 상당히 제한적인 범위에서 이루어지고 있다. 특히 언어모델을 활용할 필요가 있는 실효성 있는 문제의 발굴과 언어모델의 실질적 적용 방법에 대한 논의는 매우 미비한 실정이다. 현재 공공부문에서 다양한 유형의 인공지능 사업이 수행되고 있지만 대부분 챗봇을 활용한 민원 상담 등 기존 서비스의 개선에만 초점을 맞추고 있으며, 이에 따라 언어모델을 활용한 새로운 공공 서비스 발굴 및 활용 영역 확장을 위한 논의가 필요하다(행정안전부, 2021).

본 연구는 언어모델을 활용한 공공부문의 행정혁신을 촉진하기 위해 언어모델의 개념 및 주요 접근법과 공공부문에서의 활용 사례를 소개하고, 언어모델 활용을 위한 주요 고려 사항을 제시한다. 본 연구의 주요 목적은 특정 영역이나 방법에 대한 상세한 문헌조사 결과를 제공하는 것이 아니라, 다양한 유형의 언어모델 접근법을 개념적으로 설명하여 포괄적 이해를 돕고 공공부문에서의 구체적 활용 방안을 논의하는 것이다. 본 연구가 언어모델을 활용한 공공부문의 행정혁신 연구에만 국한되는 것이 아니라, 행정학 및 정책학 분야에서 활용하는 텍스트 데이터 분석 방법론의 고도화와 다각화에 기여할 수 있기를 기대한다.

Ⅱ. 언어모델의 개념 및 주요 접근법

언어모델은 자연어로 이루어진 텍스트 데이터를 컴퓨터가 이해할 수 있는 형태로 수치화하여 표현하는 모델을 지칭하며, 자연어를 이해하는 방식에 따라 빈도 기반, 의미 기반, 문맥 기반 언어모델로 구분된다. 빈도 기반 언어모델은 가장 초기의 언어모델로, 텍스트 내 단어의 출현 빈도를 바탕으로 텍스트를 수치화한다. 이후 인공신경망이 발전함에 따라 텍스트 데이터를 학습함으로써 단어의 의미를 고려하는 의미 기반 언어모델이 등장하였다. 최근에는 더욱 깊고 복잡해진 딥러닝 모델이 등장하면서 문맥에 따른 단어의 의미 변화를 고려하여 텍스트를 이해하는 문맥 기반 언어모델이 개발되었다. 특히, 문맥 기반 언어모델의 경우 대규모 텍스트 데이터를 학습함으로써 텍스

트 분류, 번역과 같은 전통적인 텍스트 분석뿐 아니라 질의응답, 텍스트 생성 등 고도화된 작업에서도 획기적인 성능 개선을 보이며 산업 전반에서 활발히 활용되고 있다. 본 장에서는 앞서 언급한 구분 방식을 토대로 (1) 빈도 기반 언어모델, (2) 의미 기반 언어모델, (3) 문맥 기반 언어모델로 구분하여 언어모델의 주요 접근법을 소개한다.

1. 빈도 기반 언어모델

빈도 기반 언어모델은 단어의 출현 빈도를 바탕으로 텍스트를 수치화하여 표현하는 언어모델이다. 대표적인 빈도 기반 언어모델로는 원핫인코딩(one-hot encoding), BoW(bag of words), BoN(bag of N-grams), TF-IDF(term frequency-inverse document frequency) 등이 있으며, 이들은 기본적으로 분석 대상 문서 집합(corpus) 내 모든 고유한 단어의 집합(vocabulary)을 기반으로 텍스트를 벡터 형태로 표현한다.

원핫인코딩은 가장 기초적인 빈도 기반 언어모델로, 각 단어를 단어 집합의 크기와 같은 차원을 가지는 벡터로 변환하는 방식이다. 단어 집합 내 각 단어에 고유한 정수 인덱스를 부여하고, 주어진 단어에 대해 벡터 내 해당 단어의 인덱스에는 1을, 나머지에는 0을 할당함으로써 단어를 원핫 벡터 형태로 변환한다(Harris and Harris, 2016).[2)]

BoW는 유사한 내용의 텍스트는 포함하는 단어의 종류와 빈도가 유사할 것이라는 가정 하에, 각 단어가 텍스트 내에서 출현한 횟수를 반영하여 단어를 벡터로 변환하는 방식이다(Harris, 1954). 이는 원핫인코딩과 유사하게 단어 집합의 크기와 같은 차원을 가지는 벡터를 구축하되, 벡터 내 각 인덱스에 해당 단어의 출현 횟수를 할당한다. 이에 따라 여러 단어로 이루어진 텍스트는 해당 텍스트 내 단어의 출현 빈도가 기록된 하나의 벡터로 표현될 수 있다.[3)] 그러나 BoW는 단어의 출현 빈도만을 반영하며 텍스트 내 단어가 배치되는 순서는 고려하지 않는다. 이에 단어 집합을 개별 단어가 아닌 텍스트에서 연속으로 출현한 N개의 단어를 N-gram[4)]으로 정의하고, 이들의 출현 빈도를 기반으로 텍스트를 벡터로 표현하는 BoN 방식이 제시된 바 있다.[5)]

---

2) 단어 집합이 “사과”, “바나나”, “오렌지” 3가지의 단어로 이루어진 경우, 사과는 [1,0,0], 바나나는 [0,1,0], 오렌지는 [0,0,1]과 같은 수치형 벡터로 표현된다.

3) 주어진 단어 집합이 {“the”, “cat”, “sat”, “hat”, “in”, “with”}일 때, “the cat sat”이라는 문장은 [1,1,1,0,0,0]의 벡터로 표현되고, “the cat with the hat”이라는 문장은 [2,1,0,1,0,1]의 벡터로 표현된다.

4) N-gram은 포함되는 단어의 수에 따라 bi-gram(2개), tri-gram(3개) 등 다른 명칭을 갖는다.

5) “cat eats mouse, mouse eats food”라는 문장이 있을 때, bi-gram으로 BoN을 생성하는 경우 “cat eats”, “eats mouse”, “mouse mouse”, “mouse eats”, “eats food”의 5개

TF-IDF는 단어의 출현 빈도 뿐 아니라, 단어의 중요도에 따른 가중치를 함께 고려하는 텍스트 표현 방식이다(Leskovec et al., 2014). TF-IDF는 특정 단어가 한 문서 내에서 얼마나 자주 출현하는지를 나타내는 TF(term frequency)와 해당 단어가 전체 문서 집합에서 얼마나 희귀하게 사용되는지를 나타내는 IDF(inverse document frequency)를 활용하여 텍스트를 표현한다. TF는 BoW 방식과 동일하게 특정 문서 내에서 각 단어의 출현 횟수를 의미하고, IDF는 특정 단어를 포함하는 문서 비율의 역수에 로그를 취한 값[6]으로, 전체 문서 집합 내에서 해당 단어가 포함된 문서의 수가 적을수록 높은 값을 가진다. 이를 바탕으로, TF-IDF는 단어 집합의 크기와 같은 차원을 가지는 벡터에 각 단어의 TF와 IDF를 곱한 값을 할당함으로써 텍스트를 표현한다.

이와 같이 빈도 기반 언어모델은 단어의 출현 빈도를 바탕으로 텍스트를 수치화된 벡터로 표현할 수 있으며, 이해하기 쉽고 구현이 간단하다는 장점을 가진다. 그러나 텍스트 내 단어의 출현 순서를 고려하지 않으며 각 단어에 부여된 인덱스만을 기준으로 텍스트를 표현하므로 단어 간 의미적 유사성을 반영하지 못한다는 한계점이 존재한다. 또한, 텍스트를 표현하는 벡터가 대부분의 위치에서 0의 값을 가지는 희소 벡터(sparse vector)이므로, 단어 집합이 커질수록 학습 데이터의 밀도가 급격히 감소하여 분석 성능이 저하되는 차원의 저주(curse of dimensionality) 문제가 발생할 가능성이 높다.

2. 의미 기반 언어모델

의미 기반 언어모델은 단어의 의미가 주변 단어에 의해 형성된다는 분포 가설(distributional hypothesis)을 근거로 단어가 사용된 맥락을 고려하여 텍스트를 벡터로 표현하는 언어모델이다(Harris, 1954). 주어진 맥락에 적합한 단어를 추론하도록 학습된 인공신경망을 활용하여 단어를 벡터로 표현하기 때문에 추론 기반 기법이라고도 지칭하며, 대표적인 의미 기반 언어모델로는 Word2Vec, GloVe(global vectors for word representation), fastText, Doc2Vec 등이 있다.

Word2Vec은 텍스트 내 등장하는 각 단어와 그 주변 단어가 공통의 맥락을 공유하는 단어들이라 간주하고 이들이 유사한 벡터 값을 가지도록 학습되는 인공신경망 구조의 언어모델로, 단어 간 의미적 유사성을 고려하여 텍스트를 벡터로 표현할 수 있다

의 bi-gram이 단어 집합을 이루게 되고, 이에 따라 “eats mouse”와 “mouse eats”는 같은 두 단어의 조합이지만 서로 다른 벡터로 표현된다.

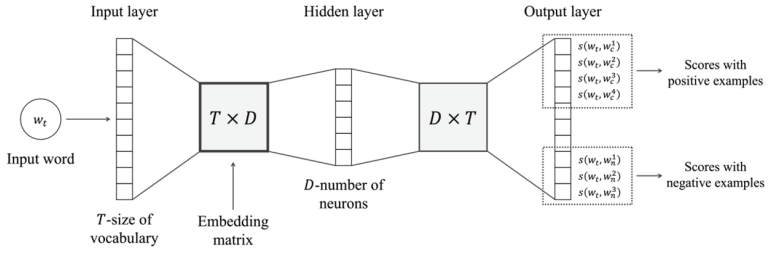
6) $IDF = \log\left(\frac{\text{전체 문서의 수}}{1 + \text{특정 단어가 포함된 문서의 비율}}\right)$

(Mikolov et al., 2013). Word2Vec은 인공신경망 구조 및 학습 방식에 따라 Skip-gram 모델과 CBOW(continuous bag of words) 모델로 구분된다. 두 방식 모두 하나의 은닉층(hidden layer)을 가지는 인공신경망 구조를 바탕으로 하지만, Skip-gram 모델은 특정 단어가 입력되었을 때 주변 단어를 예측하도록 학습되는 반면, CBOW 모델은 주변 단어가 입력되었을 때 중심 단어를 예측하도록 학습된다. 이때 주변 단어는 특정 단어의 앞뒤에 위치하는 단어를 의미하며, 해당 단어가 등장하는 맥락을 나타낸다. 주변 단어의 범위는 Word2Vec의 인공신경망이 학습해야 하는 데이터의 범위에 영향을 미치며, 윈도우 크기(window size)라는 하이퍼 파라미터(hyperparameter)를 통해 조절된다. 두 접근 방식은 입출력이 서로 반대인 것을 제외하면 유사한 방식으로 학습되므로, 본 절에서는 Skip-gram 모델에 대해서만 설명한다.

Skip-gram 모델의 인공신경망 구조는 〈그림 1〉과 같다. Skip-gram 모델의 입력층은 각 단어를 원핫 벡터의 형태로 입력받고, 출력층은 단어 집합의 모든 단어에 대해 입력된 단어의 주변에 위치할 확률을 산출한다. 이때 입력층과 은닉층, 은닉층과 출력층을 연결하는 가중치들이 각각 T × D, D × T 크기의 행렬로 구성되며, T는 단어 집합의 크기, D는 표현하고자 하는 단어 벡터의 차원을 나타낸다. 일반적인 인공신경망에서 은닉층과 출력층 사이에 활성화 함수(activation function)를 배치하여 데이터의 비선형성을 학습에 반영하는 것과 달리 Skip-gram 모델에는 활성화 함수가 존재하지 않으며, 이에 따라 각 층 사이에서 값이 전달되는 과정은 가중치 행렬을 통한 선형변환(linear transformation) 연산과 같다. Skip-gram 모델의 입력층에는 각 단어가 원핫 벡터 형태로 입력되므로 입력층과 은닉층을 연결하는 가중치 행렬 중 입력된 단어에 대응되는 행(row)만 벡터의 형태로 은닉층에 전달된다. 이는 입력 단어에 대한 D개의 가중치 값을 가지는 임베딩 벡터(embedding vector)로서, 모델의 학습 과정에서 해당 단어의 의미를 반영하도록 업데이트되며 학습이 완료된 후 최종 단어 벡터로 사용된다. 이때 임베딩 벡터의 차원은 텍스트 분석의 효율을 고려하여 일반적으로 전체 단어 집합 크기에 비해 상대적으로 적은 크기(예: 300차원)로 설정된다(Pennington et al., 2014). 은닉층을 통과한 임베딩 벡터는 은닉층과 출력층 사이 가중치 행렬을 통해 전체 단어 집합과 동일한 크기의 출력 벡터로 변환되어 전체 단어에 대해 해당 단어가 주변에 위치할 확률을 나타내게 된다. 출력층에서는 입력된 단어가 특정 단어와 같은 맥락을 공유하는지를 판단하는 점수 함수(score function)를 적용하여 최종 확률값을 산출하며, Skip-gram 모델은 이 점수 함수를 최대화하는 것을 목적으로 학습된다. 일반적으로 점수 함수는 두 단어의 임베딩 벡터 간 내적(dot product) 값으로 계산되며, 이에 따라 인공신경망의 학습이 완료되면 같은 맥락에서

자주 함께 출현한 단어들의 임베딩 벡터는 서로 유사한 값을 가지게 된다. 이때 출력층에서 이루어지는 연산을 다중 분류 문제로 간주할 경우, 학습 과정의 매 순간마다 단어 집합의 모든 단어에 대한 확률을 계산하여야 하므로 텍스트 데이터의 크기가 커질수록 학습 효율이 크게 저하된다. 그러므로 대규모 텍스트 데이터를 학습할 때 소요되는 계산 비용을 줄이기 위해, 출력층에서의 연산을 여러 개의 독립적인 이진 분류 문제로 간주하고 입력된 단어의 실제 주변 단어를 제외한 나머지 단어들에 대해서는 무작위로 추출된 일부만 학습에 반영하는 네거티브 샘플링(negative sampling) 기법이 적용된다.

〈그림 1〉 Skip-gram의 인공신경망 구조(Lee et al., 2023)



앞선 설명과 같이 Word2Vec은 인공신경망의 학습을 통해 최적화된 가중치를 바탕으로 단어 벡터를 도출하므로 각 단어 벡터는 모든 위치에서 연속값을 가지는 밀집 벡터(dense vector)이며, 이러한 벡터 표현 방식을 분산 표현(distributed representation)이라 한다. 분산 표현 방식은 단어 집합의 크기가 커질수록 벡터의 차원도 함께 커지는 빈도 기반 언어모델의 희소 벡터에 비해 높은 분석 성능과 효율을 제공한다. 또한 Word2Vec의 학습이 적절히 이루어진 경우, 유사한 의미를 가지는 단어들은 벡터 공간(vector space) 내 유사한 영역에 위치하게 된다. 이와 같은 공간적 관계를 바탕으로 Word2Vec은 단어 벡터 사이의 합(addition) 또는 차(subtraction) 연산을 통해 단어 간 의미적 관계를 나타낼 수 있다.[7]

이러한 장점에도 불구하고, Word2Vec은 몇 가지 한계점을 가지고 있다. 첫 번째로, Word2Vec은 학습 과정에서 정해진 윈도우 크기 내의 주변 단어만을 고려하기 때

7) 예를 들어, Word2Vec이 제공하는 단어 벡터 표현을 활용하면 "$king - man + woman \approx queen$"의 벡터 연산이 성립한다. 이는 Word2Vec이 단어를 벡터로 변환하는 과정에서 단어의 문맥적 의미가 보존된다는 것을 의미한다.

문에 전체 문서 집합 내 단어 간 관계를 충분히 반영하지 못한다. 이를 보완하기 위해, 단어 간 전역적 통계 정보를 활용하여 단어 벡터를 생성하는 GloVe가 제안된 바 있다(Pennington et al., 2014). GloVe는 중심 단어의 벡터와 주변 단어의 벡터의 내적값이 전체 문서 집합에서 해당 단어들의 동시 출현 확률(co-occurrence probability)과 동일한 값을 가지도록 학습된다. GloVe는 이러한 학습 과정을 통해 단어 벡터 간 의미적 유사성을 충분히 반영하면서, 전체 문서 집합에서의 통계 정보 또한 고려하여 텍스트를 벡터로 표현할 수 있다.

두 번째로, Word2Vec은 벡터 도출 시 단어의 형태학적 특성을 고려하지 않으며,[8] 학습하는 단어의 출현 빈도에 따라 학습 효과에 차이가 있다. 문서 집합 내 자주 등장한 단어의 경우 학습이 충분히 이루어져 단어의 의미가 비교적 정확하게 반영되는 반면, 출현 빈도가 적은 단어의 경우 학습이 적게 이루어져 정확한 의미를 포착하지 못한다. 특히, 전체 문서에서 한 번도 출현하지 않은 단어의 벡터를 도출하려면 모델을 재학습시켜야 한다. 이를 보완하기 위해, 각 단어를 여러 개의 하위 단어(sub-word)로 분해하여 단어뿐 아니라 각 하위 단어에 대해서도 모델을 학습시키는 fastText가 제안된 바 있다(Bojanowski et al., 2017).[9] fastText는 이와 같은 학습 방식을 통해 희귀한 단어나 한 번도 출현하지 않은 단어의 의미를 파악할 수 있어 오타가 자주 발생하거나 신조어가 포함된 텍스트 데이터에 대한 분석에도 적용이 가능하다.

마지막으로, Word2Vec은 개별 단어의 벡터 표현을 위한 언어모델이므로 다수의 단어로 구성된 긴 텍스트를 처리하는데 한계가 존재한다. Word2Vec을 활용할 때에도 텍스트에 포함된 모든 단어 벡터의 합이나 평균을 취함으로써 텍스트를 하나의 벡터로 표현할 수 있지만, 포함되는 단어의 수가 증가할수록 산출되는 벡터의 의미가 모호해져서 전체 텍스트의 의미를 정확하게 표현하지 못한다. 이러한 한계를 보완하기 위해 단어와 함께 문서의 벡터도 표현할 수 있도록 설계된 Doc2Vec이 제안되었다(Le and Mikolov, 2014). Doc2Vec은 단어와 주변 단어의 관계와 더불어 단어가 포함된 문단(paragraph)의 정보를 모델 학습에 반영함으로써 단어뿐 아니라 문서도 벡터로 변환할 수 있다. Doc2Vec은 인공신경망 구조 및 학습 방식에 따라 PV-DM(distributed

8) Word2Vec에서는 각 단어에 대한 학습이 개별적으로 이루어지기 때문에, “teach”, “teacher”, “teaching”과 같이 형태학적으로는 유사한 단어들이라도 변환된 벡터 값은 유사하지 않을 수 있다.

9) 예를 들어, fastText는 “hello”라는 단어를 “〈he”, “hel”, “ell”, “llo”, “lo〉”와 같이 여러 개의 하위 단어로 분해하고, 해당 단어뿐 아니라 하위 단어들까지 학습에 반영하여 벡터를 도출한다. 이때 각 단어의 벡터는 해당 단어에 포함되는 모든 하위 단어의 벡터를 합산한 것과 같다.

memory version of paragraph vector) 모델과 PV-DBOW(distributed bag of words version of paragraph vector) 모델로 구분되며, 두 모델은 각각 앞서 설명한 Word2Vec의 CBOW 모델과 Skip-gram 모델과 유사한 특징을 가진다. PV-DM 모델은 특정 문단에 속한 일련의 단어들과 해당 문단의 식별자(identifier)가 입력되었을 때 그 문단에서 다음에 출현할 단어를 예측하도록 학습되는 반면, PV-DBOW 모델은 특정 문단의 식별자가 입력되었을 때 해당 문단에 출현하는 일련의 단어를 예측하는 방식으로 학습된다.

의미 기반 언어모델은 단어와 주변 단어의 관계를 토대로 단어의 의미를 파악하여 텍스트를 표현할 수 있다는 장점이 있으며, 챗봇, 감성 분석, 검색 엔진 등 다양한 분야에 활용되어 왔다. 그러나 의미 기반 언어모델은 단어를 단일 벡터로 표현하기 때문에, 문맥에 따라 여러 의미를 가지는 단어의 다의성(동음이의어)을 고려할 수 없다는 한계점이 있다.[10]

3. 문맥 기반 언어모델

문맥 기반 언어모델은 단어의 문장 내 위치, 주변 단어와의 관계 등을 통해 텍스트의 문맥을 파악하고, 이를 바탕으로 단어에 대한 임베딩 벡터를 동적으로 할당하는 언어모델이다. 문맥 기반 언어모델은 크게 순환신경망(recurrent neural networks: RNN) 기반 언어모델과 트랜스포머 기반 언어모델로 구분된다.

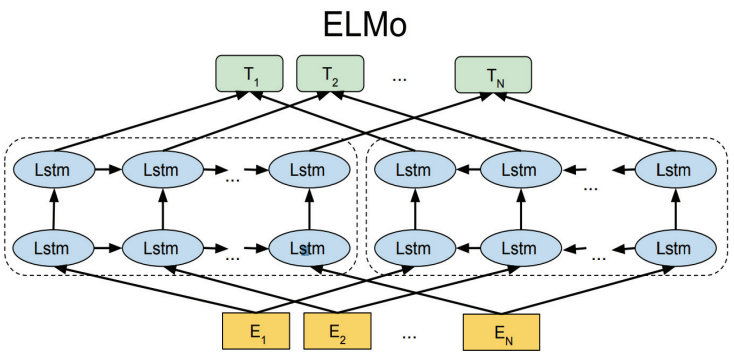
순환신경망은 인공신경망을 재귀적으로 연결하여 여러 시점에 대한 데이터를 순차적으로 입력 받고 출력된 결과를 다음 시점의 입력으로 사용함으로써 순차 데이터(sequential data) 처리에 강점을 가지는 모델이다. 순환신경망은 이러한 재귀적 구조를 통해 특정 시점의 출력값을 산출하는 과정에 이전 시점의 정보를 반영할 수 있으며, 이에 따라 순차 데이터의 시간적 의존성과 비선형적인 추세를 파악하는데 효과적인 것으로 알려져 있다. 하지만 기본적인 순환신경망의 경우 입력 데이터의 길이가 길어질수록 데이터의 초기 부분을 잘 기억하지 못하는 장기 의존성 문제(long-term dependency problem)가 발생하여 성능이 저하된다는 한계를 가지고 있다. 이를 보완하기 위해, 입력된 정보를 선택적으로 전달하는 다수의 게이트(gate)를 순환신경망 구조에 포함시킨 LSTM(long short-term memory)과 GRU(gated recurrent unit)

10) 두 문장 “Apple is red”와 “Apple is a famous company”에서 “Apple”은 첫 번째 문장에서는 사과를, 두 번째 문장에서는 특정 기업을 의미하나, 의미 기반 언어모델은 이를 동일한 하나의 벡터로 변환하기 때문에 의미적 차이를 구분하지 못한다.

등의 파생 모델이 개발되었다(Cho et al., 2014; Hochreiter and Schmidhuber, 1997). 텍스트 데이터는 여러 단어가 순차적으로 연결된 형태를 가지고 있으므로, 순차 데이터의 일종이다. 따라서 텍스트 데이터를 활용하여 순환신경망을 학습시키면, 일련의 문장이 주어졌을 때 이전까지의 문맥을 고려하여 다음 단어를 예측하는 언어모델로 활용할 수 있다. 이러한 순환신경망을 활용한 대표적인 문맥 기반 언어모델로는 ELMo(embeddings from language model)가 있다.

ELMo는 〈그림 2〉와 같이 입력된 문장에 대해 순방향 LSTM(그림의 좌측)과 역방향 LSTM(그림의 우측)을 동시에 학습하는 언어모델이다(Peters et al., 2018). 순방향 LSTM은 문장의 첫 단어부터 현재 단어까지 학습하여 문장 내 현재 단어 이전의 문맥을 반영하고, 역방향 LSTM은 이와 반대 방향으로 학습함으로써 문장 내 현재 단어 이후의 문맥을 반영하여 단어 벡터를 출력한다. 이러한 구조를 바탕으로 ELMo는 주어진 단어가 문장 내에서 쓰인 정확한 의미를 반영하여 단어를 벡터로 표현할 수 있다.[11] 학습이 완료된 ELMo는 두 LSTM이 출력한 단어 벡터를 결합하여 최종 단어 벡터를 도출한다. 하지만 ELMo는 단어의 순차적인 위치에 기반한 문맥을 고려하기 때문에 단어 간의 물리적인 위치가 가까울수록 높은 관계성을 부여하는 경향이 있어, 복잡한 언어적 구조를 가지는 텍스트나 멀리 떨어진 단어 사이의 의미적 관계를 파악하기 어렵다는 한계점이 존재한다.

〈그림 2〉 ELMo의 인공신경망 구조(Devlin et al., 2019)



11) 'He went to the bank and then sat down to fish'에서 'bank'는 강변을 의미하지만, 순방향 LSTM은 'bank' 다음에 오는 'and then sat down to fish'라는 문맥을 반영할 수 없기 때문에 정확한 의미를 파악할 수 없다. 하지만 양방향 LSTM은 단어의 뒤에 오는 문맥을 고려할 수 있어 'bank'가 강변을 의미한다는 것을 유추할 수 있다.

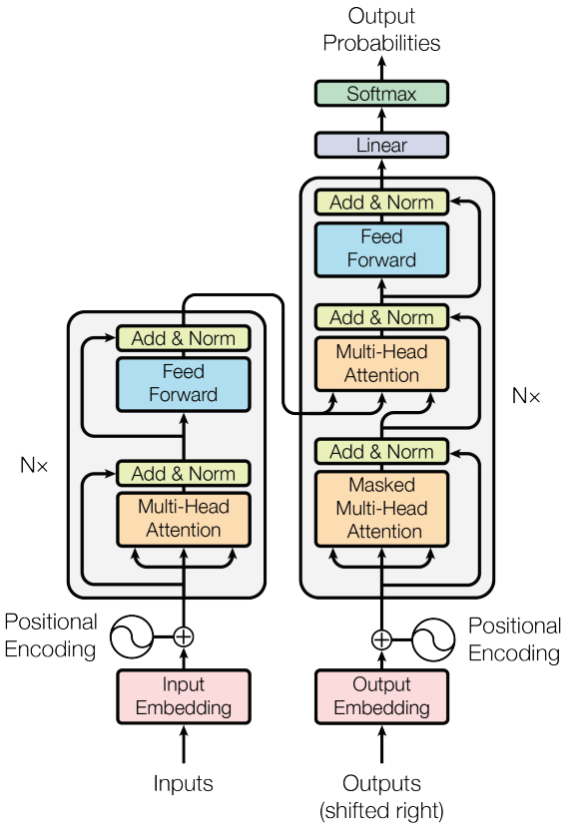
트랜스포머는 순환신경망의 순차적 데이터 처리 방식을 벗어나, 입력된 문장 내 모든 단어 간의 관계를 동시에 고려하는 셀프 어텐션(self-attention) 매커니즘을 기반으로 동작하는 인공신경망이다(Vaswani et al., 2017). 트랜스포머는 〈그림 3〉과 같이, 인코더(encoder)(그림의 좌측)와 디코더(decoder)(그림의 우측)로 구성되며 각각은 여러 개의 인코더 층과 디코더 층이 쌓인 구조이다. 여기서 인코더는 문장을 입력 받아 해당 문장의 의미를 나타내는 잠재 벡터를 출력하는 역할을 수행한다. 인코더에 입력된 문장은 먼저 의미를 가지는 가장 작은 단위인 토큰(token)[12]으로 분리되고, 각 토큰은 각자의 고유한 의미를 나타내는 토큰 임베딩 벡터로 변환된다. 이때 트랜스포머는 순환신경망처럼 데이터를 순차적으로 입력 받는 것이 아니라 문장 전체를 한 번에 입력 받기 때문에, 입력된 문장 내 토큰 간의 순서를 반영하기 위해 각 토큰의 위치 정보를 임베딩 벡터에 포함시키는 위치 인코딩(positional encoding) 과정[13]을 거친다. 각 토큰의 임베딩 벡터는 같은 문장 내 다른 모든 토큰과의 유사도를 나타내는 셀프 어텐션 벡터로 변환되며, 트랜스포머는 이러한 셀프 어텐션 연산을 통해 입력된 문장의 맥락을 반영한다. 이때 여러 번의 셀프 어텐션 연산을 동시에 병렬로 수행하는 멀티 헤드 어텐션(multi-head attention)이 적용되어, 문장의 맥락을 다양한 관점에서 포착할 수 있다. 이후 각 토큰에 대한 셀프 어텐션 벡터는 피드 포워드 신경망(feed-forward neural network)을 통해 학습 목적에 맞게 정제되고, 입력된 정보의 손실을 최소화하기 위한 잔차 연결(residual connection)과 계층 정규화(layer normalization)라는 추가적인 연산을 거쳐 최종 토큰 벡터로 변환된다. 이러한 과정을 통해 생성된 토큰 벡터는 다음 인코더 층의 입력으로 사용되고, 층을 거듭할수록 벡터는 더 정교한 문맥적 정보를 포함하게 된다. 디코더는 인코더가 처리한 텍스트에 대응하는 출력 문장을 순차적으로 생성하는 역할을 수행한다. 디코더의 출력 문장에 대한 토큰 임베딩 벡터는 인코더와 동일하게 위치 인코딩 과정을 통해 형성된다. 그러나 인코더와 달리 디코더의 목적은 텍스트를 생성하는 것이므로, 각 토큰에 대한 어텐션 연산 시 해당 토큰 이후에 등장하는 토큰의 정보를 반영하지 않도록 미래의 토큰 정보를 가리고 연산을 수행하는 마스크드(masked) 멀티 헤드 어텐션이 적용된다. 이후 인코더에 입력된 문장에 대한 토큰 벡터와 디코더의 출력 문장에 대한 토큰 임베딩

12) 토큰은 언어모델의 사용 목적에 따라 상이한 형태로 정의될 수 있다. 주로 어절, 단어, 형태소 등으로 정의된다.

13) 인코더가 입력받는 문장과 같은 길이의 벡터를 구축하여 문장 내 각 위치를 구분하는 고유한 값을 미리 할당하고, 이를 입력된 토큰의 임베딩 벡터에 원소별(element-wise)로 더해주는 과정을 말한다.

벡터를 대상으로 또 다른 멀티 헤드 어텐션을 적용함으로써 인코더를 통해 전달된 문맥적 정보를 반영하여 출력 문장에 대한 최종 토큰 벡터를 도출한다. 마지막으로, 디코더를 통해 도출된 출력 문장의 최종 토큰 벡터는 소프트맥스(softmax)층을 통과함으로써 확률값의 형태로 변환되고, 이를 통해 다음에 등장할 단어에 대한 예측을 수행한다. 이와 같은 학습 과정을 완료한 후, 트랜스포머는 입력된 문장의 문맥을 파악하여 그에 대응하는 적절한 문장을 새롭게 생성할 수 있는 언어모델로 활용될 수 있다.

〈그림 3〉 트랜스포머의 인공신경망 구조(Vaswani et al., 2017)



트랜스포머의 이러한 장점을 토대로 개발된 대표적 언어모델로는 BERT와 GPT가 있다. BERT는 트랜스포머의 인코더만 여러 층 쌓은 구조로 설계된 언어모델로 텍스트

의 언어적 특성을 이해하기 위해 문장의 일부 토큰을 가리고 모델이 이를 예측하도록 하는 마스크드 언어 모델링(masked language modeling)과 더 넓은 범위의 문맥과 문장 간의 논리적 연속성을 학습하기 위해 주어진 두 문장이 서로 이어지는 문장인지 판별하는 다음 문장 예측(next sentence prediction)의 방식으로 학습된다(Devlin et al., 2019). 반면 GPT는 트랜스포머의 디코더만 여러 층 쌓은 구조로 설계된 언어모델로 텍스트 생성 작업에 강점을 가지고 있으며, 주어진 문장 내 토큰들의 문맥을 바탕으로 다음에 등장할 문장을 예측하는 방식으로 학습된다(Radford et al., 2018).

BERT와 GPT는 위키피디아, 책, 웹 텍스트 등 대규모의 텍스트 데이터를 활용하여 모델이 자연어로 이루어진 텍스트를 포괄적으로 이해할 수 있도록 학습되었으며, 이 과정을 사전 학습(pre-training)이라 한다. 사전 학습이 완료된 언어모델은 각 토큰에 포함된 복잡한 문맥 정보를 반영하여 텍스트를 벡터로 표현할 수 있으며, 사전 학습만으로도 다양한 텍스트 분석 작업에서 일정 수준 이상의 성능을 보인다. 이에 더하여 특정한 자연어 처리 작업에 대한 성능을 개선하기 위해 해당 작업의 목적에 맞는 데이터를 활용하여 사전 학습이 완료된 언어모델에 추가적인 학습을 수행할 수 있으며, 이 과정을 미세 조정이라 한다.

일반적으로 미세 조정은 해당 작업에 대한 정답이 존재하는 데이터를 활용하여 특화하려는 작업(예: 텍스트 분류)을 위한 추가적인 인공신경망(예: 다중 퍼셉트론 분류기)을 학습함으로써 수행된다. 미세 조정을 통해 모델은 주어진 작업에 최적화되며, 특히 사전 학습 과정에서 얻은 언어적 이해 능력을 바탕으로 질의응답, 감정 분석, 텍스트 분류 등의 다양한 자연어 처리 작업에서 뛰어난 성능을 발휘할 수 있다.

이와 같은 트랜스포머 기반 언어모델들은 공통적으로 모델 구조가 매우 크고 복잡하기 때문에 직접 학습을 수행하기 위해서는 많은 시간과 컴퓨팅 자원이 필요하다. 따라서 대규모 텍스트 데이터를 통해 사전 학습된 ELMo, BERT, GPT 모델을 활용하여 자신이 원하는 목적에 맞게 미세 조정하는 방식을 택하는 것이 효율적일 수 있다. 특히 트랜스포머 기반 언어모델의 경우 Python 라이브러리인 HuggingFace[14]에서 모델을 직접 다운로드하지 않고도 입력한 텍스트에 대한 출력을 호출받을 수 있는 API(application programming interface)를 제공하고 있어, 개인 연구자나 개발자도 비교적 적은 비용으로 활용이 가능하다.

14) https://huggingface.co/

Ⅲ. 언어모델의 최근 연구 동향

본 장에서는 언어모델의 최신 연구 동향에 대해 살펴본다. 특히, 최근 여러 산업에서 큰 영향을 주고 있는 초거대 언어모델(large language model: LLM)을 중심으로 학술적 등장 배경을 설명하고 최근 학습 방법을 소개한다. 또한, 초거대 언어모델의 활용 전략을 크게 3가지로 나누어 설명한다.

1. 언어모델의 거대화와 창발 능력

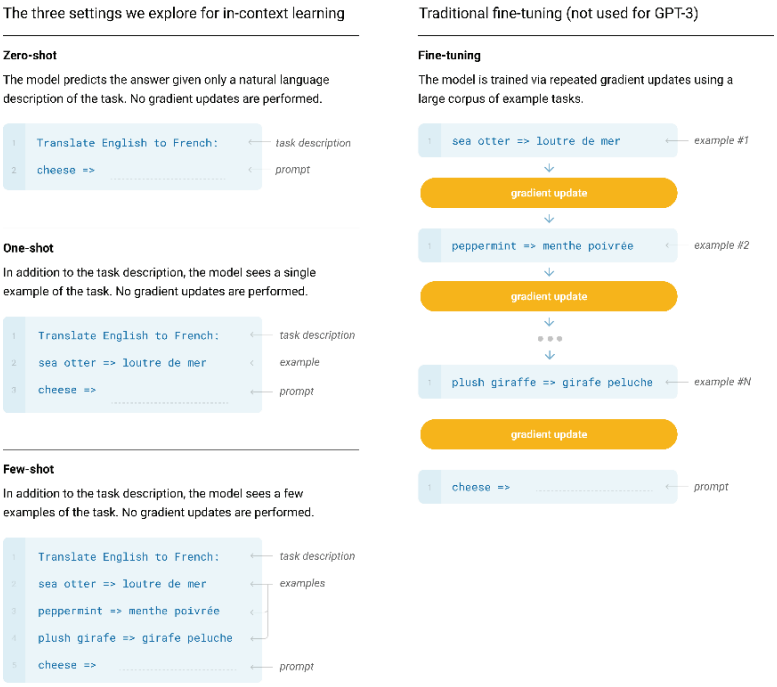
트랜스포머 및 어텐션 메커니즘이 제안된 이후로, 많은 연구에서 언어모델의 매개변수(parameter) 수와 학습 데이터의 크기가 증가함에 따라 모델 성능이 향상된다(scaling law)는 결과를 보고한 바 있다(Brown et al., 2020; Kaplan et al., 2020; Raffel et al., 2020). 이에 따라 GPT-3(1750억개 매개변수), LaMDA(1600억개 매개변수), PaLM(5400억개 매개변수) 등 방대한 양의 모델 매개변수와 데이터로 학습한 초거대 언어모델이 등장하였다. 〈표 1〉은 현재까지 공개된 주요 초거대 언어모델을 정리한 표이다.

〈표 1〉 주요 초거대 언어모델

모델명	기업	모델 크기(매개변수 수)	공개일자
GPT-3	OpenAI	1750억	2020년 5월
LaMDA	Google	1370억	2021년 7월
Chinchilla	DeepMind	700억	2022년 3월
PaLM	Google	5400억	2022년 4월
Galactica	Meta	1200억	2022년 11월
ChatGPT	OpenAI	1750억	2022년 11월
LLaMa	Meta	650억	2023년 2월
Bard	Google	미공개	2023년 2월
GPT-4	OpenAI	미공개	2023년 3월
PaLM2	Google	5400억	2023년 5월
LLaMa2	Meta	700억	2023년 7월
Mistral-7B	Mistral AI	70억	2023년 9월
Solar-0-70b	Upstage	107억	2023년 9월
Gemini	Google	미공개	2023년 12월

흥미롭게도, 비교적 작은 규모의 기존 언어모델과 달리, 초거대 언어모델은 텍스트 형태의 명령(프롬프트)만으로 주어진 문제를 해결하는 이른바 인-컨텍스트 학습(in-context learning) 능력을 보인다. 기존 자연어 처리 기법의 경우 특정 작업을 위해 많은 양의 학습 데이터를 심층학습 모델에 훈련시킬 필요가 있었다. 그러나 초거대 언어모델은 인-컨텍스트 학습이 가능하므로 모델 파라미터에 내재된 지식을 바탕으로 주어진 프롬프트만으로도 문맥을 이해하고 문제를 해결할 수 있다. 〈그림 4〉는 그 예시로써, 초거대 언어모델은 학습 데이터 없이도 주어진 문제가 번역 문제임을 알려주거나 몇 개의 사례를 제공하는 것만으로도 번역 작업을 수행할 수 있음을 보여준다.

〈그림 4〉 초거대 언어모델의 인-컨텍스트 학습 예시(Brown et al., 2020)



이처럼 비교적 작은 크기의 기존 언어모델에서는 나타나지 않았으나 초거대 언어모델에서 나타난 능력을 창발 능력(emergent ability)이라고 한다(Wei et al., 2022). 초거대 언어모델의 창발 능력을 최대화하기 위해서는 모델의 입력을 전략적으로 설계

하는 작업이 필요하며 이를 프롬프트 엔지니어링이라고 일컫는다(Zhou et al., 2023). 예를 들면, 사용자가 풀고자 하는 문제의 이상적인 입력과 출력 예시 복수 개를 프롬프트에 제공하는 퓨샷 러닝(few-shot learning) 방식이나, 분류값이나 제약 조건 등의 구체적인 내용을 프롬프트에 명시하는 방식을 통해 언어모델의 문맥 이해를 높이고 성능을 향상시킬 수 있다. 최근에는 주어진 입력에 따라 어떤 예시를 프롬프트에 제공하는 게 성능 향상에 도움이 되는지를 분석한 연구(Wang, Shuhe et al., 2023)와, 초거대 언어모델을 통해 이와 같은 탐색 과정을 자동화하는 연구(Yang et al., 2023)도 수행된 바 있다.

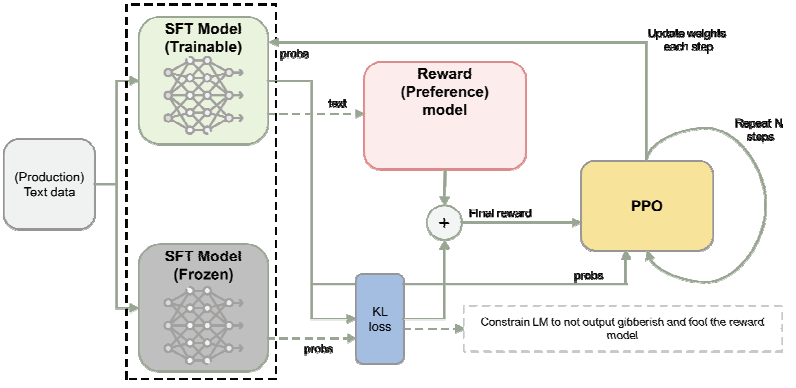
2. 언어모델의 강화학습

초거대 언어모델이 서비스로 제공된 이후로, 사용자들에 의해 많은 문제점이 보고된 바 있다. 특히 프롬프트에 명시된 사용자의 의도와 다르게 단순히 프롬프트와 유사한 문장을 반복해서 생성하거나, 거짓 또는 해로운 내용, 전혀 관련이 없는 내용을 출력하는 문제가 다수 발생하였다. 사용자의 의도에 더욱 부합하는 서비스를 제공하기 위해, 초거대 언어모델에 강화학습(reinforcement learning)을 도입하여 위와 같은 문제를 해결하려는 노력이 이루어졌으며, 이를 적용한 대표적인 모델로는 GPT-3를 기반으로 하는 InstructGPT와 GPT-3.5를 기반으로 하는 ChatGPT가 있다(Ouyang et al., 2022). 강화학습은 주어진 환경(environment)에서 주체(agent)가 현재의 상태(state)를 기준으로 선택할 수 있는 행동(action) 중 최대의 보상(reward)을 가져다주는 행동을 선택하도록 학습하는 것을 의미한다. 초거대 언어모델을 위한 강화학습 방식으로는 RLHF(reinforcement learning from human feedback) 방식이 있으며, RLHF 방식에서 주체는 기존의 초거대 언어모델, 환경은 사용자의 입력, 행동은 모델의 답변 생성이고, 보상은 사용자의 답변에 대한 평가 점수를 기반으로 생성되는 기대 점수이다.

RLHF 방식의 강화학습은 〈그림 5〉와 같이 크게 3단계로 구분하여 수행된다. 첫 번째 단계로, 사전 학습된 초거대 언어모델을 사용자가 선호하는 방향으로 미세 조정하기 위해 프롬프트와 그에 따른 이상적인 답변으로 이루어진 고품질의 데모 데이터 셋(약 12000~15000여건)을 구축한다. 이때 프롬프트는 정답을 매기는 라벨러(labeller)나 개발자가 준비한 프롬프트와 OpenAI의 API로 제출된 실제 사용자의 프롬프트 집합으로부터 표본 추출하여 구성되며, 답변은 프롬프트를 토대로 라벨러가 직접 작성한 이상적인 출력을 사용한다. 이와 같은 데모 데이터 셋을 활용하여 기존의

초거대 언어모델을 미세 조정함으로써 SFT(supervised fine-tuning) 모델을 개발한다. 두 번째로, 앞서 개발한 SFT 모델에 사전에 준비한 질문을 입력하여 복수 개의 출력(4~9개)을 생성하고, 라벨러가 직접 답변에 대한 순위를 매긴 후, 출력과 순위를 기반으로 라벨러의 선호도를 모방하는 보상 모델(reward model)을 개발한다. 마지막으로, 보상 모델을 통해 SFT 언어모델이 제시하는 출력의 품질을 평가하고, 그 결과를 기반으로 사람이 선호하는 답변을 생성하도록 언어모델을 강화한다.

〈그림 5〉 초거대 언어모델의 강화학습 과정[15)]



강화학습이 적용된 초거대 언어모델은 일반적인 초거대 언어모델에 비해 답변의 해로움, 진실성, 고객 응대성 등을 평가하는 벤치마크 데이터 셋에 대해 우수한 성능을 보인 바 있다. 그러나 모델에 입력된 사용자의 설명에 의존하여 답변이 생성되므로 사용자의 의도에 따라 편향되거나 폭력성이 포함된 결과를 산출할 수 있으며, 학습 데이터의 한계로 인해 인종, 성별, 지역, 언어 등의 측면에서 다양한 사용자의 선호를 충분히 만족시키기 어렵다는 한계가 존재한다. 이러한 한계를 보완하기 위해, 프롬프트의 전제가 잘못되었거나 유해한 내용이 작성될 경우 답변을 하지 않도록 학습시키거나 지속적인 사용자 피드백을 통해 선호도의 편향을 개선하는 등 다양한 노력들이 이어지고 있다.

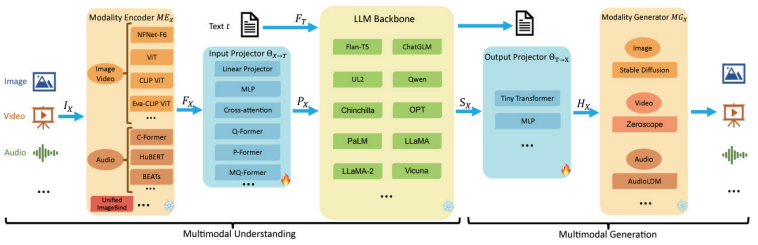
15) Ouyang et al. (2022)을 바탕으로 도식화.

3. 언어모델의 멀티모달 학습

최근 초거대 언어모델들은 텍스트와 더불어 오디오, 비디오, 이미지 등 다른 형태(modality)의 데이터를 함께 활용하여 보다 복잡한 지식을 학습하는 초거대 멀티모달 모델로 발전하고 있다. 이에 따라 OpenAI의 GPT-4(Achiam et al., 2024)와, 구글의 Gemini(Anil et al., 2024), 그리고 Apple의 Ferret(You et al., 2023) 등과 같이 이질적(heterogeneous) 형태의 데이터를 입력 및 출력으로 자유롭게 다룰 수 있는 초거대 모델이 등장하고 있다. 이와 같은 초거대 멀티모달 모델들은 텍스트에 국한되지 않은 다양한 형태의 프롬프트를 입력받음으로써 차트 이미지의 해석, 글과 그림이 모두 포함된 물리학 문제 해결, 이미지에 대한 자동 캡션 생성, 비디오에 대한 질의응답, 자동 통역, 음성 인식 등 대부분의 분야에 활용될 수 있다.

초거대 멀티모달 모델은 이미지-텍스트, 비디오-텍스트, 오디오-텍스트와 같이, 텍스트와 다른 모달리티의 쌍으로 이루어진 데이터 셋을 활용하여 같은 대상을 표현하는 다른 형태의 데이터를 동일한 벡터 공간에 표현하고 그 관계를 학습한다(Zhang et al., 2024). 초거대 멀티모달 모델은 일반적으로 〈그림 6〉과 같이 모달리티 인코더(modality encoder), 입력 프로젝터(input projector), 백본 초거대 언어모델(LLM Backbone), 출력 프로젝터(output projector), 그리고 모달리티 생성기(modality generator)라는 다섯 가지 주요 구성 요소를 포함한다. 이와 같은 요소들을 통해 모델은 텍스트, 이미지, 오디오 등 다양한 모달리티에 대한 정보를 처리하고 생성할 수 있다.

〈그림 6〉 초거대 멀티모달 모델의 일반적인 구조 도식화(Zhang et al., 2024)



구체적으로 각 요소의 역할은 다음과 같다. (1) 모달리티 인코더는 다양한 모달리티의 입력(I_X)을 인코딩하여 특성(F_X)을 추출하는 역할을 한다. 각 입력의 모달리티에 따라 사전 학습된 인코더 모델이 존재한다. 예를 들어, 이미지 데이터에 대해서는 ViT,

NFNet-F6, CLIP ViT와 같이 컴퓨터 비전 모델들이(Brock et al., 2021; Dosovitskiy et al., 2020; Radford et al., 2023), 오디오 데이터에 대해서는 C-Former, HuBERT, BEATs 등의 모델이 주로 인코더로 사용된다(Chen et al., 2022; Chen et al., 2023; Hsu et al., 2021). (2) 입력 프로젝터($\Theta_{X\to T}$)는 모달리티 인코더에서 추출한 특징 벡터(F_X)를 텍스트의 특징 공간(T)에 투영해서 프롬프트(P_X)를 생성하는 역할이다. 즉, 모달리티 입력(I_X)과 이를 설명하는 텍스트(t) 데이터가 있는 상태에서, 텍스트에 대한 특징 벡터(F_T)와 유사한 프롬프트(P_X)를 생성하도록 학습한다. 입력 프로젝터는 간단하게 선형변환이나 다중 퍼셉트론으로 구현할 수도 있고, Cross-attention, Q-Former, P-Former 등의 복잡한 모델을 이용하여 모달리티 간의 연계성을 높일 수 있다(Alayrac et al., 2022; Jian et al., 2024; Li et al., 2023). (3) 백본 초거대 언어모델은 다양한 모달리티로부터 입력 데이터의 의미를 이해 및 추론하고 이에 대한 결정을 내리는 역할을 한다. 백본 모델은 입력된 모달리티의 정보를 기반으로 생성한 텍스트를 직접 출력할 뿐만 아니라, 멀티모달 출력을 생성할지 여부와 생성할 출력의 종류를 지정하는 신호 토큰(S_X)을 생성한다. 여기서, Chinchilla, LLaMa, Vicuna 등의 초거대 언어모델이 백본 모델로 활용될 수 있다(Chiang et al., 2023; Hoffmann et al., 2022; Touvron et al., 2023). (4) 출력 프로젝터($\Theta_{T\to X}$)는 입력 프로젝터와 반대로, 백본 모델의 출력으로 생성된 신호 토큰(S_X)에 따라 특정 모달리티로 출력을 표현하는 역할을 한다. 모달리티 입력(I_X)과 이를 설명하는 텍스트(t) 데이터 쌍이 있는 데이터 셋을 기준으로, 백본 모델의 텍스트 출력을 다음 단계인 모달리티 생성기가 이해할 수 있는 형태(H_X)로 맵핑한다. 이때, 출력 프로젝터는 작은 트랜스포머(tiny transformer) 또는 MLP(multi-layer perceptron)을 주로 이용한다. (5) 모달리티 생성기는 다양한 모달리티로 출력을 생성하는 역할로, 사전에 준비된 생성모델을 주로 활용한다. 예를 들어, 이미지 합성에는 Stable Diffusion(Rombach et al., 2022), 비디오 합성에는 Zeroscope(Wang, Jiuniu et al., 2023), 오디오 합성에는 AudioLDM-2(Liu et al., 2023) 모델이 활용될 수 있다.

4. 언어모델의 경량화

기업들이 방대한 양의 데이터와 매개변수를 통해 초거대 언어모델들을 경쟁적으로 개발할 때, 비교적 작은 크기지만 비슷한 성능을 내는 언어모델들이 공개되어 각광을 받고 있다. 여기에는 몇 가지 배경이 있는데 첫 번째는 scaling law에 대한 인식 변화이다. 모델 크기를 기업들이 경쟁적으로 높이던 시기에, 딥 마인드는 가장 큰 모델보

다는 가장 많은 데이터로 학습한 모델이 더 높은 성능을 보인다고 보고했다(Hoffmann et al., 2022). 즉, 비교적 작은 크기이더라도 오랜 기간 더 많은 데이터를 훈련시킨 모델이 모델 크기는 크지만 상대적으로 적은 양의 학습 데이터를 훈련시킨 모델보다 높은 성능을 보인다는 것이다. 예를 들어, Chinchilla 모델(700억개 매개변수)은 Gopher(2800억개 매개변수), GPT-3(1750억개 매개변수), Megatron(5300억개 매개변수)과 같은 다른 큰 모델들에 비해 작지만 상대적으로 더 많은 데이터를 학습한 모델로, 다른 모델과 비교하여 미세 조정이나 하위 작업에서 더 낮은 연산량과 더 높은 성능을 보여주었다. 이러한 연구 결과는 초거대 언어모델을 개발할 때 모델의 크기보다는 추론 비용에 대한 고려가 필요하다는 사실을 강조하며, 모델 경량화에 대한 관심을 높였다(Touvron et al., 2023).

두 번째는 자유롭게 미세 조정이 가능한 초거대 언어모델의 필요성이다. OpenAI나 Google에 의해 제안된 초거대 언어모델들은 압도적인 성능을 보였으나, 낮은 접근성과 소유주의와 같은 한계가 존재했다(Bahrini et al., 2023; Chowdhery et al., 2023). 초거대 언어모델을 개발한 대부분의 기업은 모델 사용을 유료화하였고, 이에 따라 내부 구조나 학습 과정에 대한 상세 정보를 제공하지 않아 미세 조정을 수행하더라도 개발된 모델에 대한 내부 접근이 불가했다. 따라서 연구자의 연구 목적에 맞는 자유로운 활용에 어려움이 있었으며, 기업이나 연구원의 입장에서는 사내 정보 및 연구 자료의 사용에 따른 보안 문제 등이 발생했다. 이에 따라, 자체적으로 미세 조정이 가능한 초거대 언어모델의 필요성이 대두되었다.

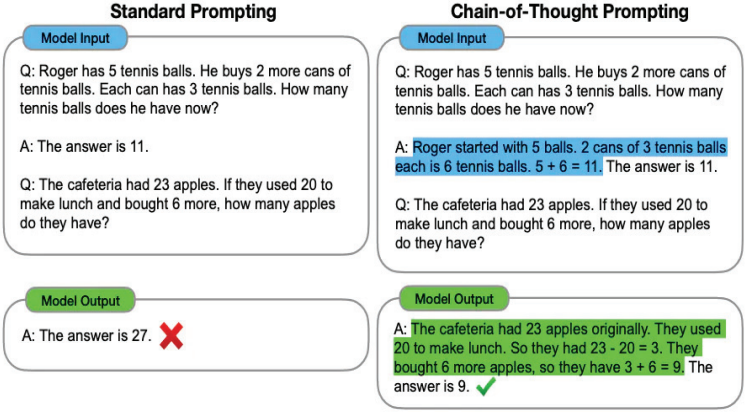
이러한 배경에 따라 성공적인 경량화를 이루어낸 초거대 언어모델의 대표적인 예시로는 Meta가 공개한 LLaMA 시리즈 모델이 있다(Touvron et al., 2023). 해당 모델은 GPT-3(1750억개 매개변수)보다 훨씬 적은 수의 매개변수(70억, 130억, 330억, 650억개)를 가졌지만, 모든 벤치마크 데이터 셋에서 GPT-3를 능가하는 성능을 보였다. LLaMA는 학습 데이터의 품질 개선과 더불어 다양한 기술적 노력을 바탕으로 성공적인 경량화를 이루어냈다. LLaMA는 학습 과정에서 다른 초거대 언어모델의 학습에 사용된 데이터 셋을 재사용하되, 공개적으로 사용할 수 있는 데이터 셋만을 선별적으로 활용하였으며, 대화, 책, 논문, 코드 등 다양한 유형의 데이터를 활용하여 학습 데이터의 다양성을 높이기 위해 노력했다. 또한, LLaMA는 개발 당시 최고 성능을 보이는 초거대 언어모델들이 활용한 다양한 기술(예: prenormalization, rotary embedding, SwiGLU)들을 적극적으로 모델 아키텍처에 적용하였다. 이에 따라 LLaMA는, 앞서 언급한 Chinchilla 모델 연구 결과와 유사하게, 다양하고 많은 데이터 셋을 작은 모델로 학습하는 것이 큰 매개변수를 가진 기존의 초거대 언어모델보다

성능이 높다는 것을 다시 한번 보여주었다. 특히 Meta는 LLaMA의 모델과 내부 매개변수를 공개함으로써 여러 후속 모델의 발전을 비롯하여 학술적으로도 많은 기여를 하였다(Chiang et al., 2023; Taori et al., 2023).

5. 언어모델의 활용 전략

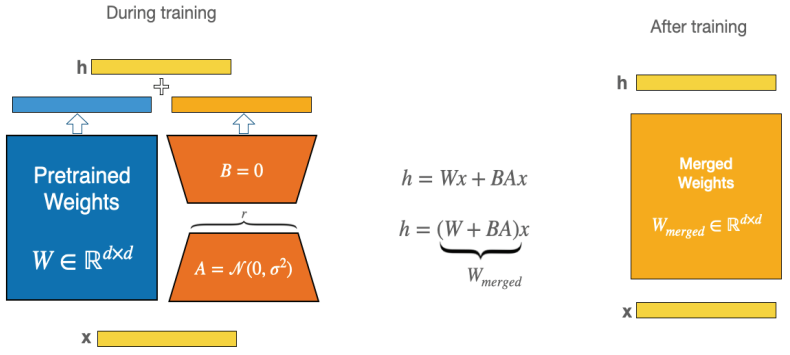
초거대 언어모델의 기존 프롬프트 엔지니어링 방식은 단순한 텍스트 생성에는 효과적이나, 지식이 필요한 추론 또는 산술 작업에 있어서는 낮은 성능을 보였다. 이와 같은 한계를 해결하기 위해, 단계별로 추론하는 과정을 프롬프트에서 설명하는 CoT (chain-of-thought)와 같은 프롬프트 엔지니어링 방법이 제시되었다(Wei et al., 2023). 이는 산술 문제와 같은 복잡한 작업을 해결하기 위해, 입력과 출력 사이의 과정을 연쇄적 단계로 분해하여 모델에게 제공하는 방식이다. 예를 들어, 〈그림 7〉은 예시 문제와 정답 쌍을 제공하는 기존의 프롬프트 엔지니어링 방식(그림의 좌측)에서는 정답을 맞히지 못하지만, 최종 답변이 나오기까지의 풀이 과정을 프롬프트에 제공한 경우(그림의 우측)에는 새로운 산술문제에 대한 정답을 맞히는 것을 보여준다. 이와 같은 CoT 프롬프트 방식은 기본 프롬프트 방식과 함께 초거대 언어모델을 평가하기 위한 벤치마킹 데이터 셋에서 기본 전략으로 활용되고 있으며, 복수 개의 추론 결과를 산출한 뒤 자기 일관성이 높은 답변을 고르는 방식(Wang, Xuezhi et al., 2023)이나 트리 구조로 연쇄 추론을 수행하는 방식으로 발전하기도 했다(Yao et al., 2024).

〈그림 7〉 CoT 프롬프트와 기본프롬프트의 산술 추론능력 차이(Wei et al., 2023)



초거대 언어모델을 특정 목적에 따라 미세 조정하는 것은 컴퓨팅 소스 및 데이터의 양을 고려할 때 비효율적이고 어려운 작업이다. 이에 따라, 미세 조정을 위해 비효율적으로 수많은 모델 매개변수를 조정하기보다는 성능을 유지하면서 학습 가능한 매개변수의 수를 줄이는 효율적인 학습 방법으로 PEFT(parameter efficient fine-tuning) 방식이 제안된 바 있다(Houlsby et al., 2019). PEFT 방식을 적용한 대표적인 모델은 HuggingFace에서 공개한 LoRA(low rank adaptation) 모델이 있다(Hu et al., 2021). LoRA의 학습은 〈그림 8〉과 같이, 초거대 언어모델의 가중치 행렬을 두 개의 작은 행렬(low-rank)로 분해하여 미세 조정하는 방식으로 이루어진다. LoRA는 기존의 사전 학습된 가중치 행렬 W는 동결한 채로, 작은 행렬로 분해된 A와 B만 미세 조정 작업에 맞춰 학습한다. 이때 계수(rank)는 행렬의 행과 열로 생성할 수 있는 공간의 차원을 의미하며, 행렬 A와 B는 W보다 작은 계수를 가지므로 A×B는 W의 근사치이다.

〈그림 8〉 LoRA 모델 메커니즘(Hu et al., 2021)

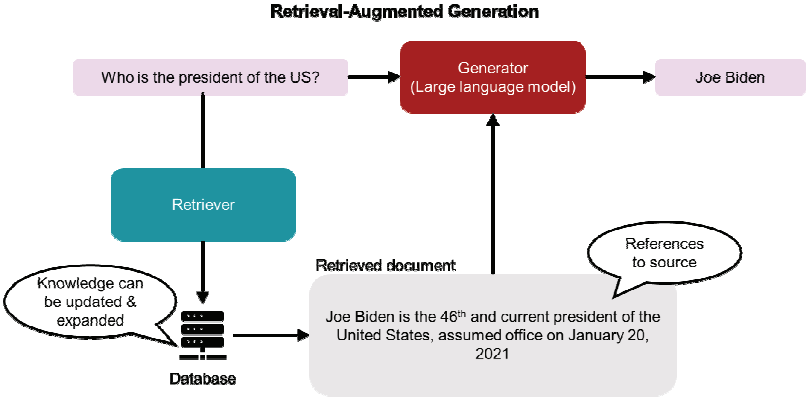


LoRA의 가중치 행렬의 크기는 〈그림 8〉에서 r이라는 변수로 나타난다. 기존 가중치 행렬이 500*500의 크기를 가지고 있으면, 매개변수는 250,000이다. 이를 r = 4로 설정하고 분해하면 A는 500*4, B는 4*500이므로 총 매개변수의 수는 기존 매개변수의 1.6%인 4,000이 된다. 즉, LoRA는 특정 작업에 맞춰 조정된 언어모델을 저장할 때 비교적 작은 크기의 가중치 행렬인 A, B를 저장하므로, 새로운 작업도 비교적 효율적으로 수행할 수 있으며 가중치 업데이트에 필요한 메모리 및 스토리지 비용을 절감할 수 있다. Hu et al. (2021)은 전체 매개변수를 미세 조정한 GPT-2 모델과 0.1%의 매개변수를 미세 조정한 LoRA가 동일하거나 비슷한 수준의 성능을 보임을 밝혔다.

그 외에도, 기존의 초거대 언어모델 구조에 어댑터(adaptor)라는 작은 인공 신경망 모듈을 추가하는 방식이나(Hu et al., 2023), 입력 시퀀스에 작업을 나타내는 접두사(prefix) 또는 학습 가능한 프롬프트를 추가함으로써 특정 작업에 최적화된 미세 조정을 수행하는 방식(Lester et al., 2021; Li and Liang, 2021), 또는 Nvidia에서 제안한 IA3 모델과 양자화(quantization) 방식 등 다양한 연구가 이루어진 바 있다. PEFT 방식은 초거대 언어모델을 미세 조정 하기 위한 자원이 부족한 개인 연구자나 기업에게 적절한 전략이다. 그러나 이와 같은 시도는 어디까지나 효율을 고려한 시도임으로 충분한 데이터 셋이 없을 경우 성능이 낮을 수 있고, 기존의 초거대 언어모델이 가지고 있던 지식이나 능력이 감퇴하는 현상(catastrophic forgetting)이 발생할 수 있다.

한편, 기존 초거대 언어모델의 학습 데이터에 포함되지 않을 가능성이 높은 의료, 법률 등에 대한 전문적인 지식이나 사내 기밀 정보 혹은 최신 정보가 필요한 경우, 초거대 언어모델의 활용 가치는 하락할 수 있다. 이에 대한 대안으로, 〈그림 9〉와 같이 사용자의 질문에 답할 때 필요한 정보를 외부 데이터베이스에서 제공하는 방식, 즉 초거대 언어모델과 정보 검색 알고리즘이 결합된 형태의 검색증강생성(retrieval-augmented generation: RAG) 방식이 제안된 바 있다(Gao et al., 2024). 이와 같은 접근 방식은 추가 학습을 수행할 만큼 많은 데이터가 존재하지 않고 외부 지식이 많이 필요한 문제를 풀 때 효율적이며, 사내 데이터나 자료를 직접 초거대 언어모델에 학습 데이터

〈그림 9〉 검색증강생성 시스템 예시[16)]



16) Lewis et al. (2021)을 바탕으로 도식화.

로 제공하지 않기 때문에 정보 보안에도 유리하다. 또한, 데이터베이스를 업데이트함에 따라, 최신 정보를 프롬프트에 반영할 수 있어서 학습 시점의 정보만 가진 초거대 언어모델보다 유연하다. 검색증강생성은 사용자 질문에 적합한 정보를 검색하는 방법뿐만 아니라, 주어진 쿼리가 데이터베이스에서 답을 찾을 수 있는지 판단하는 방식으로 발전하고 있다.

요약하자면, 사용자는 초거대 언어모델을 활용할 때 풀고자 하는 문제의 특수성과 활용 가능한 자원을 고려하여 적절한 활용 전략을 선택해야 한다. 외부 지식이 크게 필요 없는 문제일 경우 ChatGPT나 GPT-4와 같이 매개변수가 700억 개 이상인 모델을 활용하여 프롬프트 엔지니어링을 먼저 수행할 수 있다. 이때 예제 없이 작업을 설명하거나 예제를 추가하는 프롬프트 전략을 활용하거나 나아가 CoT 등의 전략을 도입할 수 있다. 또한, 프롬프트 엔지니어링만으로 충분한 성능이 확보되지 않는 경우에는 PEFT와 같은 자원 효율적인 미세 조정 방식을 대안으로 도입하여 작업에 특화된 언어모델을 개발할 수 있다. 마지막으로, 외부 지식이 많이 요구되는 작업을 수행할 경우에는 검색증강생성 방식을 도입하여 효율성을 높일 수 있다.

Ⅳ. 언어모델의 공공부문 활용 사례

본 장에서는 〈표 2〉에 정리된 것과 같이 과학기술 분야와 관련한 다양한 문제를 해결하기 위해 앞서 기술한 언어모델의 주요 접근법들을 활용한 사례를 소개한다. 최근 언어모델이 다양한 자연어 처리 분석에 대해 강력한 성능과 범용성을 증명함에 따라 공공부문의 전 영역에서 언어모델의 도입 및 활용에 대한 활발한 논의가 이루어지고 있으나, 아직 기술적·정책적으로 초기 단계에 머물러 있으며 실질적인 활용 사례 또한 과학기술 분야와 같은 일부 영역에 국한되어 있는 실정이다. 이에 따라 본 장에서는 사용한 언어모델의 기술적 가치 및 현업 적용 가능성에 대한 체계적인 검증을 통해 공공부문에서의 활용이 가능할 것으로 판단되는 사례들을 선별하여 소개하며, 각 사례는 (1) 공공부문에서 적용 가능한 시스템을 개발한 사례와 (2) 공공부문의 기존 유사 시스템 및 플랫폼 고도화에 활용할 수 있는 사례로 나누어 소개한다.

〈표 2〉 과학기술 분야의 언어모델 활용 사례 요약

목적	데이터	언어모델 및 접근법	결과	활용가능기관
기술 아이디어 평가(Hong et al., 2022)	특허 초록 정보	Word2vec, Text CNN을 통한 인용 예측	기술 아이디어 평가 결과 (고인용/저인용 예상)	기술보증기금
기술 수요자와 공급자의 중개 (Lee et al., 2023)	논문, 특허, 저서 등의 연구자 성과와 기업의 기술 수요 정보	fastText, 유사도 분석	기술이전 발생 가능성이 높은 연구자-기업 쌍	한국산업기술진흥원
도메인 특화 연구 동향 분석(Choi and Lee, 2024)	특정 도메인의 과학기술 문헌 텍스트 및 메타 정보	Doc2vec, 밀도 기반 클러스터링	도메인 내 연구 세부 토픽 및 기관/국가별 동향	한국과학기술연구원
소비자 불만 이슈 탐지(Lee et al., 2021)	온라인 제품 리뷰	Sentence-BERT, 군집 분석과 감성 분석	소비자 불만 이슈의 시급성 및 심각성 점수	한국과학기술정보연구원
지정상품간 유사도 측정 (한국지식재산연구원, 2023)	지정상품의 상품해설서	KoBERT, KoGPT, KR-SBERT 등 한국어 언어모델, 코사인 유사도	지정상품간 상품해설서 항목 수준의 유사도 및 상품 수준의 유사도	특허청

1. 의미 기반 언어모델과 합성곱 신경망을 활용한 기술 아이디어 평가

기술 아이디어 스크리닝(screening)은 기업이 가치 없는 아이디어에 투자하거나 반대로 가치 있는 아이디어를 놓치는 일을 방지하기 위한 기술 검토 작업으로, 초기 기술 개발의 핵심 단계이다. 전통적으로 기술 아이디어 스크리닝은 전문가 중심의 방법에 의존하였으나, 다수의 아이디어를 평가하는 데 상당한 시간과 비용이 소요되므로 효율적이지 않았다. 또한, 기술 아이디어의 융합과 복잡성이 증가하는 최근의 추세는 전문가의 의견만으로 기술 아이디어를 평가하는 것을 더욱 어렵게 만들었다. 이에 따라, 기술 아이디어 스크리닝의 효과 및 효율을 개선하기 위한 정량적 분석 방법이 많

은 주목을 받았다. 기존 문헌에서는 기술 아이디어의 가치를 평가하기 위해 주로 특허 데이터 기반 접근법이 제시되었으며, 입력 데이터에 따라 인용 정보, 서지 정보, 텍스트 정보를 활용한 방법으로 구분된다. 인용 정보를 활용한 방법은 주로 곡선 적합(curve fitting) 기법과 확률 모델을 활용하여 미래의 특허 전방 인용 수를 추정하였고, 서지 정보를 활용한 방법은 다양한 특허 지표와 전방 인용 수 사이의 관계를 기계학습 모델을 활용해 추정하였다. 하지만 두 접근법은 특허가 등록된 이후에만 사용할 수 있으므로 기술 개발의 초기 단계에 적용하는 데 한계가 있었다. 특허 텍스트를 기반으로 한 연구는 특허가 포함하고 있는 단어의 중요도와 빈도를 기반으로 특허를 벡터로 표현한 뒤, 기계 학습을 통해 전방 인용 수와의 관계를 추론하였다. 이러한 연구는 특허의 등록 여부와 관계없이 기술 아이디어에 대한 텍스트 정보만 있어도 분석이 가능하기 때문에 초기 단계의 기술 아이디어 스크리닝에도 적용이 가능하다는 장점이 있지만, 텍스트의 의미와 맥락 정보를 고려하지 않기 때문에 자세한 기술적 요소를 반영하지 못한다는 한계점 또한 존재한다.

이에 Hong et al. (2022)은 (1) 기술 개발 초기 단계의 기술 아이디어 스크리닝 과정에는 특허의 기술적 요소 파악이 중요하다는 점, (2) 특허의 단어 사용 패턴을 통해 기술 아이디어를 식별할 수 있다는 점, (3) 특허의 전방 인용 수는 그 특허가 설명하는 기술 아이디어의 가치 평가를 위한 단서를 제공한다는 점을 바탕으로, 특허의 텍스트 정보를 활용하여 특허의 전방 인용 수를 예측하는 방법론을 제시하였다. 이 방법론은 크게 4가지 단계를 통해 수행된다. 첫 번째 단계에서는 기술 아이디어 스크리닝의 대상이 될 특허를 수집하고, 텍스트 정보(예: 초록, 본문, 청구항)에 대한 전처리를 수행한다. 전처리는 소문자 변환, 불용어(예: “a”, “is”, “the”) 제거, 어간 추출, 흔히 사용되는 단어 제거(예: 수집한 특허의 70% 이상 등장하는 단어) 등의 과정으로 이루어지며, 이를 통해 특허가 설명하는 기술 아이디어를 표현하는 데 중요하게 사용되는 단어들만 활용할 수 있게 된다. 두 번째 단계에서는 전처리가 완료된 특허의 텍스트 정보를 활용하여 Word2Vec을 학습시키고, 이를 통해 특허 텍스트에 포함된 모든 단어의 벡터를 추출한다. 세 번째 단계로, 각 특허의 텍스트 정보에 포함되는 모든 단어의 벡터를 쌓아 올려 2차원 이미지를 만들고, 합성곱 신경망(convolutional neural network: CNN)을 구축하여 특허의 실제 전방 인용 수를 예측하도록 학습시킨다. 마지막으로, 방법론의 예측 결과와 실제 결과를 비교하여 방법론을 통한 기술 아이디어의 가치 평가의 성능을 검증한다. 이러한 과정을 통해 학습이 완료된 Word2Vec 모델과 합성곱 신경망 모델을 활용하여 기술 개발자나 연구자가 고안한 새로운 기술 아이디어의 잠재적 가치를 특허 출원 전 단계에서 미리 확인할 수 있다.

해당 연구에서 제안한 방법론을 제약 기술 분야의 35,376개 특허에 적용하여 실험을 수행한 결과, 기존의 특허 텍스트 정보를 기반으로 한 기술 아이디어 스크리닝 방법론보다 정확도와 신뢰도 측면에서 우수함을 확인하였으며, 특히 낮은 잠재력을 가진 대부분의 아이디어를 식별하는 것을 확인하였다. 해당 연구에서 제안한 방법론은 기술 개발 초기 단계의 기술 아이디어 스크리닝을 지원함으로써, 자원이 한정적인 중소기업의 기술 혁신 역량을 강화시키고 혁신 기회를 효율적으로 탐색할 수 있게 하는 지능 플랫폼 시스템의 핵심 요소로서 사용될 수 있을 것으로 기대된다.

2. 의미 기반 언어모델을 활용한 대학 기술 이전 공급자-수요자 중계

대학의 기술 이전은 연구 성과 확산을 통해 경제적 이익을 얻고자 하는 기술 공급자(inventor)와 외부 기술 도입을 바탕으로 사업의 효율성을 개선하고자 하는 기술 수요자(licensee)의 상호 보완적인 이해관계에 근거하여 학계와 산업의 지속적인 관심을 받고 있다. 기업은 대학에서 개발된 기술과 연구 성과를 활용하여 상대적으로 낮은 비용과 시간 투자로 신제품 개발, 기술 혁신, 시장 경쟁력 강화의 기회를 얻을 수 있으며, 대학과의 협력을 통해 연구개발의 불확실성을 완화하고 연구 인프라 및 전문지식을 공유 받음으로써 직면한 기술적 문제를 효과적으로 해결할 수 있다. 대학 역시 보유기술 및 연구 성과의 상업화를 통해 지역산업 발전에 기여할 뿐 아니라 기술료 수익 창출을 바탕으로 연구 및 교육 인프라 개선을 도모한다. 성공적인 기술 이전을 위해서는 상호 이해관계에 부합하는 기술 공급자와 수요자를 식별하는 과정이 필수적이다. 이에 대학의 기술이전 담당 부서(technology licensing office: TLO)는 대학이 제공할 수 있는 기술적 기능과 기업이 필요로 하는 사업 요구사항을 연결함으로써 대학 기술의 잠재적인 구매자를 탐색한다. 하지만, 대부분의 과정이 기술이전 전담인력에 의해 수행됨에 따라 많은 시간이 소요되고 노동 집약적이며, 개인의 인적 네트워크를 기반으로 하여 좁은 탐색 범위에 국한될 수 있다는 한계점이 존재한다. 특히, 이와 같은 전문가 중심의 기존 접근 방법은 대학이 보유한 기술의 수와 기술 지식의 복잡성이 증가함에 따라 기술 이전 과정의 효율 및 효과를 감소시키는 원인이 되고 있다.

이에 Lee et al. (2023)은 의미 기반 언어모델을 활용하여 대학 보유기술의 기술적 기능과 기술 수요 기업의 사업적 요구사항 간의 의미적 관계를 파악하고, 이를 바탕으로 대학의 기술 이전을 위한 적절한 기술 공급자와 수요자를 식별하는 분석 프레임워크를 제안하였다. 우선적으로, 대학 연구자들에 의해 생산된 연구 결과물과 대학의 기

술 이전 계약 문서로부터 텍스트 형태로 이루어진 기술적 기능과 사업 요구사항 데이터베이스를 구축하고, 텍스트 클리닝 기법을 활용하여 각 텍스트가 핵심적인 기술 및 사업 용어로 구성되도록 전처리하였다. 다음으로, 구축된 데이터베이스를 사용하여 의미 기반 언어모델의 일종인 fastText를 학습시킴으로써 기술 및 사업 용어를 고차원 벡터 공간에 임베딩하고, 이를 바탕으로 기술적 기능-사업 요구사항 지형(technology characteristic-business requirement landscape)을 형성하였다. 특히, fastText는 단어뿐만 아니라 단어를 구성하는 문자 N-gram에 대해서도 학습이 이루어지므로, 학습에 사용되지 않은 단어에 대한 벡터 또한 추론할 수 있다. 이러한 특징은 미래 지향적인 개념을 포함함에 따라 기존 기술 용어를 재구성하는 경우(예: 3D Bioprinting, Neuro-robotics)가 잦은 대학 기술의 특성을 고려할 수 있어, 대학 기술 이전 측면에서 기술적 기능과 사업 요구 사항 간의 의미적 관계를 효과적으로 포착할 수 있다. 이후 기술적 기능-사업 요구사항 지형으로부터 각 단어들의 벡터를 추출하고, 이들의 평균을 취함으로써 각 기술적 기능 및 사업 요구사항의 대표 벡터를 도출하였다. 특히, 공통의 맥락을 공유하는 기술 및 사업 용어들이 비슷한 벡터를 가지므로, 각 기술적 기능과 사업 요구사항 간 유사성을 벡터 유사도를 기반으로 파악할 수 있다. 이에 따라 코사인 유사도 지수를 사용한 유사성 분석을 수행하여 기술 공급자가 제공할 수 있는 기술적 기능과 기술 수요자가 얻고자 하는 사업 요구사항 사이의 의미적 유사성을 측정하고, 이를 바탕으로 잠재적인 기술 공급자-수요자 쌍을 식별하였다. 마지막으로, 실제 대학의 기술 이전 담당 부서에서 실시한 기술 이전 계약 기록을 기준으로, 해당 방법론에 의해 식별된 기술 공급자-수요자 쌍의 매칭률(matching rate)을 산출하여 대학의 기술 이전 과정을 지원하는 도구로서의 신뢰성과 타당성을 평가하였다.

이 연구에서 제안하는 방법론의 성능을 개선하기 위해서는 활용하는 언어모델의 고도화, 특허 또는 논문 등의 추가 데이터 활용, 단어 벡터 표현 자체에 대한 성능 검증, 다수의 기술 공급자 및 수요자에 대한 대응 등 지속적인 노력이 필요하다. 그럼에도 불구하고, 이 연구는 정량적 데이터를 바탕으로 하는 체계적이고 자동화된 시스템을 제공함으로써 기술 이전을 위한 잠재적인 기술 공급자-수요자 식별에 필요한 시간과 비용을 절감하는데 도움이 될 것으로 판단된다.

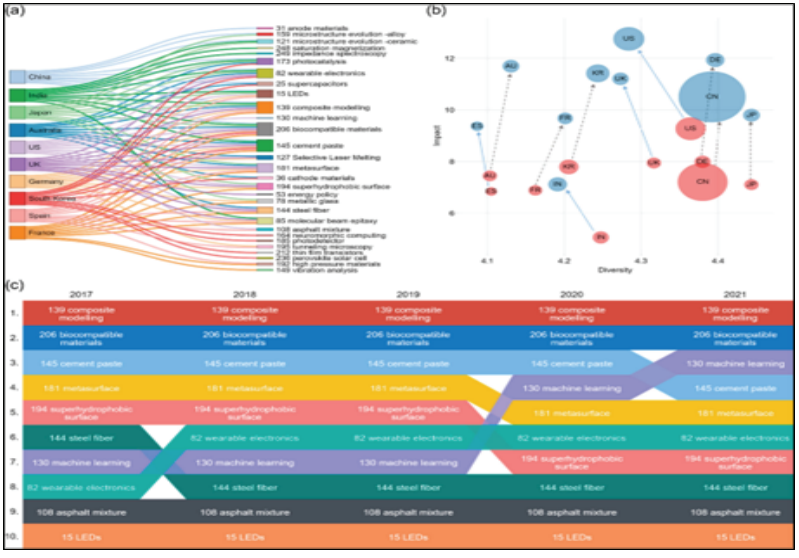
3. 문서 군집화를 통한 국가/기관별 연구 동향 분석

연구 동향 분석은 특정 분야의 최신 연구 개발 현황 및 신흥 주제에 대한 정보와 새로운 연구 방향 탐색을 위한 전략적 통찰력을 습득하기 위한 필수적인 작업이다. 기

존에는 특정 분야의 연구 동향을 분석하기 위해 전문가의 의견에 의존했지만, 과학기술 문헌의 양과 다양성이 증가함에 따라 전문가 기반 분석은 상당한 시간과 비용을 요구하고 분석 결과가 일관적이지 못하다는 한계를 보여주었다. 이에 따라, 과학기술 문헌의 텍스트 정보를 수치화하여 체계적이고 자동화된 방법으로 연구 동향을 파악하기 위한 분석 프레임워크의 필요성이 대두되었다.

이에 Choi and Lee (2024)는 재료과학 분야의 연구 동향을 파악하기 위해 논문의 텍스트 정보에 언어 모델과 군집화 알고리즘을 적용한 분석 프레임워크를 제시하였다. 먼저, 약 30만 건의 재료과학 논문의 초록을 활용하여 Doc2Vec(Le and Mikolov, 2014) 모델을 훈련시켰고, 이를 통해 각 논문의 연구 내용을 대표하는 논문 벡터를 추출하였다. 이후 밀도 기반 군집화 알고리즘인 HDBSCAN(hierarchical density-based spatial clustering of applications with noise)을 적용하여 비슷한 연구 내용을 다루는 논문을 군집화하고, 군집별 연구 내용을 파악하여 257개의 재료 과학 관련 주제를 도출하였다(McInnes et al., 2017). HDBSCAN으로 군집화 되지 못한 데이터는 이상치로 분류되었으며, 주로 다학제적인 연구 주제를 가지는 논문으로 확인되었다. 마지막으로, 특정 연구 주제에 속한 논문 벡터의 평균을 해당 연구 주제를 대표하는 벡터로 간주하고, 이와 함께 각 논문의 논문 벡터 및 메타 데이터를 활용하여 연구 동향 분석을 실시하였다. 이를 통해 〈그림 10〉과 같이, 각 국가 및 연구 지원 기관, 학술지

〈그림 10〉 문서 군집화를 통한 국가/기관별 연구 동향 분석 결과(Choi and Lee, 2024)



의 연도별 주요 연구 관심사를 분석하고 연구 주제 간 관련성 네트워크를 구축하여 재료과학 관련 연구 주제의 융합 현황을 파악하였다.

이 연구는 언어모델을 활용하여 논문을 벡터화하고 이에 대해 군집화 알고리즘을 적용함으로써 키워드 분석 및 통계 분석을 기반으로 하는 기존 접근법의 한계점으로 지적되는 연산량 문제를 완화하였다. 이를 통해 논문 데이터를 바탕으로 연구 동향을 파악하기 위한 분석 프레임워크를 체계적이고 재현 가능한 방식으로 제시하였으며, 이는 연구자와 정책 입안자의 연구 분야 모니터링 과정에 효율성을 제공할 것으로 기대된다.

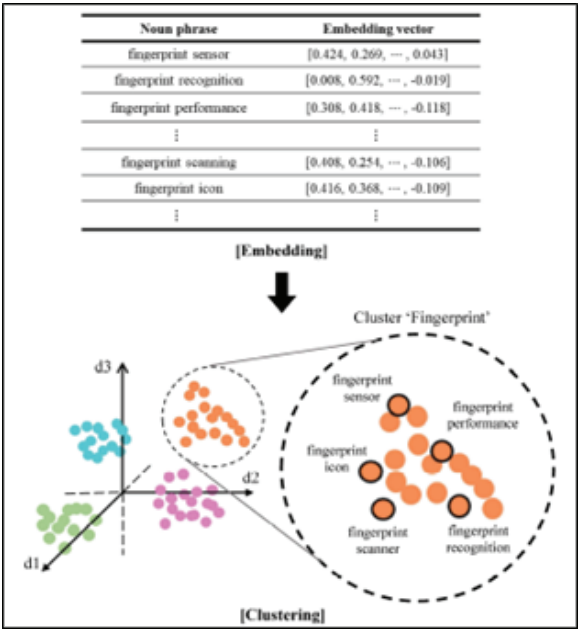
4. 감성 분석을 통한 제품 출시 초기 소비자 불만 모니터링

기업은 제품 출시 이후 실제로 제품을 사용한 소비자들이 제기하는 의견 및 불만사항을 활용하여 제품의 품질 관리 전략을 수립한다. 특히 제품 출시 초기의 소비자 반응은 제품 기획 과정에서의 기술적 결함이나 제품-서비스 시스템 구조의 불완전성과 밀접한 관련이 있으므로 더욱 중요하게 여겨지며, 이를 분석함으로써 기업은 제품 출시와 관련하여 장기적으로 발생할 수 있는 문제점을 파악하고 이에 대한 개선 방향성을 결정할 수 있다. 기존에는 이러한 소비자들의 직접적인 의견을 접할 수 있는 채널이 한정적이었기 때문에, 기업들은 소비자 반응을 수집하기 위해 많은 시간과 비용을 투자하였다. 그러나 인터넷이 발달하면서 현대의 소비자들은 제품을 사용한 경험이나 의견을 소셜 미디어 혹은 온라인 커뮤니티를 통해 적극적으로 표현하는 경향을 보였고, 이전에 비해 더 많은 소비자 의견을 더 쉽게 접할 수 있게 되었다. 이에 따라 학계에서는 소셜 미디어와 온라인 커뮤니티에서 소비자 의견을 수집하고 감성 분석(sentiment analysis)을 수행함으로써 제품의 품질에 대한 소비자 여론을 체계적으로 파악하기 위한 연구가 활발히 이루어졌다. 이러한 기존 연구들이 제품 품질 관리 전략 수립과 관련한 의사 결정을 지원하기에 유용함을 보였으나, 주로 문서 단위의 감성 분석을 수행함으로써 분석의 정확성 측면에서 한계가 있었다. 특히 소셜 미디어에 업로드 되는 소비자 의견은 한 가지 주제에 대한 일관적인 내용 보다 여러 주제에 대한 다양한 평가를 내포하는 경우가 많으므로, 문서가 아닌 문장을 대상으로 하는 보다 세부적인 수준의 감성 분석 기반 접근이 요구되고 있다.

이를 위해, Lee et al. (2021)은 제품 출시 초기에 소셜 미디어에 업로드 된 소비자 리뷰에 대해 문장 단위로 감성 분석을 수행하는 방법론을 제안하였다. 제안된 방법은 소비자 리뷰에서 제품의 기능이나 특성을 나타낼 수 있는 최소 단위를 명사구로 정

의하고, 이를 자연어 처리 기반 키워드 추출 알고리즘을 적용하여 추출한다. 이후 사전 학습된 Sentence-BERT(Reimers and Gurevych, 2019)를 활용하여 앞서 추출된 명사구들을 임베딩 벡터로 변환하고, 이에 대해 K-평균 군집화(K-means clustering) 알고리즘을 적용하여 공통된 내용을 다루는 명사구들끼리 군집화를 실시한다. 이때 〈그림 11〉과 같이, 군집을 이루는 명사구들을 소비자들의 잠재 불만 요소로 정의한다. 다음으로, 어휘 및 규칙 기반 감성 분석 도구인 VADER(valance aware dictionary and sentiment reasoner)를 도입하여 각 잠재 불만 요소에 대한 소비자의 반응이 긍정적인지 부정적인지를 정량적으로 측정한다. 긍정/부정 정도가 측정된 명사구를 바탕으로 각 잠재 불만 요소의 중요도를 시급성과 심각성 두 가지 척도로 나누어 분석한다.[17] 최종적으로, 시급성과 심각성이 동시에 높은 잠재 불만 요소를 식별하여 우선적으로 개선이 필요한 제품 개선 과제로 간주한다. 이 연구에서는 제안된 방법을 삼성

〈그림 11〉 사전 학습된 언어모델과 군집 분석을 활용한 잠재 불만 요소 추출 과정(Lee et al., 2021)



17) 시급성은 해당 잠재 불만 요소 중 부정 정도가 높은 명사구의 빈도로 정의되고, 심각성은 해당 잠재 불만 요소에 포함된 명사구의 부정 정도의 평균값으로 정의된다.

갤럭시 S10 제품의 출시 초기 소비자 리뷰 데이터에 적용하였고, '삼성페이 기능의 서비스 연결 및 로딩 실패', '지문인식 장치의 위치의 불편함과 낮은 지문 인식률', '적은 배터리 용량'을 개선 필요성이 높은 소비자 불만 사항으로 식별하였다.

이 연구는 대량의 텍스트로부터 공통된 내용을 파악하고 그 내용에 대한 감성 분석을 자동화된 방식으로 수행하여 제품 출시 초기에 개선되어야 할 사항에 대해 기업들이 빠르게 대응할 수 있는 분석 프레임워크를 제시하였다. 이러한 접근 방식은 소비자의 불만 사항을 시기적절하게 개선하여 기업들이 시장에서의 제품 성공률을 높이고, 장기적으로 높은 소비자 만족도를 유지할 수 있는 고객 관리 시스템의 기반을 마련할 수 있을 것으로 판단된다.

5. 한국어 언어모델을 활용한 지정상품 간 유사도 측정

상품심사는 상표 등록 절차 중 하나로, 상표 소유자의 권리를 보호하고 상품 출처의 혼동을 예방하기 위해 출원한 지정상품과 선출원 상표의 지정상품이 동일 또는 유사한지 판단하는 과정을 나타낸다. 특허청은 객관적이고 공정한 상품 심사를 위해 상품의 품질, 형상, 용도 등의 상품 속성과 생산 및 판매 부문, 수요자 범위 등의 거래 실정을 종합적으로 고려하여 상품 간 유사 여부를 판단하고 있으며, 상품심사의 신속성, 객관성 및 공정성을 유지하기 위해 유사군 코드를 활용한 유사상품 심사 기준을 제시하고 있다. 즉, 출원한 지정상품과 선출원 상표의 지정상품이 동일한 유사군 코드에 해당되면 양 상품은 서로 유사한 것으로 판단된다. 하지만 최근 다기능 혹은 통합기능 상품이 등장함에 따라 유사군 코드와 상관없이 현시점의 거래 통념에 따라 상품 간 유사 여부를 판단하는 사례가 증가하고 있다.

이에 한국지식재산연구원 (2023)은 지정상품의 유사 판단에 참고할 수 있도록 정의, 속성, 거래 실정 등 상품의 항목별 특징을 기술하는 상품해설서를 구축하고, 언어모델을 활용하여 지정상품 간 유사도를 측정하고 유사도에 대한 판단 근거를 수치로 제공하는 방법론과 시스템을 개발하였다. 우선적으로, 상품해설서의 한국어를 처리하기 위해 한국어 언어모델을 활용하여 상품의 특징을 나타내는 벡터를 항목별로 추출하고, 이를 활용하여 항목 수준과 상품 수준에서 유사도를 측정하였다. 다음으로, 항목 수준에서의 유사도 측정을 위해 모든 토큰에 대해 벡터를 추출하고 이를 평균하여 해당 항목의 벡터를 정의하였으며, 벡터 간 유사도 지표를 계산하여 두 지정상품의 항목별 유사도를 도출하였으며, 이를 종합하여 상품 수준에서의 유사도를 도출하였다. 최종적으로, 두 상품의 해당 항목에 포함된 단어 쌍별 유사도를 각 쌍이 항목 수준 유사도에

대한 기여도로 산정하여 상품의 항목 수준에서 유사도 판단의 근거를 제시하였다.

개발한 방법론의 성능은 활용하는 벡터 추출 방법, 유사도 지표, 상품 수준 유사도 측정 방법에 따라 달라질 수 있다. 이에 (1) 벡터 추출 방법의 경우 KoBERT, KLUE-RoBERTa, KoGPT-J, Ajoublue-GPT2 등을 포함하여 총 15종의 BERT와 GPT 계열의 방법론을, (2) 유사도 지표의 경우 코사인 유사도, 유클리디안 거리 등을 포함하는 3종의 유사도 지표를, (3) 상품 수준 유사도 측정 방법의 경우 항목별 벡터를 연결하여 산출된 대표 벡터 간 유사도 지표를 계산하는 벡터 연결 방법과 항목별 유사도에 가중치를 적용하는 항목별 유사도 종합 방법을 활용하여 총 90가지 모델에 대한 비교 실험을 수행함으로써 최적 모델을 도출하였다. 최적 모델을 도출하는 과정에서 (1) 동일 유사군 코드에 속한 상품 쌍과 그렇지 않은 상품 쌍에 대한 유사도 비교 분석, (2) 유럽 연합 지식재산권 사무소(European Union Intellectual Property Office: EUIPO)의 상품 유사성 판결례와의 비교 분석, (3) 국내 상품 유사성 판단 판결례와의 비교 분석을 수행하였다.

이를 바탕으로 〈그림 12〉에 도식화된 것과 같이 지정상품간 유사도 측정과 근거 제시를 위한 플랫폼 시스템을 개발하였다. 영문 데이터에 대한 처리, 활용하는 언어모델의 고도화 등 지속적인 성능 개선에 대한 노력이 필요하지만, 이러한 시스템은 상품해설서 내용을 토대로 일관성 있는 지정상품 간 유사도 측정을 가능하게 하며, 상품심사

〈그림 12〉 비료와 고토비료의 유사도 측정 및 근거 제시 예시

항목	01-000001	01-000009	유사도
상품명	비료	고토비료	63.24%
영문명	fertilisers	dolomitic fertilizers	80.65%
정의	경작지에 뿌리는 영양 물질. 토지의 생산력을 높이고 식물의 생장을 촉진하는 물질로 질소, 인산, 칼륨을 비료의 3요소라고 함. (@네이버사전)	작물에게 유효한 마그네슘이 다량 함유되어 있는 비료	73.63%
속성(기능/용도)	경작지에 뿌리는 영양물질로, 식물의 성장에 필요한 영양소를 공급하여 성장을 촉진시킨다. 식물의 건강을 유지하고 수확량을 증가시키는데 중요한 역할을 함.	고토비료는 화분재배, 텃밭재배, 농업재배 등 다양한 용도로 사용됨	42.23%
속성(형상)	과립형태	고토비료는 액체비료, 고형비료, 엽면시비용 비료 등 다양한 형태	23.97%
속성(원재료)	인소, 질소, 칼륨 등 혼합물	마그네사이트, 돌고래뼈, 해조류 등	9.82%
속성(사용방식)	적정량을 토양에 살포하여 흙과 잘 섞어줌	적정량을 흙에 섞어서 토양에 살포	91.18%
속성(기타 특이사항)	적정량과 올바른 비료를 선택하여 사용하는 것이 중요	사용할 때는 식물의 종류, 토양의 종류, 기후 조건 등을 고려하여 적절한 양을 사용해야 함	64.82%
KSIC(생산부문)	20311 질소 화합물, 질소 · 인산 및 칼리질 화학비료 제조업 20312 복합비료 및 기타 화학비료 제조업 20313 유기질 비료 및 상토 제조업	20312 복합비료 및 기타 화학비료 제조업 20313 유기질 비료 및 상토 제조업	-
KSIC(판매부문)	46732 비료 및 농약 도매업	46732 비료 및 농약 도매업	-
유사군코드	G0101	G0101	100.00%
심판결례 및 협의심사사례	(비유사) 2014원004680 비료 VS 미가공 단백질 플라스틱(Unprocessed protein plastics), 미가공 아크릴수지(Unprocessed acrylic resins), 미가공 인조수지(Unprocessed artificial resins), 미가공 중합 플라스틱(Unprocessed polymerization plastics), 미가공 축합 플라스틱(Unprocessed condensation plastics), 미가공 플라스틱(Unprocessed plastics), 셀룰로오스 플라스틱 수지(Cellulose plastic resins) (통념) 이 사건 출원상표의 지정상품들과 상표법시행규칙 제40조 제1항 등의 규정에 따른 유사상품군이 다르고, 상품 자체의 속성인 품질, 형상, 용도와 생산 부문, 판매 부문, 수요자의 범위 등 거래의 실정이 차이가 있어 일반 거래의 통념상 서로 유사하지 아니하므로, 이 사건 출원상표는 선등록상표와 그 지정상품이 유사하지 않음		-
상품 수준 유사도			57.29%

의 공정성과 신뢰성을 높이는 데 도움이 될 것으로 판단된다.

Ⅴ. 언어모델 활용을 위한 주요 고려 사항

본 장에서는 공공부문에서 언어모델의 활용 영역 확장과 새로운 유형의 공공 서비스 발굴에 기여하기 위해, 효과적인 언어모델 학습을 위한 텍스트 데이터의 구조와 품질 기준을 설명하고 공공부문에서 활용 가능한 공개 데이터 셋을 소개한다. 이어서 다양한 분석 기법과의 결합을 바탕으로 하는 다양한 언어모델 활용 방안을 제시하고, 언어모델의 효과성을 검증하기 위한 성능 평가 방법을 소개한다.

언어모델의 학습은 대규모의 텍스트 데이터를 통해 이루어지며, 학습에 사용되는 텍스트의 구조나 포함되는 내용에 따라 언어모델의 성능이 달라질 수 있다. 기본적으로 언어모델은 책이나 기사와 같이 정답이 존재하지 않는 일반적인 서술문 형태의 자연어 텍스트에 대해 사전 학습을 수행하며, 이를 통해 포괄적인 자연어 이해 능력을 가질 수 있게 된다. 대량의 텍스트를 통해 사전 학습된 언어모델은 텍스트 전처리, 텍스트 분류, 정보 검색 등 기본적인 작업을 비롯하여 텍스트 생성, 정보 추출, 챗봇 등 고도화된 작업까지 다양한 형태의 자연어 처리 작업에 활용될 수 있다. 언어모델이 활용되는 주요 자연어 처리 작업에 대한 설명은 부록의 〈표 A1〉에 정리되어 있다. 특히 사전 학습이 완료된 언어모델을 특정한 자연어 처리 작업에 최적화하여 적용하기 위해서는 미세 조정 과정이 필요하며, 이를 위해 해당 작업의 목적에 맞는 데이터를 활용하여 추가적인 학습이 수행된다. 사전 학습에 사용되는 텍스트 데이터와 달리, 미세 조정을 위한 텍스트 데이터에는 특정한 형태의 라벨(label)이 할당된다. 이를 통해 언어모델은 자연어 처리 작업의 논리를 파악하고, 해당 작업을 수행하는 데 필요한 학습 기반을 갖추게 된다. 라벨은 자연어 처리 작업의 종류에 따라 텍스트에 대한 분류, 속성, 태그, 대응되는 다른 텍스트(예: 번역문, 요약문) 등으로 정의될 수 있으며, 〈표 3〉은 주요 자연어 처리 작업에 대한 텍스트 데이터의 라벨 구조 예시를 보여준다.

〈표 3〉 언어모델의 주요 자연어 처리 작업별 텍스트 데이터 라벨 구조 예시

대분류	소분류	라벨 대상
텍스트 전처리	문법적 오류 해결	문법적 오류가 포함된 문장에 대한 교정 문장 혹은 문장 내 오류 위치와 교정값
	품사 태깅	문장 내 품사 위치 및 품사 태그

텍스트 분류	문서 분류	사전에 정의된 카테고리나 클래스 기준에 속하는 분류값
	감성 분석	사전에 정의된 감성 분류 기준에 속하는 분류값
정보 검색	질의 응답	질문과 답변을 구분할 수 있는 태그(tag) 및 질문의 유형(예: 단답형, 리스트형) 정보
텍스트 생성	기계 번역	원문을 다른 언어로 번역한 텍스트
	텍스트 요약	원문의 주요 내용에 대해 추출 또는 생성한 요약문
정보 추출	개체명 인식	원문 내 개체명의 위치 및 개체 클래스 분류값
	관계 추출	원문 내 관계 분류값 및 관계를 가지는 주격/목적격 대상(단어, 구, 문장 등)의 위치

언어모델을 실제 서비스에 활용하기 위해서는 학습에 사용되는 텍스트 데이터의 품질에 대한 평가가 필수적으로 선행되어야 한다. 특히 특정 분야의 일부 사용자들을 대상으로 하는 것이 아니라 공공의 이익을 위해 누구나 사용 가능한 공공 서비스에 활용되는 언어모델의 경우, 모델이 부적절한 결과를 도출하지 않도록 모델 학습을 위한 텍스트 데이터에 대해 더욱 엄격한 품질 기준을 마련할 필요가 있다. 이와 관련하여, 〈표 4〉와 같이 텍스트 데이터가 가지는 다양한 특성에 따라 품질을 평가하기 위한 기준과 평가 방안들이 제시된 바 있다.

〈표 4〉 텍스트 데이터의 품질 특성별 평가 기준 및 방안

품질특성	평가 기준	평가 방안
다양성	텍스트 데이터의 고유 어휘, 문형 등의 규모와 빈도	어절 단위 통계
중복성	텍스트 데이터 내 정보의 중복성 또는 유사한 정도	텍스트 데이터 내 중복 어절수(N-gram) 통계
구문 정확성	언어모델 학습에 유의한 텍스트 데이터 구조 적합성 및 데이터 간 형식(타입, 클래스 유효값, 정규식 패턴 등) 일치성	정의된 구조로 파싱(parsing)이 불가한 텍스트 데이터 검출 및 형식 이상치 검출
의미 전달성	텍스트 데이터의 언어 전달력(예: 맞춤법) 및 표현의 유창성(fluency)	맞춤법 검사 API 활용 또는 데이터 샘플에 대한 인력 검수
의미 적정성	자연어 처리 작업에 대한 텍스트 데이터 라벨의 적정성	전문 인력 검수
신뢰성	(사실성) 사건, 현상, 수치 등에 대해 텍스트 데이터가 실제와 부합하는지 여부 (적시성) 시의성을 가지는 텍스트 데이터가	(사실성) 사실 여부 또는 제공된 근거의 타당성에 대한 전문 인력 검수

	현재 시점에서 가장 최신의 정보인지 여부	(적시성) 데이터 생성, 수집, 현재 시점 사이 발생한 정보의 변경 여부에 대한 검수
편향성	인종, 성별, 연령 등 사회문화적 관점에서 텍스트 데이터 내 인식 편향의 정도	공정성 평가[18] 또는 편향 감지 언어모델 활용
유해성	욕설, 음란한 발언 등 저속한 언어 및 혐오 표현의 사용 정도	사전(dictionary) 기반 유해 표현 검출 모델 활용
민감정보비식별화	식별 가능한 개인정보 또는 민감한 정보를 포함하는 정도	개체명 인식 모델을 통한 검출

일부 공공기관에서는 공공부문에서의 언어모델 활용을 지원하기 위해 앞서 소개한 품질 평가 기준을 바탕으로 검증이 완료된 학습용 텍스트 데이터 셋을 공개하고 있다. 〈표 5〉는 이와 같이 공공 서비스에 활용되는 언어모델의 학습을 위한 텍스트 데이터 셋들을 정리한 자료이며, 모두 AI-Hub를 통해 공개 및 관리되고 있다.[19]

〈표 5〉 AI-Hub 공공부문 텍스트 데이터 셋 요약

데이터명	데이터 개요
행정 문서 대상 기계독해 데이터	비정형 텍스트인 행정 문서를 이용하여 표와 일반 텍스트 데이터에 대한 다양한 형식의 질의응답을 지문-질문-답변으로 구성한 데이터 셋
공공 분야 고객 응대 데이터	보건・복지, 도시・교통, 전자상거래 등 6개 공공 분야에 대한 전사 텍스트와 감정 및 의도 태깅, 요약문 등으로 이루어진 데이터 셋
민원 업무 자동화 인공지능 언어 데이터	창원시 콜센터 및 국민신문고의 민원에 대해 개체명 태깅, 의도 분류, 부서 예측, 키워드 분류를 부여한 질의 응답 형태의 데이터 셋
산업정보 연계 주요국 특허 영-한 데이터	미국, 유럽, 일본, 중국의 특허명세서에 대해 주요 내용을 한국어로 번역하고 KSIC(표준산업분류) 코드를 부여한 데이터 셋
특허 분야 자동분류 데이터	고른 분포의 국제・국내 산업분류별 원문을 대상으로 특허 코드를 부여한 데이터 셋
법률/규정(판결서, 약관 등) 텍스트 분석 데이터	판결문의 기초사실, 주장 등을 가공하고, 판례 기반 판결문 분석 정보 및 약관의 유・불리 조항 판단을 위한 위법성, 유리 판단 사유를 기록한 데이터 셋

공공 서비스는 교육, 의료, 교통, 환경 등 사회 전반에 걸쳐 다양한 형태로 제공되

18) 공정성 평가 방안으로는 그룹 공정성과 반사실적 공정성 평가가 있다.
19) 한국지능정보사회진흥원.

고 있으며, 이에 따라 텍스트뿐 아니라 이미지나 음성 등 서로 다른 비정형 데이터들이 복합적으로 활용되는 경우가 많다. 또한, 공공부문의 영역마다 각자의 목적에 맞는 서비스를 제공하기 위해 다양한 형태의 의사결정 과정이 수반된다. 이러한 특성으로 인해, 범용으로 개발된 언어모델을 공공부문에 단독으로 적용하는 것은 언어모델의 실효성 감소와 활용 범위의 제약을 초래한다. 그러므로 공공부문과 관련한 주요 의사결정 문제를 해결하는 데 사용될 수 있는 다양한 형태의 데이터와 분석 기법을 식별하고, 이를 언어모델과 연계하여 분석 범위와 시사점을 다각화할 필요가 있다. 그러나 대부분의 공공부문에서는 챗봇을 활용한 민원 상담 등 언어모델의 기본적인 기능에만 국한되어 있어 여전히 활용도가 제한적인 실정이다. 따라서 이미지, 음성 데이터와 같은 다양한 형태의 데이터 활용 및 군집화 알고리즘, 이상치 탐지 기법과 같은 다각적인 의사결정을 지원할 수 있는 방법론과의 결합을 통해 공공부문에서의 언어모델 활용 범위를 증대할 수 있을 것으로 판단된다. 예를 들어, 언어모델과 이상치 탐지 기법을 결합하여, 공공보건 분야에서 의료 기록 텍스트 데이터로부터 특이 질환자를 신속하게 식별하거나, 군집화 기법을 이용해 교통 분야에서 대중교통 이용자 리뷰를 분석하고, 이를 바탕으로 교통 계획의 개선안을 제시하는 의사결정 지원 도구를 개발할 수 있다. 또한, 언어모델과 기계 학습 기반 분류 모델을 결합하여, 공공행정 분야에서 고객의 대화 내용과 녹음을 분석함으로써 각 민원의 유형별 감성을 파악할 수 있다. 이를 통해 민원 처리의 우선순위를 결정하기 위한 객관적 기준을 마련할 수 있다.

국내에서 생산된 공공부문 텍스트 데이터는 대부분 한국어로 작성되어 있다는 점을 고려할 때 이 데이터를 효율적으로 처리하고 분석하기 위해서는 한국어를 정확하게 이해하고 처리할 수 있는 언어모델이 필수적이다. 이를 위해, 개인 연구자를 비롯한 기업, 연구기관, 대학 등 다양한 연구 주체가 국내 뉴스, 위키피디아, 책, 소셜미디어 등 대규모 한국어 텍스트 데이터로 언어모델을 사전 학습하고, 오픈 소스의 형태로 이를 배포하고 있다. 〈표 6〉은 오픈 소스의 형태로 활용할 수 있는 17종의 트랜스포머 기반 한국어 언어모델을 정리한 것이다. 아래 모델들은 학습에 사용된 데이터와 모델 크기에 따라 한국어 자연어 처리 성능이 다르므로 언어모델간 성능 비교 분석이 필요할 수 있다.

〈표 6〉 트랜스포머 기반 한국어 언어모델 요약

구분	모델 이름	개발 주체	학습 데이터	모델 크기 (학습 파라미터 수)
BERT 계열	KorBERT	ETRI	뉴스, 백과사전	~110M
	KoBERT	SKT	위키피디아	~92M
	HanBERT(IP)	투블럭 AI	특허문서	~128M
	KoreALBERT	삼성 SDS	위키피디아, 뉴스, 책 줄거리 등	~12M
	KLUE-BERT	KAIST Upstage	모두의 말뭉치, 나무위키 등	~111M
	KR-BERT	서울대	위키피디아, 뉴스	~96M
	KR-SBERT	서울대	위키피디아, 뉴스, KLUE-NLI	~101M
	DistillKoBERT	개인	위키피디아, 뉴스 등	~28M
	KcBERT	개인	네이버 뉴스 댓글	~109M
	KcELECTRA	개인	네이버 뉴스 댓글	~124M
	KoELECTRA	개인	모두의 말뭉치, 위키피디아, 뉴스 등	~110M
	KoBigBird	개인	위키피디아, 뉴스	~113M
GPT 계열	KoGPT2	SKT	위키피디아, 뉴스, 모두의 말뭉치, 청와대 국민청원 등	~125M
	GPT3-kor -small	개인	리뷰 데이터, 블로그, 모두의 말뭉치, 위키피디아, 나무위키 등	~119M
	KoGPT-J	개인	AIHub 대화	~124M
	Ajoublue-GPT2	개인	AIHub 대화	~125M
	KoAlpaca	개인	한국어 Instruction-following 데이터	~65B
	KoVicuna	개인	shareGPT	~7B
	Polyglot-ko	EleutherAI	위키피디아, 뉴스, 블로그 등	~12.8B
	KULLM	고려대	위키피디아, 뉴스 및 한국어 instruction-following 데이터	~12.8B
	HyperCLOVA	네이버	네이버 뉴스, 카페, 블로그, 모두의 말뭉치, 위키피디아 등	~82B

공공부문에서 언어모델을 실질적으로 활용하기 위해서는 학습이 완료된 언어모델의 적정성과 실효성에 대한 판단이 필요하다. 따라서 언어모델의 학습이 충분히, 목적에 맞게 수행되었는지 검증하는 성능 평가가 실시되며, 이는 모델의 성능을 수치화하여 나타내는 정량적 평가 혹은 모델이 도출하는 결과를 사람이 직접 검증하는 정성적 평

가를 통해 이루어진다. 먼저 정량적 평가와 관련하여, 텍스트 분류, 개체명 인식, 품사 태깅 등 언어모델이 도출하는 결과가 특정 대상에 대한 식별이나 분류인 경우 F1 점수를 계산하여 정량적인 성능을 평가할 수 있다. F1 점수는 정밀도(precision)와 재현율(recall)의 조화 평균으로 계산되며, 정밀도는 모델의 예측값 중 정확히 예측한 값의 비율을 나타내고 재현율은 실제값 중에서 모델이 정확히 예측한 비율을 나타낸다. 텍스트 생성이나 질의 응답 작업과 같이 언어모델이 도출하는 결과가 텍스트 형태로 나타나는 경우 주로 BLEU(bilingual evaluation understudy)와 ROUGE(recall-oriented understudy for gisting evaluation)를 사용하여 성능 평가를 수행한다. BLEU는 언어모델이 생성한 텍스트와 정답 텍스트 간 일치하는 단어의 수와 순서를 정밀도를 기반으로 계산하는 지표이며, 이때 단어의 순서는 연속된 N개의 단어의 일치 여부로 판단한다(Papineni et al., 2001). ROUGE는 BLEU와 유사하게 두 텍스트 간의 단어 단위 일치 수준을 나타내는 지표이지만, 이를 재현율을 기반으로 계산한다는 차이가 있다(Lin, 2004). 또한, 언어모델 자체의 성능을 평가하는 지표로 퍼플렉시티(perplexity)가 있다. 퍼플렉시티는 모델이 텍스트를 생성하는 과정에서의 불확실성을 수치화한 것으로, 언어모델이 생성한 각 단어의 예측 확률의 역수를 모두 곱하고 단어의 총 수에 해당하는 제곱근을 적용하여 계산된다(Jelinek et al., 1977). 따라서 퍼플렉시티 값이 낮을 수록 언어모델의 성능이 높다고 평가한다. 한편, 정성적 평가는 언어모델이 도출하는 결과의 적합성과 신뢰성을 판단하기에 적절한 평가 항목을 구성하고, 이를 바탕으로 사람이 평가 점수를 측정하는 방식으로 이루어진다. 이를 위해, 명목 척도로써 구글에서 발표한 자율 발화 모델에 대한 평가 지표인 SSA(sensible and specificity average)를 사용할 수 있다(Adiwardana et al., 2020). SSA는 언어모델이 도출하는 결과가 입력된 질문의 의도에 부합하는지와 구체적인 답변을 제공하는지에 대해 0, 1의 값으로 그 여부를 판단한다. 또한, 서열 척도로써 리커트(Likert) 척도를 사용하여 평가하거나 동일한 데이터에 대해 여러 모델이 도출하는 결과 간의 선호를 비교하여 평가할 수 있다. 이와 같은 평가 항목으로는 언어적 유창성(fluency), 사용자 의도 부합성(engagingness), 답변 일관성(consistency) 등이 사용되고 있다.

Ⅵ. 결론

본 연구는 언어모델에 대한 이해를 높이고 공공부문에서의 활용을 촉진하기 위한 목적으로 수행되었다. 이를 위해 언어모델의 개념을 제시하고 BoW, Word2Vec,

ELMo, BERT, GPT 등 언어모델의 주요 접근법에 대해 설명하였다. 또한 언어모델의 거대화, 경량화, 강화학습의 적용과 같은 최신 언어모델 연구동향을 조사하였고, 언어모델의 공공부문 행정혁신 사례를 소개하였다. 이에 더하여 공공부문에서 언어모델을 도입할 때 고려해야하는 주요 사항을 제시하고, 이를 바탕으로 향후 언어모델의 활용 방향을 제안하였다. 공공부문에서의 언어모델 활용과 관련하여 필수적으로 요구되는 내용을 모두 다루고자 하였으나, 연구의 범위와 지면의 제약으로 인해 몇몇 접근법과 사례는 포함하지 못하였다. 특히 현재 실질적으로 활용되고 있는 주요 접근법을 중점적으로 소개함에 따라 언어모델 관련 최신 기술 및 시범적 적용 사례에 대해서는 상세한 내용을 소개하지 못하였다. 또한, 데이터 수집 및 관리를 비롯하여 제도·정책적 이슈, 그리고 언어모델을 활용한 공공 서비스의 배포 및 운영 등 공공부문의 언어모델 활용과 관련한 실무적 측면에서의 고려 사항을 충분히 다루지 못하였다. 이러한 한계에도 불구하고, 본 연구가 언어모델의 개념과 주요 접근법에 대한 이해를 토대로 공공부문에서의 활발한 언어모델 활용에 기여할 수 있기를 바란다.

▣ 참고문헌

이창용. 2023. “인공지능의 개념 및 공공부문 활용 사례: 주요 접근법 소개 및 향후 연구 방향에 대한 제언”. ≪한국행정학보≫, 57(3): 395-425.

한국지능정보사회진흥원. 2023. ≪공공부문 거대언어모델(LLM) 오픈소스 활용방안≫.

한국지식재산연구원. 2023. ≪지정상품간 유사도 측정 방법론 개발≫.

행정안전부. 2021. ≪공공분야 인공지능 도입을 위한 실무자 안내서≫.

Achiam, Josh, Adler, Steven, Agarwal, Sandhini, Ahmad, Lama, Zoph, Barret, et al. 2024. “GPT-4 Technical Report”. arXiv, arXiv:2303.08774. *arXiv.org*.

Adiwardana, Daniel, Luong, Minh-Thang, So, David R., Hall, Jamie, Fiedel, Noah, Thoppilan, Romal, Yang, Zi, Kulshreshtha, Apoorv, Nemade, Gaurav, Lu, Yifeng, & Le, Quoc V. 2020. “Towards A Human-Like Open-Domain Chatbot”. arXiv, arXiv:2001.09977. *arXiv.org*.

Alayrac, Jean-Baptiste, Donahue, Jeff, Luc, Pauline, Miech, Antoine, Barr, Iain, Hasson, Yana, Lenc, Karel, Mensch, Arthur, Millican, Katie, Reynolds, Malcolm, Ring, Roman, Rutherford, Eliza, Cabi, Serkan, Han, Tengda, Gong, Zhitao, Samangooei, Sina, Monteiro, Marianne, Menick, Jacob, Borgeaud, Sebastian, Brock, Andrew, Nematzadeh, Aida, Sharifzadeh, Sahand, Binkowski, Mikolaj, Barreira, Ricardo, Vinyals, Oriol, Zisserman, Andrew, & Simonyan, Karen. 2022. “Flamingo: A Visual Language Model For Few-Shot Learning”. arXiv, arXiv:2204.14198. *arXiv.org*.

Anil, Rohan, Borgeaud, Sebastian, Alayrac, Jean-Baptiste, Yu, Jiahui, Vinyals, Oriol, et al. 2024. “Gemini: A Family Of Highly Capable Multimodal Models”. arXiv, arXiv:2312.11805. *arXiv.org*.

Bahrini, Aram, Khamoshifar, Mohammadsadra, Abbasimehr, Hossein, Riggs, Robert J., Esmaeili, Maryam, Majdabadkohne, Rastin Mastali, & Pasehvar, Morteza. 2023. “ChatGPT: Applications, Opportunities, And Threats”. arXiv, arXiv:2304.09103. *arXiv.org*.

Bojanowski, Piotr, Grave, Edouard, Joulin, Armand, & Mikolov, Tomas. 2017. “Enriching Word Vectors With Subword Information”. arXiv, arXiv:1607.04606. *arXiv.org*.

Brock, Andrew, De, Soham, Smith, Samuel L., & Simonyan, Karen. 2021. "High-Performance Large-Scale Image Recognition Without Normalization". arXiv, arXiv:2102.06171. *arXiv.org*.

Brown, Tom B., Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared, Dhariwal, Prafulla, Neelakantan, Arvind, Shyam, Pranav, Sastry, Girish, Askell, Amanda, Agarwal, Sandhini, Herbert-Voss, Ariel, Krueger, Gretchen, Henighan, Tom, Child, Rewon, Ramesh, Aditya, Ziegler, Daniel M., Wu, Jeffrey, Winter, Clemens, Hesse, Christopher, Chen, Mark, Sigler, Eric, Litwin, Mateusz, Gray, Scott, Chess, Benjamin, Clark, Jack, Berner, Christopher, McCandlish, Sam, Radford, Alec, Sutskever, Ilya, & Amodei, Dario. 2020. "Language Models Are Few-Shot Learners". arXiv, arXiv:2005.14165. *arXiv.org*.

Chen, Feilong, Han, Minglun, Zhao, Haozhi, Zhang, Qingyang, Shi, Jing, Xu, Shuang, & Xu, Bo. 2023. "X-LLM: Bootstrapping Advanced Large Language Models By Treating Multi-Modalities As Foreign Languages". arXiv, arXiv:2305.04160. *arXiv.org*.

Chen, Sanyuan, Wu, Yu, Wang, Chengyi, Liu, Shujie, Tompkins, Daniel, Chen, Zhuo, & Wei, Furu. 2022. "BEATs: Audio Pre-Training With Acoustic Tokenizers". arXiv, arXiv:2212.09058. *arXiv.org*.

Chiang, Wei-Lin, Li, Zhuohan, Lin, Zi, Sheng, Ying, Wu, Zhanghao, Zhang, Hao, Zheng, Lianmin, Zhuang, Siyuan, Zhuang, Yonghao, Gonzalez, Joseph E., & others. 2023. "Vicuna: An Open-Source Chatbot Impressing Gpt-4 With 90%* Chatgpt Quality". *See Https://Vicuna. Lmsys. Org (Accessed 14 April 2023)*, 2(3): 6.

Cho, Kyunghyun, van Merrienboer, Bart, Gulcehre, Caglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, & Bengio, Yoshua. 2014. "Learning Phrase Representations Using RNN Encoder-Decoder For Statistical Machine Translation". arXiv, arXiv:1406.1078. *arXiv.org*.

Choi, Jaewoong, & Lee, Byungju. 2024. "Quantitative Topic Analysis Of Materials Science Literature Using Natural Language Processing". *ACS Applied Materials & Interfaces*, 16(2): 1957-1968.

Chowdhery, Aakanksha, Narang, Sharan, Devlin, Jacob, Bosma, Maarten, Fiedel, Noah, et al. 2022. "PaLM: Scaling Language Modeling With

Pathways". arXiv, arXiv:2204.02311. *arXiv.org*.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, & Toutanova, Kristina. 2019. "BERT: Pre-Training Of Deep Bidirectional Transformers For Language Understanding". arXiv, arXiv:1810.04805. *arXiv.org*.

Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, Uszkoreit, Jakob, & Houlsby, Neil. 2021. "An Image Is Worth 16x16 Words: Transformers For Image Recognition At Scale". arXiv, arXiv:2010.11929. *arXiv.org*.

Engstrom, David Freeman, Ho, Daniel E., Sharkey, Catherine M., & Cuéllar, Mariano-Florentino. 2020. "Government By Algorithm: Artificial Intelligence In Federal Administrative Agencies". *SSRN Electronic Journal*.

Gao, Yunfan, Xiong, Yun, Gao, Xinyu, Jia, Kangxiang, Pan, Jinliu, Bi, Yuxi, Dai, Yi, Sun, Jiawei, Wang, Meng, & Wang, Haofen. 2024. "Retrieval-Augmented Generation For Large Language Models: A Survey". arXiv, arXiv:2312.10997. *arXiv.org*.

Harris, Sarah L., & Harris, David Money. 2016. "Digital Design And Computer Architecture". ARM® Edition, Amsterdam: Elsevier/Morgan Kaufmann.

Harris, Zellig S. 1954. "Distributional Structure". *WORD*, 10(2-3): 146-162.

Hochreiter, Sepp, & Schmidhuber, Jürgen. 1997. "Long Short-Term Memory". *Neural Computation*, 9: 1735-1780.

Hoffmann, Jordan, Borgeaud, Sebastian, Mensch, Arthur, Buchatskaya, Elena, Cai, Trevor, Rutherford, Eliza, Casas, Diego de Las, Hendricks, Lisa Anne, Welbl, Johannes, Clark, Aidan, Hennigan, Tom, Noland, Eric, Millican, Katie, Driessche, George van den, Damoc, Bogdan, Guy, Aurelia, Osindero, Simon, Simonyan, Karen, Elsen, Erich, Rae, Jack W., Vinyals, Oriol, & Sifre, Laurent. 2022. "Training Compute-Optimal Large Language Models". arXiv, arXiv:2203.15556. *arXiv.org*.

Hong, Suckwon, Kim, Juram, Woo, Han-Gyun, Kim, Young-Choon, & Lee, Changyong. 2022. "Screening Ideas In The Early Stages Of Technology Development: A Word2vec And Convolutional Neural Network Approach". *Technovation*, 112: 102407.

Houlsby, Neil, Giurgiu, Andrei, Jastrzebski, Stanislaw, Morrone, Bruna, de Laroussilhe, Quentin, Gesmundo, Andrea, Attariyan, Mona, & Gelly, Sylvain. 2019. "Parameter-Efficient Transfer Learning For NLP". arXiv, arXiv:1902.00751. *arXiv.org*.

Hsu, Wei-Ning, Bolte, Benjamin, Tsai, Yao-Hung Hubert, Lakhotia, Kushal, Salakhutdinov, Ruslan, & Mohamed, Abdelrahman. 2021. "HuBERT: Self-Supervised Speech Representation Learning By Masked Prediction Of Hidden Units". arXiv, arXiv:2106.07447. *arXiv.org*.

Hu, Edward J., Shen, Yelong, Wallis, Phillip, Allen-Zhu, Zeyuan, Li, Yuanzhi, Wang, Shean, Wang, Lu, & Chen, Weizhu. 2021. "LoRA: Low-Rank Adaptation Of Large Language Models". arXiv, arXiv:2106.09685. *arXiv.org*.

Hu, Zhiqiang, Wang, Lei, Lan, Yihuai, Xu, Wanyu, Lim, Ee-Peng, Bing, Lidong, Xu, Xing, Poria, Soujanya, & Lee, Roy Ka-Wei. 2023. "LLM-Adapters: An Adapter Family For Parameter-Efficient Fine-Tuning Of Large Language Models". arXiv, arXiv:2304.01933. *arXiv.org*.

Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. 1977. "Perplexity—A Measure Of The Difficulty Of Speech Recognition Tasks". *The Journal of the Acoustical Society of America*, 62(S1): S63–S63.

Jian, Yiren, Gao, Chongyang, & Vosoughi, Soroush. 2023. "Bootstrapping Vision-Language Learning With Decoupled Language Pre-Training". arXiv, arXiv:2307.07063. *arXiv.org*.

Kaplan, Jared, McCandlish, Sam, Henighan, Tom, Brown, Tom B., Chess, Benjamin, Child, Rewon, Gray, Scott, Radford, Alec, Wu, Jeffrey, & Amodei, Dario. 2020. "Scaling Laws For Neural Language Models". arXiv, arXiv:2001.08361. *arXiv.org*.

Le, Quoc V., & Mikolov, Tomas. 2014. "Distributed Representations Of Sentences And Documents". arXiv, arXiv:1405.4053. *arXiv.org*.

Lee, Gyumin, Lee, Sungjun, & Lee, Changyong. 2023. "Inventor–Licensee Matchmaking For University Technology Licensing: A fastText Approach". *Technovation*, 125: 102765.

Lee, Seunghyun, Choi, Jaewoong, & Yoon, Janghyeok. 2021. "A Social Media

Mining Approach For Monitoring Customer Complaints In The Early Stage Of Product Launch". *Journal of the Korean Institute of Industrial Engineers*, 47(3): 289-301.

Leskovec, Jurij, Rajaraman, Anand, & Ullman, Jeffrey D. 2014. "Mining Of Massive Datasets". Second edition, Cambridge: Cambridge University Press.

Lester, Brian, Al-Rfou, Rami, & Constant, Noah. 2021. "The Power Of Scale For Parameter-Efficient Prompt Tuning". arXiv, arXiv:2104.08691. *arXiv.org*.

Lewis, Patrick, Perez, Ethan, Piktus, Aleksandra, Petroni, Fabio, Karpukhin, Vladimir, Goyal, Naman, Küttler, Heinrich, Lewis, Mike, Yih, Wen-tau, Rocktäschel, Tim, Riedel, Sebastian, & Kiela, Douwe. 2021. "Retrieval-Augmented Generation For Knowledge-Intensive NLP Tasks". arXiv, arXiv:2005.11401. *arXiv.org*.

Li, Junnan, Li, Dongxu, Savarese, Silvio, & Hoi, Steven. 2023. "BLIP-2: Bootstrapping Language-Image Pre-Training With Frozen Image Encoders And Large Language Models". arXiv, arXiv:2301.12597. *arXiv.org*.

Li, Xiang Lisa, & Liang, Percy. 2021. "Prefix-Tuning: Optimizing Continuous Prompts For Generation". arXiv, arXiv:2101.00190. *arXiv.org*.

Lin, Chin-Yew. 2004. "Rouge: A Package For Automatic Evaluation Of Summaries". *In Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, 74-81.

Liu, Haohe, Tian, Qiao, Yuan, Yi, Liu, Xubo, Mei, Xinhao, Kong, Qiuqiang, Wang, Yuping, Wang, Wenwu, Wang, Yuxuan, & Plumbley, Mark D. 2023. "AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining". arXiv, arXiv:2308.05734. *arXiv.org*.

McInnes, Leland, Healy, John, & Astels, Steve. 2017. "Hdbscan: Hierarchical Density Based Clustering". *The Journal of Open Source Software*, 2(11): 205.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, & Dean, Jeffrey. 2013. "Efficient Estimation Of Word Representations In Vector Space". arXiv,

arXiv:1301.3781. *arXiv.org*.

Ouyang, Long, Wu, Jeff, Jiang, Xu, Almeida, Diogo, Wainwright, Carroll L., Mishkin, Pamela, Zhang, Chong, Agarwal, Sandhini, Slama, Katarina, Ray, Alex, Schulman, John, Hilton, Jacob, Kelton, Fraser, Miller, Luke, Simens, Maddie, Askell, Amanda, Welinder, Peter, Christiano, Paul, Leike, Jan, & Lowe, Ryan. 2022. "Training Language Models To Follow Instructions With Human Feedback". arXiv, arXiv:2203.02155. *arXiv.org*.

Papineni, Kishore, Roukos, Salim, Ward, Todd, & Zhu, Wei-Jing. 2001. "BLEU: A Method For Automatic Evaluation Of Machine Translation". *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL'02*, Philadelphia, Pennsylvania: Association for Computational Linguistics, 311.

Pennington, Jeffrey, Socher, Richard, & Manning, Christopher. 2014. "Glove: Global Vectors For Word Representation". *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 1532–1543.

Peters, Matthew E., Neumann, Mark, Iyyer, Mohit, Gardner, Matt, Clark, Christopher, Lee, Kenton, & Zettlemoyer, Luke. 2018. "Deep Contextualized Word Representations". arXiv, arXiv:1802.05365. *arXiv.org*.

Radford, Alec, Narasimhan, Karthik, Salimans, Tim, & Sutskever, Ilya. 2018. "Improving Language Understanding By Generative Pre-Training".

Raffel, Colin, Shazeer, Noam, Roberts, Adam, Lee, Katherine, Narang, Sharan, Matena, Michael, Zhou, Yanqi, Li, Wei, & Liu, Peter J. 2023. "Exploring The Limits Of Transfer Learning With A Unified Text-To-Text Transformer". arXiv, arXiv:1910.10683. *arXiv.org*.

Rombach, Robin, Blattmann, Andreas, Lorenz, Dominik, Esser, Patrick, & Ommer, Björn. 2022. "High-Resolution Image Synthesis With Latent Diffusion Models". arXiv, arXiv:2112.10752. *arXiv.org*.

Taori, Rohan, Gulrajani, Ishaan, Zhang, Tianyi, Dubois, Yann, Li, Xuechen, Guestrin, Carlos, Liang, Percy, & Hashimoto, Tatsunori B. 2023.

"Stanford Alpaca: An Instruction-Following LLaMA Model". *GitHub Repository*, GitHub.

Touvron, Hugo, Lavril, Thibaut, Izacard, Gautier, Martinet, Xavier, Lachaux, Marie-Anne, Lacroix, Timothée, Rozière, Baptiste, Goyal, Naman, Hambro, Eric, Azhar, Faisal, Rodriguez, Aurelien, Joulin, Armand, Grave, Edouard, & Lample, Guillaume. 2023. "LLaMA: Open And Efficient Foundation Language Models". arXiv, arXiv: 2302.13971. *arXiv.org*.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, & Polosukhin, Illia. 2017. "Attention Is All You Need". arXiv, arXiv:1706.03762. *arXiv.org*.

Wang, Jiuniu, Yuan, Hangjie, Chen, Dayou, Zhang, Yingya, Wang, Xiang, & Zhang, Shiwei. 2023. "ModelScope Text-To-Video Technical Report". arXiv, arXiv:2308.06571. *arXiv.org*.

Wang, Shuhe, Sun, Xiaofei, Li, Xiaoya, Ouyang, Rongbin, Wu, Fei, Zhang, Tianwei, Li, Jiwei, & Wang, Guoyin. 2023. "GPT-NER: Named Entity Recognition Via Large Language Models". arXiv, arXiv:2304.10428. *arXiv.org*.

Wang, Xuezhi, Wei, Jason, Schuurmans, Dale, Le, Quoc, Chi, Ed, Narang, Sharan, Chowdhery, Aakanksha, & Zhou, Denny. 2023. "Self-Consistency Improves Chain Of Thought Reasoning In Language Models". arXiv, arXiv:2203.11171. *arXiv.org*.

Wei, Jason, Wang, Xuezhi, Schuurmans, Dale, Bosma, Maarten, Ichter, Brian, Xia, Fei, Chi, Ed, Le, Quoc, & Zhou, Denny. 2023. "Chain-Of-Thought Prompting Elicits Reasoning In Large Language Models". arXiv, arXiv:2201.11903. *arXiv.org*.

Wei, Jason, Tay, Yi, Bommasani, Rishi, Raffel, Colin, Zoph, Barret, Borgeaud, Sebastian, Yogatama, Dani, Bosma, Maarten, Zhou, Denny, Metzler, Donald, Chi, Ed H., Hashimoto, Tatsunori, Vinyals, Oriol, Liang, Percy, Dean, Jeff, & Fedus, William. 2022. "Emergent Abilities Of Large Language Models". arXiv, arXiv:2206.07682. *arXiv.org*.

Yang, Chengrun, Wang, Xuezhi, Lu, Yifeng, Liu, Hanxiao, Le, Quoc V., Zhou, Denny, & Chen, Xinyun. 2024. "Large Language Models As

Optimizers". arXiv, arXiv:2309.03409. *arXiv.org*.

Yao, Shunyu, Yu, Dian, Zhao, Jeffrey, Shafran, Izhak, Griffiths, Thomas L., Cao, Yuan, & Narasimhan, Karthik. 2023. "Tree Of Thoughts: Deliberate Problem Solving With Large Language Models". arXiv, arXiv: 2305.10601. *arXiv.org*.

You, Haoxuan, Zhang, Haotian, Gan, Zhe, Du, Xianzhi, Zhang, Bowen, Wang, Zirui, Cao, Liangliang, Chang, Shih-Fu, & Yang, Yinfei. 2023. "Ferret: Refer And Ground Anything Anywhere At Any Granularity". arXiv, arXiv:2310.07704. *arXiv.org*.

Zhang, Duzhen, Yu, Yahan, Li, Chenxing, Dong, Jiahua, Su, Dan, Chu, Chenhui, & Yu, Dong. 2024. "MM-LLMs: Recent Advances In MultiModal Large Language Models". arXiv, arXiv: 2401.13601. *arXiv.org*.

Zhou, Yongchao, Muresanu, Andrei Ioan, Han, Ziwen, Paster, Keiran, Pitis, Silviu, Chan, Harris, & Ba, Jimmy. 2023. "Large Language Models Are Human-Level Prompt Engineers". arXiv, arXiv: 2211.01910. *arXiv.org*.

부록

〈표 A1〉 언어모델의 주요 작업에 대한 설명

대분류	소분류	설명
텍스트 전처리 (text pre-processing)	품사 태깅 (part of Speech tagging: POS tagging)	문장이나 문서에서 명사, 동사, 형용사, 부사, 대명사, 전치사 등 각 단어의 문법적인 역할(즉, 품사)을 태깅하는 과정
	토큰화 (tokenization)	문장이나 문서를 단어, 구문, 형태소 등의 작은 단위인 토큰으로 나누는 작업으로, 토큰은 문장, 단어, 형태소, 문자 등 텍스트를 나누는 최소 단위를 의미하며, 토큰화 과정은 기계학습 모델이 텍스트를 처리하고 이해하는 데 필수적임
	상호 참조 해결 (coreference resolution)	주어진 텍스트 내에서 동일한 대상을 참조하는 개체들을 식별하는 작업으로, 주로 대명사와 명명된 개체 등이 어떤 단어나 구에 참조되는지 파악하는 데 사용됨
	단어 의미 중의성 해소 (word sense disambiguation)	특정 단어가 여러 의미를 가질 때, 주어진 문맥에서 해당 단어의 정확한 의미를 결정하는 작업
	문법적 오류 해결 (grammatical error correction)	주어진 텍스트에서 발생한 문법 오류(예: 맞춤법, 구문, 동사-주어 일치, 시제, 대명사 오류)를 감지하고 수정하는 작업
텍스트 분류(text classification)	문서 분류 (document classification)	문서를 사전에 정의한 카테고리 혹은 클래스로 할당하는 작업으로, 정보 검색, 스팸 필터링, 뉴스 기사 분류 등 여러 목적으로 응용됨
	감성 분석 (sentiment analysis)	주어진 문장이나 문서의 감정을 파악하는 것으로, 단순하게는 긍정, 부정, 중립으로 분류하거나 더 나아가 문장의 어조를 식별
	가짜 뉴스 탐지 (fake news detection)	주어진 뉴스나 정보가 진실인지 혹은 가짜인지를 판별하는 작업으로, 텍스트

		내 언어적 특성, 문체, 단어 사용 패턴과 같은 텍스트 내부 특징과 미디어에서의 뉴스 전파 패턴, 사용자 의견 등의 외부 특징을 주로 사용함
	토픽 모델링 (topic modelling)	주어진 문서나 텍스트 코퍼스에 어떤 주제들이 존재하는지를 식별하고, 각 문서가 각 주제에 어떤 비율로 구성되어 있는지를 모델링하는 과정
정보 검색 (information retrieval)	문서 유사도 (document similarity)	주어진 문서들이 얼마나 비슷한 내용이나 주제를 다루는지를 정량적으로 파악하기 위한 과정으로, 문서의 벡터 산출 방식에 따라 코사인, 자카드, 유클리드, 헬링거 유사도 등의 방식으로 유사도를 측정함
	질의 응답 (question answering)	질문과 답변이 포함된 문맥이 함께 모델에 제공되어, 답변에 해당하는 토큰의 위치(인덱스)를 맞추는 추출형 질의응답(extractive QA)과 정확한 문맥 이해와 추론을 통해 답변을 생성하는 생성형 질의응답(generative QA)으로 나뉨
텍스트 생성(text-to-text generation)	기계 번역 (machine translation)	하나의 언어로 작성된 텍스트가 주어지면, 문맥 이해를 바탕으로 다른 언어로 자동으로 변환하는 작업으로 모델링 방식에 따라 규칙 기반, 통계 기반, 혹은 신경망 기반으로 나뉨
	텍스트 생성 (text generation)	기계가 주어진 문맥에 맞게 자연어로 텍스트를 생성하는 작업으로, 최근에는 언어모델을 통한 방식이 각광받고 있음
	텍스트 요약 (text summarization)	주어진 텍스트에서 중요한 문장이나 단어를 식별하여 요약하는 추출적 요약(extractive summarization)과 원문의 내용을 이해하고 새로운 문장을 생성하여 요약하는 추상적 요약(abstractive summarization)으로 나뉨
	어휘 정규화	텍스트의 다양한 언어적 변형, 축약, 줄

	(lexical normalization)	임말, 철자 오류 등을 통일하여 데이터를 일관된 형태로 만들어 처리하는 작업으로, 텍스트 데이터의 노이즈를 줄이고 일관성을 높이는 목적으로 활용됨
정보 추출 (information extraction)	개체명 인식 (named entity recognition)	문장 내에서 사람, 장소, 날짜, 조직 등의 카테고리에 해당하는 단어 혹은 명사구(개체)를 인식하는 방법인데, 구체적으로 개체의 범위를 식별하고 어떤 카테고리에 속하는지 분류
	관계 추출 (relation extraction)	주어진 텍스트에서 개체 간의 관계를 식별하고 분류하는 작업으로, 문장 수준에서의 관계식별 및 분류 외에도 문서 수준에서의 멀리 떨어진 개체 간 관계 매칭 및 증거 제공도 함께 다루기도 함
	사건 추출 (event extraction)	주어진 텍스트에서 특정 사건이나 이벤트를 인식하고 추출하는 작업으로, 주로 도메인에 따라 이벤트의 성격이 정의되는 편임
추론(inference)	상식 추론 (common sense reasoning)	주어진 텍스트에 내재된 상식적인 사실과 지식을 이용하여 특정한 가정을 세우거나 문맥을 이해하고 새로운 정보를 유추하는 작업
	자연어 추론 (natural language inference)	'전제(premise)'와 '가설(hypothesis)' 두 문장이 주어졌을 때, 전제 문장을 바탕으로 가설 문장이 함의, 모순, 중립 관계인지 분류하는 작업
챗봇(chatbots)	슬롯 채우기 (slot filling)	주어진 문장에서 특정 정보를 추출하여 미리 정의된 슬롯(slot)에 매핑하는 작업으로, 대화형 시스템, 예약 시스템 등의 분야서 응용됨
	발화 의도 예측 (intent detection)	주어진 문장이나 발화에서 사용자의 의도를 판별하는 작업으로, 주로 대화형 시스템이나 자연어 처리 기반 응용 프로그램에서 사용자의 의도를 파악하여 적절한 응답이나 작업을 수행하는 데 활용됨

이종 데이터 처리(text-to-data and vice versa)	텍스트-이미지 (text-to-image)	텍스트로 주어진 문맥이나 설명을 이해하고, 그에 맞는 이미지를 생성하는 작업으로 컴퓨터 비전과 자연어 처리 기술이 결합된 모델을 통해 수행함
	텍스트-음성 (text-to-speech 또는 speech-to-text)	텍스트를 음성 신호로 혹은 그 반대로 변환하는 작업을 일컬으며, 음성 검색 및 명령, 음성 안내 및 도우미, 자동 번역, 장애인 보조 기술 등의 목적으로 사용됨
	데이터의 텍스트화 (data-to-text)	표, 차트와 같은 구조화된 데이터로부터 이를 해석하고 요약하는 텍스트를 생성하는 작업으로, 시각화 설명, 금융 리포팅 등의 목적으로 사용됨

Language Models and Administrative Innovation in the Public Sector: Concept, Approaches, and Considerations

Jungwon Park & Gyumin Lee & Daeseong Jeon &
Jaewoong Choi & Changyong Lee

With the establishment of a framework for managing unstructured data in the public sector, research on administrative innovation through language models is gaining attention. However, there is a lack of discussion regarding significant problems that necessitate the use of language models, as well as potentially applicable methods. This study aims to bridge this gap by delineating the concept and approaches of language models and presenting key considerations for their deployment in the public sector. Initially, we elucidate the concept and approaches of language models, focusing on recent research trends centered around large language models. Subsequently, we present examples of language model applications in the public sector, with a particular emphasis on the science and technology sectors. Lastly, we explore considerations for their deployment from the perspectives of quality improvement and the expansion of application areas. This study is expected to not only stimulate research on administrative innovation in the public sector through the use of language models, but also enhance and diversify text analysis methods employed in the fields of public administration and policy studies.

※ Key words: language model, public sector, administrative innovation