

자연어 처리(NLP) 기술 동향 및 전망

KDB미래전략연구소 산업기술리서치센터
최 성 호 (sungho.choi@kdb.co.kr)

- ◆ 자연어 처리(NLP) 기술은 대규모 언어모델(LLM) 및 챗GPT(ChatGPT) 등장으로 많은 관심을 받으며 산업과 기업 전반의 AI 적용 및 확산에 중심점이 되고 있음
- ◆ LLM의 오픈소스화로 NLP "기술 생태계 전반의 성장 가속화"와 "개방성에 의한 위험 수반"이라는 양면성에 대해 향후 대처가 중요

□ NLP 기술은 LLM의 등장과 함께 각 산업계의 AI 활용영역 확대에 중요한 역할 담당

- NLP는 언어정보처리와 관련된 분야를 중심으로 빠르게 발전하고 있는 대표적인 인공지능 기술
 - NLP(Natural Language Processing, 자연어 처리)는 사람과 사람 간 지식, 정보 등을 전달하는 소통 수단인 언어를 컴퓨터가 이해할 수 있도록 변환하여 처리하는 인공지능의 한 분야임
 - NLP 기술은 주로 기계 학습과 딥러닝 접근법을 활용하며 대규모 텍스트 데이터를 학습하여 패턴을 인식하고 예측하는 모델을 구축함. 순환 신경망(Recurrent Neural Network, RNN*), 합성곱 신경망(Convolution Neural Network, CNN**), 생성형 사전훈련(Generative Pre-trained Transformer, GPT***) 등 다양한 모델이 사용되고 있음
 - * 인공신경망의 한 종류로, 유닛간의 연결이 순환적 구조를 갖는 특징
 - ** 심층 신경망 기법으로 행렬로 표현된 필터의 각 요소가 데이터 처리에 적합하도록 자동으로 학습되는 과정을 통해 이미지를 분류하는 기법
 - *** OpenAI社가 만든 인공신경망 기반의 대형언어모델
 - 음성 인식, 자동 번역, 대화체 질의응답, 시맨틱 웹 등 사람이 쓰거나 말하는 원리와 이해를 위한 핵심 기술로 활용되고 있으며, 최근 OpenAI 社의 ChatGPT가 널리 알려지며 관심이 증대되고 있음
- NLP 기술의 핵심인 언어모델(Language Model)은 LLM의 등장으로 산업 전반의 AI 확산에 중요한 역할
 - LM은 과거 통계적 모델(SLM)에서 인공신경망(트랜스포머) 모델로 발전하며 대규모 언어모델(Large Language Model, LLM)의 초석을 마련
 - LLM은 많은 양의 텍스트 데이터로 사전 훈련된 언어모델로서 NLP 작업 중 문장, 문단 또는 긴 텍스트의 흐름을 이해하고 문법, 의미, 문맥을 파악할 수 있는 특징을 갖고 있음

LLM이 NLP에 적용될 수 있는 주요 단계

구분	설명
언어 모델링 (Language Modeling)	주어진 문맥에서 다음 단어를 예측하는 언어 모델링 작업에 사용. 문장의 일부를 주면 LLM은 다음 단어를 예측하고 문장을 자연스럽게 완성
기계 번역 (Machine Translation)	입력 문장을 다른 언어로 번역하는 과정에서 LLM은 문맥을 고려하여 더 자연스러운 번역 결과를 생성
질의 응답 (Question Answering)	질문에 대한 답변을 생성하는 질의응답 시스템에 사용. LLM은 주어진 질문과 관련된 문맥을 이해하고 정확한 답변을 생성
텍스트 생성(Text Generation)	주어진 문맥에서 자연스러운 텍스트를 생성하는 작업에 사용. 소설, 시, 기사 등의 범주에서의 생성
요약 (Summarization)	긴 문서를 요약하는 작업에 사용. 입력된 문서의 핵심 내용을 파악하고 중요한 정보를 요약하여 더 간결한 형태로 제공
감정 분석(Sentiment Analysis)	텍스트의 감정을 분석하는 작업에 사용. 문장의 감정적인 톤을 이해하고 긍정적/부정적/중립적인지 등을 판별

자료 : AWS, Google Tech Blog, 당행 재작성

- LLM의 가장 성공적인 애플리케이션 중 하나인 트랜스포머(Transformer)* 모델은 AI에 인간의 언어를 가르치고, 소프트웨어 코드 작성, NLP 애플리케이션(번역, 챗봇, AI 비서 등)의 진화를 가속화하는 등 산업과 기업 전반에 AI가 활용되는 범위를 확대 중

* Attention 또는 Self-attention 이라 불리며 진화를 거듭하는 수학적 기법을 응용해 순차 데이터 내의 관계를 추적해 맥락과 의미를 학습하는 신경망(17년 Google의 논문에 첫 등장)

□ LLM의 오픈소스 공개로 NLP “기술 생태계 성장 가속화”와 “개방성에 의한 위협”이라는 양면성에 대한 대응이 중요할 전망

- 글로벌 AI 기업들의 LLM 오픈소스 공개로 관련 파생 서비스 및 애플리케이션을 만드는 스타트업이 증가함에 따라 자연어처리 기술 생태계의 성장 가속화 전망
- 또한, Google, OpenAI처럼 특정 기업이 독점하는 형태로 제공되는 LLM의 대안으로 부상되고 있음

글로벌 주요 LLM 오픈소스

기업명	오픈소스 프로젝트명(출시)
Meta	- LLaMA('23.2)
Databricks	- Dolly2.0('23.4)
Amazon	- Alexa Teacher Model, AlexaTM('22.8)
Stability AI	- StableLM-Alpha('23.4)

자료 : Github, 당행 재작성

- LLM 기술의 발달이 오픈소스라는 개방성에 의해 위협이 될 수 있음을 고려한 정책적, 사회적 대응 시급
- Meta가 공개한 LLaMA('23.2) 오픈소스가 성인용 챗봇(Allie)에 제한없이 사용된 사례와 같은 부정적 사용 방지 및 차단을 위한 논의가 시급