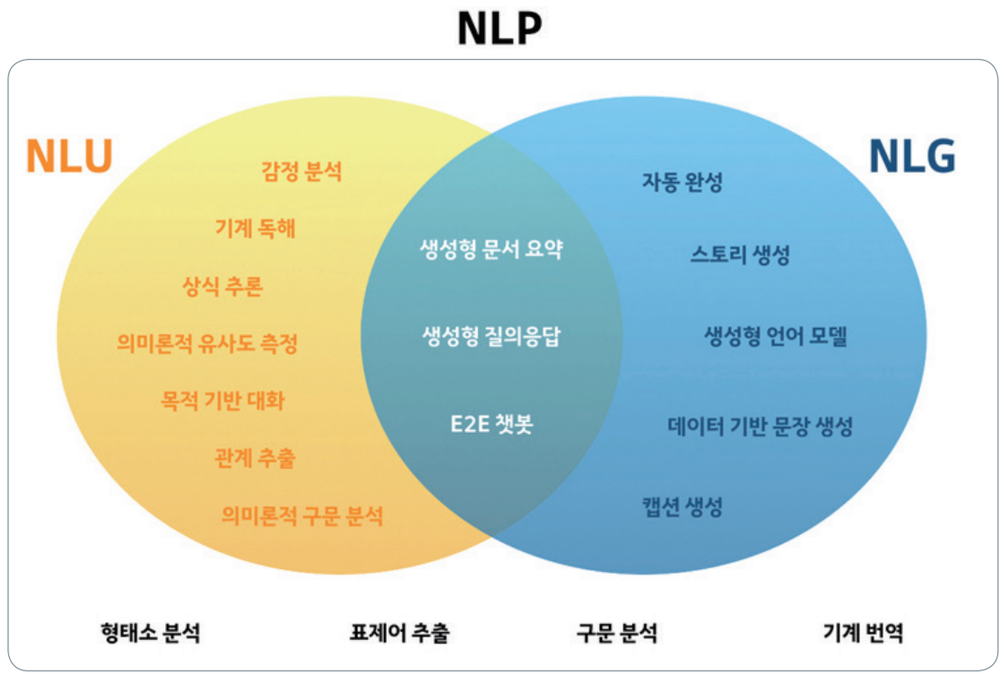
초거대 인공지능 언어모델 동향 분석

임수종
한국전자통신연구원 언어지능연구실 책임연구원
/이학박사
isj@etri.re.kr

Ⅰ. 서론

자연어처리(Natural Language Processing)는 인간의 언어 현상을 컴퓨터를 이용하여 모사할 수 있도록 연구하고 구현하는 인공지능의 주요 분야이다. 자연어처리는 기본적으로 언어를 이해하기 위해 형태소 분석, 개체명 인식, 어휘의미 분별, 구문분석, 의미역 인식, 상호참조 해결, 생략어 복원 등의 기술이 필요하다. 자연어처리 기술은 감정 분석, 기계 독해, 의미론적인 유사도 측정과 같은 자연어 이해(natural language understanding)와 자동완성, 스토리 생성, 캡션 생성과 같은 자연어 생성(natural language generation)으로 크게 나눌 수 있으며 생성형 문서 요약, 질의응답, 챗봇 대화 기술은 이 두 가지 기술이 모두 필요하다.

[그림 1] 자연어처리 기술 종류



자료: 카카오 브레인(https://www.kakaobrain.com/blog/118)

이러한 자연어처리를 구현하기 위해서 수학적, 통계적 방법을 많이 활용하며, 자연어처리는 기계학습 기법을 많이 사용하는 대표적인 분야이다. 기계학습 기법을 적용하려면 반드시 학습 데이터가 필요하며 특히 지도 학습(supervised learning)에 필요한 학습 데이터를 구축하려면 많은 시간과 비용이 필요하다. 2000년대 중반 이후로 영상 인식 분야에서 딥러닝 기법이 주목받기 시작하면서 이러한 학습 데이터 부족 현상은 심화되었다. 이런 데이터 부족을 해결하기 위해서 더 소량의 학습 데이터로 학습이 가능한 방법을 연구하거나, 상대적으로 구하기 쉬운 데이터를 사용하고 사전 학습(pre-training) 방법을 활용해 인공지능 언어모델을 구성한 후 이를 이용하는 전이 학습(transfer learning)을 연구하기도 하였다.[1)]

인공지능 언어모델은 대용량 텍스트에서 언어 이해 능력과 지식을 학습하는 것으로, 2020년 7월 OpenAI[2)]에서 발표한 GPT-3(Generative Pre-trained Transformers 3)를 발표한 이후 폭발적인 관심을 받고 있다. 초대형 컴퓨팅과 데이터로 생성된 GPT-3 학습모델을 활용하여 소량의 학습 데이터로 응용 태스크에 적용이 가능해진다. 이렇듯 인공지능 언어모델 기술은 매우 빠르게 개발되고 있으며, 기존 방식보다 성능이 좋고 학습 데이터 구축비용이 절감되는 장점이 있어 상용화에 대한 기대가 높다.

인공지능 언어모델 구축을 위해서는 양질의 대용량 텍스트 데이터, 전문적인 전처리 과정, 대규모 컴퓨팅 파워 등이 필요하기 때문에 연구 및 산업 활용을 위한 경쟁력을 확보하는 것은 어려워 보여, 소수의 대기업을 제외하고는 진입이 어렵다. 또한 알려진 것과는 다르게 특정 대형 언어모델은 저효율성, 실서비스 불가[3)] 등의 한계로 정부 주도의 대규모 인프라 구축에 사용하는 것이 실효성이 있는지를 두고 논란이 일고 있다.

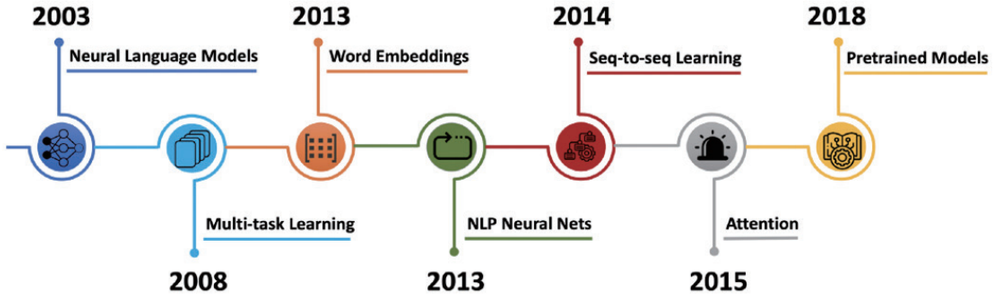
이러한 논란에도 불구하고 최근 글로벌 AI 기술 경쟁이 언어 분야로 집중되는데, 대형 사전 학습 모델의 효과성을 고려하면 이미 대용량 컴퓨팅 파워와 데이터를 확보한 글로벌 기업과의 기술 경쟁에서 한국어 처리 기술의 우위를 유지하기 위해 연구 개발이 필요하다. 현재 네이버 및 SKT, 카카오 등 대기업들은 자체적으로 한국형 GPT-3를 구축할 예정이다.

1) 임수종 · 김현기(2019), 「자연어처리를 위한 딥러닝 사전 학습 현황 및 한국어 적용 방안: 구글 BERT 사례를 중심으로」
2) https://openai.com/
3) GPT-3의 경우 많은 컴퓨팅 비용을 투입해도 우수한 작문 능력만 검증되기 때문에 이를 기반으로 하는 활용 서비스는 제약이 있는 것으로 알려졌다.

II. 딥러닝과 인공지능 언어모델

언어모델은 단어들의 열인 문장이 등장할 확률을 계산해 놓은 것이다. 문장을 구성하는 단어들이 나타날 확률을 알고 있다면 단어를 적절하게 선택하거나 문장을 순차적으로 생성해야 할 경우 여러 후보 단어 중에서 가장 그럴듯한(가장 확률이 높은) 단어를 선택하는 데 유용하다. 이는 특정 언어(이를테면 영어, 한국어)를 이해하고 말할 수 있다는 뜻이다. 언어모델은 충분한 양의 데이터 수집과 확률을 구하기 위해 방대한 양을 계산해야 했기 때문에 용도가 굉장히 제한적이었으나, 많은 양의 데이터 수집이 가능한 빅데이터와 컴퓨팅 파워의 증가로 딥러닝 시대에 접어들면서 각광을 받기 시작하였다. 2003년 초기 뉴럴 언어모델이 제안되었고 2013년에 간단한 언어모델인 워드 임베딩 모델 Word2Vec이 제안되었으며 전이 학습 개념이 소개되면서 2018년 이후로 사전 학습 기반의 딥러닝 인공지능 언어모델이 활발하게 연구 및 구축되고 있다.

[그림 2] 딥러닝 자연어처리 기술 흐름



자료: https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-2-f5e575e8e37

딥러닝 기반의 인공지능 언어모델은 대용량 텍스트 데이터에서 빈칸 단어나 다음 단어 맞히기 등의 자기 지도 학습(self-supervised learning)을 이용하는데, 이 방법으로 범용적 의미 표현을 사전 학습하고 다양한 응용 태스크에 활용할 수 있다.

이러한 대형 인공지능 언어모델은 앞에서 소개한 자연어처리 기술의 종류와 비슷하게 언어이해 모델(예: BERT), 언어생성 모델(예: GPT), 언어 이해 및 생성 모델(예: T5)로 구분되며 표 1과 같은 특징이 있다.

[표 1] 인공지능 언어 모델 유형별 특성

구분	언어이해 모델	언어생성 모델	언어 이해 및 생성 모델
개념	자연어 문장에서 단어의 주변 문맥(context)을 사전 학습하여 입력 문장에 포함된 단어의 문법과 의미를 이해	자연어 문장을 사전 학습하여 순서대로 주어진 단어 열에 가장 적합한 다음 단어를 예측하여 생성	언어 이해와 생성을 같이 사용하는 모델로 입력 문장을 이해한 결과를 바탕으로 출력 문장을 생성하는 모델
특징	활용 사례 및 후속 연구가 가장 활발한 모델	자동번역, 요약 같은 언어 생성 태스크에 가장 적합한 모델	언어 이해 및 언어 생성 모델을 모두 포함
학습방법	주변 단어를 이용하여 타겟 단어를 예측	이전 단어(들)를 기반으로 다음에 나올 단어를 예측	입력된 문장에 해당하는 문장을 출력

1 언어이해 모델

이 모델은 대용량 데이터에서 단어의 문맥을 사전 학습하여 입력 문장에 포함된 단어의 문법과 의미를 이해하는 모델로, 언어모델 중 활용 사례 및 후속 연구가 가장 많다. 기계 독해(Machine Reading Comprehension), 문서 분류(Document Classification), 언어분석(Language Analysis) 등 언어이해 유형의 다양한 태스크에서 우수한 성능을 보이며, 기본(base) 모델의 경우 실서비스 적용이 가능한 수준의 처리 시간을 보인다. 경량화 및 긴 길이 언어 모델 등과 같은 후속 연구가 다수 진행되고 있다. 그러나 언어이해 모델의 한계로 대화, 요약, 자동번역과 같은 생성형 태스크에 적용하기는 어려우며, 응용 태스크마다 태스크에 특화된 별도의 딥러닝 모델을 추가하기 때문에 별도의 학습이 필요하다. 이는 다중 작업 학습(Multi Task Learning, MTL)을 연구하는 데 단점으로 작용한다.

구글에서 2018년에 제안한 BERT는 가장 대표적인 언어이해 모델이다. 일반적으로 언어모델을 구축할 때는 문장 내에서 n개의 단어를 이용하여 n+1 번째 단어를 순차적으로 예측하도록 설계되어 있는데, BERT는 조금 다른 두 가지 방법론을 채택하여 신경망의 모든 레이어에서 전체 문맥을 확인하면서 언어모델을 학습하도록 하였다.[4)]

첫째는 문장 내에서 순차적으로 나올 단어가 아닌 임의로 등장하는 단어를 마스킹(masking)하고 이를 예측하는 Masked Language Model(MLM)을 채택하였다. 실제로 학습 과정에서 15%의 단어를 변경하였는데 15% 중 80%는 [MASK] 토큰으로 변경하고, 10%는 임의의 단어로 변경하였고, 나머지 10%는 변경하지 않은 채로 학습을 수행하였다. [MASK] 토큰은 사전 학습에서만 사용되었고, 특정 태스크를 위한 추가 학습(fine-tuning)에서는 사용되지 않는다.

둘째는 문장 내에서 단어 단위가 아닌 문장 단위로 학습 단위를 확장하고 두 문장을 동시에 입력하여 두 문장이 연속된 문장 여부를 학습하는 Next Sentence Prediction(NSP) 기법을 채택하였다. 학습 문장은 실제 다음 문장 50%, 관계가 없는 문장 50%로 구성되었다.

아래는 BERT에서 두 가지 방법을 사용하여 실제 학습할 때 문장의 예이다. MLM은 [MASK] 단어를 예측하고, NSP는 LABEL을 이진(binary)으로 예측하기 위해 학습한다.

Input = [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP] LABEL = IsNext

Input = [CLS] the man [MASK] to store [SEP] penguin [MASK] are flight ##less birds [SEP] LABEL = NotNext

BERT는 특정 태스크 처리를 위해 새로운 신경망을 추가할 필요 없이 모델 자체의 추가 학습을 통해 해당 태스크의 State-Of-The-Art(SOTA) 성능을 달성하였다. 또한 사전

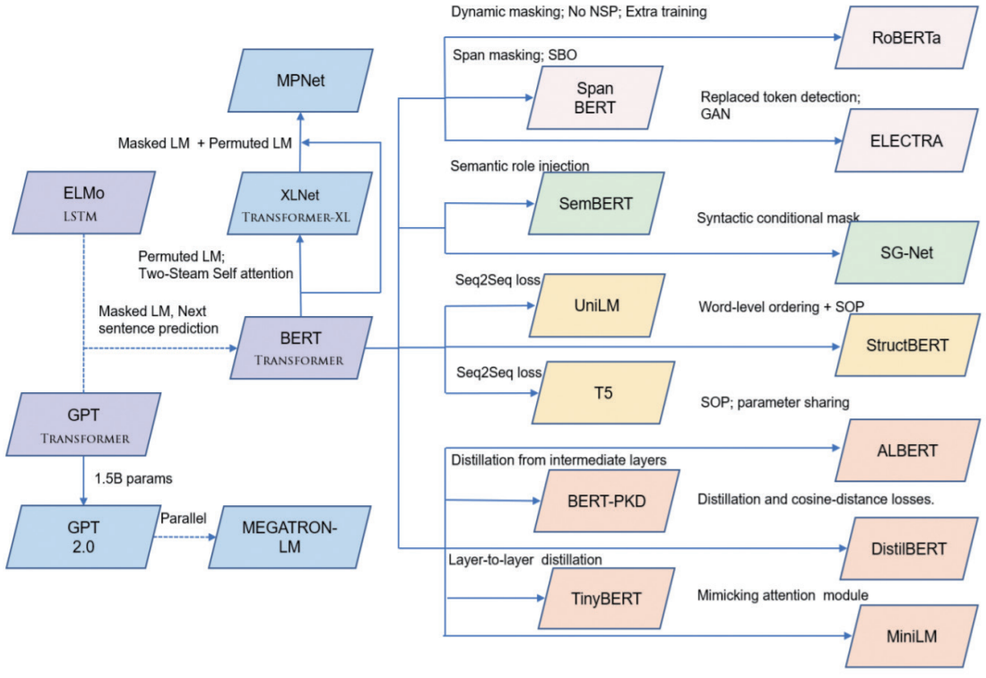
4) Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.(2018), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

학습과 추가 학습 시 구조(architecture)를 다르게 하여 전이 학습을 용이하게 하였다. 그러나 학습 방법이 복잡하기 때문에 기존의 사전 학습 방법보다 더 대용량의 학습 데이터가 필요하고 이로 인해 학습 시간도 증가된다.

2019년 10월에 구글에서 검색 결과를 재순위화하는 태스크에 BERT 모델을 적용하였다. 국내에서는 ETRI, SKT, 투블럭AI 등에서 한국어 BERT 기반 사전 학습 모델을 공개하였고, 이는 KT 기가지니, 스캐터랩 등에서 기업 고객 센터 및 챗봇에 활용된다.

BERT 모델이 SOTA 성능을 달성한 이후, 그림 3과 같이 BERT를 개선하여 더 좋은 모델을 구축하려는 노력이 구글의 경량화 모델인 ALBERT뿐만 아니라, 페이스북의 RoBERTa, Allen 인공지능 연구소의 SpanBERT, 스탠퍼드대학교의 ELECTRA에서도 활발하게 이루어지고 있다.

[그림 3] BERT 이후 딥러닝 사전학습 언어모델 연구 동향(Zhuosheng, 2020)



자료: https://arxiv.org/pdf/2005.06249.pdf

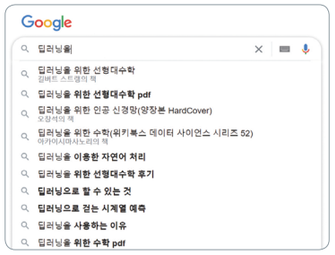
2 언어생성 모델

이 모델은 대용량 데이터를 미리 학습하여 주어진 단어 열에 가장 적합한 다음 단어를 예측하는 모델로 이미 알고 있는 이전 단어를 이용하여 어떤 단어가 가장 좋은지 예측할 수 있도록 신경망으로 다음 수식과 같이 학습하는 모델이다.

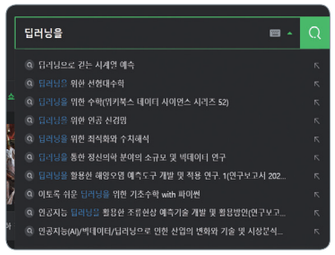
$$L_1(u) = \sum_i \log P(u_i | u_{i-k}, ..., u_{i-1}; \Theta)$$

이러한 언어생성 모델은 단어나 문장을 적절하게 선택하거나 생성해야 하는 여러 후보 중에서 더 좋은(확률이 높은) 후보를 선택하는 것을 가능하게 하며, 이를 간단히 응용한 예를 들면 다음과 같이 검색 서비스에서 자동완성을 하는 기능이다. 사용자가 "딥러닝을"까지 입력했을 때 각 검색 서비스가 구축한 한국어 언어생성 모델을 이용하여 가장 확률이 높은 문장이나 단어 열을 제시하는 방식이다.

[그림 4] 언어생성 모델을 이용한 자동완성 기능



자료: www.google.com



자료: www.naver.com

자료: www.daum.net

언어생성 모델은 입력된 언어를 이해한 후 새로운 언어로 생성해야 하는 자동번역, 문장 요약, 대화나 챗봇, 스토리 생성 등의 언어생성 유형의 태스크에서 잘 작동한다. 이 모델로는 OpenAI의 GPT 계열의 모델, CMU(Carnegie Mellon University)와 구글 브레인이 공동으로 개발한 XLNet, 페이스북의 BART(Bidirectional and Auto-Regressive Transformer) 등이 있다.

언어생성 모델의 단점은 언어이해 유형의 태스크에서 언어이해 모델보다 성능이 낮다는 점이다. 또한 생성 문장의 사용자 제어 가능성(controllability) 및 사실 일관성(factual consistency) 유지 관점에서 후속 연구가 필요한데, 특히 최근 '이루다' 사태로 AI 생성 문장의 편향성에 대한 사회적 관심이 높아졌다. 또한 알려진 n개의 단어로 n+1 번째 단어를 생성하도록 학습되어 있기 때문에 입력 문장에 대해 단방향 연산만 가능한 구조적 한계가 있다.

국내에서는 SKT와 아마존이 협업하여 한국어 GPT-2(KoGPT-2)를 공개하였고, Chit-chat 챗봇 서비스 자동 요약과 같은 언어생성 태스크를 연구 중이지만 문장 완성도에 한계점이 있어 상용화를 위해서는 정교한 후처리 작업이 필요하다.

3 언어 이해 및 생성 모델

이 모델은 언어 이해 모델과 언어 생성 모델을 같이 사용하는 모델로 입력 문장을 이해하여 출력 문장을 생성한다. 이 모델을 학습하려면 입력 문장을 이용하여 정답인 문장을 출력하는 방식을 써야 한다. BERT의 타깃 단어 예측과 GPT의 다음 단어 예측 모두 적용 가능한 방법으로, 실험 결과는 타깃 단어 예측이 더 우수한 성능을 보인다.

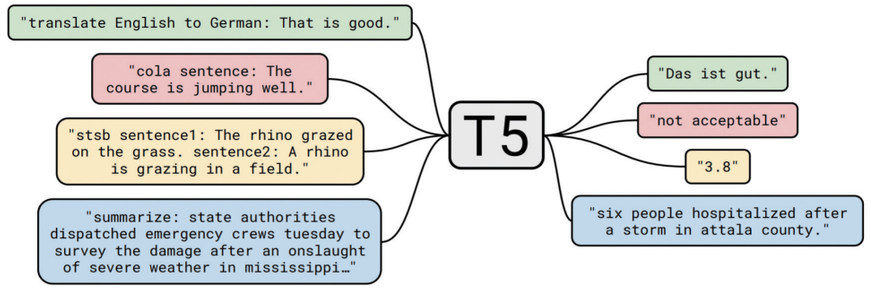
언어이해 태스크 및 생성 태스크 모두에서 BERT 및 GPT와 비슷한 수준의 성능을 나타내는데, 언어생성 태스크 적용이 어려운 BERT와 언어이해 태스크에서는 낮은 성능을 나타내는 GPT가 이 단점을 보완하여 두 방법의 좋은 대안으로 떠오르고 있다. 이는 범용적 언어처리 입력-출력 프레임워크로 BERT와 달리 응용 태스크마다 별도의 출력 레이어가 불필요하며 자연스럽게 다중 작업 학습(MTL)이 가능하다.

언어이해 결과에 기반한 언어생성 모델이기에 이해와 생성을 동시에 학습하지만 BERT나 GPT와 같은 단일 모델 대비 10%가량의 추가 연산만 필요하다. 이 추가 연산은 다음 단어 예측에 중요한 입력 문장 단어 찾기 연산이다.

최근 언어모델을 지식 저장 도구, 지식/상식 추론 도구로 활용하는 추세인데, 다수 연구에서 구글의 T5(Text-To-Text Transfer Transformer) 모델을 활용하고 있다. T5는 번역, 요약, 질의응답 등의 자연어처리 태스크가 자연어 문장을 입력하면 자연어 문장이 출력이 되는 형태이기 때문에 이를 'Text-to-Text' 태스크로 간주하고 입출력에 제약이 없는 모델을 개발하여 공개하였다.[5)]

언어모델로 특정 태스크를 수행할 때 각 태스크를 위해 추가 학습을 한 모델은 독립적으로 존재하는 언어이해 모델과 달리 언어 이해 및 생성 모델은 태스크를 특정하는 문장(a task-specific prefix)을 입력 문장의 앞에 추가하여 구분하면 하나의 언어모델로 여러 가지 태스크를 동시에 수행하는 MTL이 가능하다. 그림 5는 영어-독일어 자동번역, 문장의 비문 여부, 문장 유사도 측정, 요약 등의 서로 다른 태스크를 한 개의 T5 언어 모델로 수행하는 예를 보여 준다.

[그림 5] 구글 T5 text-to-text framework diagram



자료: https://arxiv.org/pdf/1910.10683.pdf

Ⅲ. 초거대 언어모델과 AI 패러다임 변화

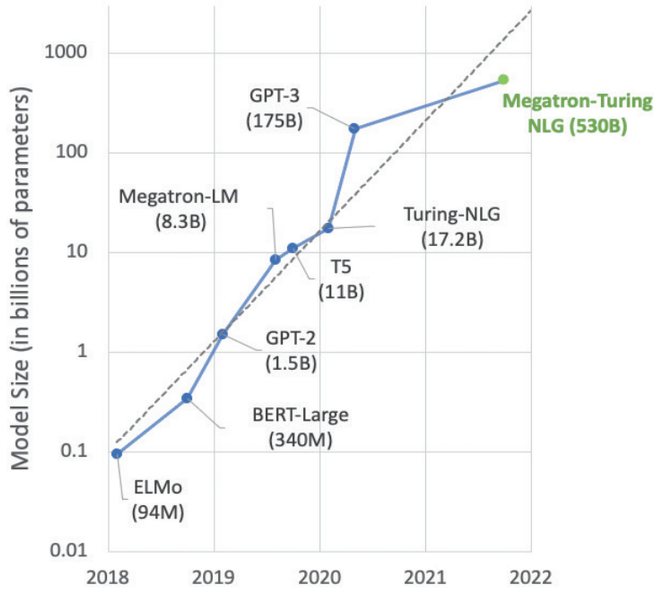
초거대 언어모델이란 언어생성 모델과 동일한 모델로 다음 단어 예측 정확도가 크게 향상되고, 대용량 연산이 가능한 인프라를 기반으로 데이터를 스스로 학습해 사람처럼 스스로 사고할 수 있도록 설계된 언어모델이다. 초거대는 언어모델을 학습하기 위해 필요한 매개변수의 수를 말하며, 보통 1,000억(100B)을 넘을 경우 초거대 언어모델이라고

5) Colin Raffel et al. (2019), Exploring the limits of transfer learning with a unified text-to-text transformer

분류하는데, 이는 현시점의 기준일 뿐이고 학습 가능한 데이터양과 컴퓨팅 파워가 증가하면서 얼마든지 변할 수 있다.

2020년 7월 GPT-3가 퓨샷 러닝의 가능성을 열면서 최근까지 초거대 언어모델에 대한 경쟁이 불붙었는데 현재는 약 5,300억(530B) 개 매개변수가 최대 모델이지만 구글이 1조(1 trillion) 개 매개변수를 갖는 Switch-C 언어모델을 구축하는 것으로 알려졌다.[6] 국내에서도 2020년 4월 SKT가 아마존웹서비스(AWS)와 협력하여 GPT-2에 상응하는 한국어 버전인 KoGPT-2를 공개하였고, 네이버는 매개변수 개수에서 GPT-3를 능가하는 204B 규모의 CLOVA 언어모델을 개발했다고 발표하였다. LG와 KT 또한 2021년 하반기에 유사한 규모의 인공지능 언어모델을 공개할 예정이라고 밝혔다.

[그림 6] 딥러닝 인공지능 언어모델 크기 변화



자료: https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

6) https://venturebeat.com/2021/01/12/google-trained-a-trillion-parameter-ai-language-model/

1 GPT-3

미국 OpenAI에서 개발한 언어모델로 대용량 데이터를 대규모 컴퓨팅 자원을 활용하여 학습함으로써 범용 인공지능의 가능성을 제시한 모델이다. 기존 자사 모델인 GPT-1, GPT-2와 동일한 방법론을 사용하였으나 Microsoft에서 제공한 상위 5위 수준의 슈퍼컴퓨터[7]를 이용하여 학습데이터의 양 및 딥러닝 모델의 크기를 대폭 증가 시켰다. 학습 방법은 언어생성 모델과 같이 이전 단어(들)을 보고 다음 단어를 예측하는데, 단어별 실수 개수 및 계산량을 기존 GPT-2 대비 약 116배 증가시켰다. 최적화 단계에서 10여 개의 샘플만으로 응용 태스크 적용이 가능한 퓨샷 학습(few shot learning)을 선보였는데, 이것은 별도의 추가 학습 없이 응용 태스크에 적용이 가능하여 범용 인공지능의 가능성을 제시하였다. 실제로 이것은 사람이 구분하기 어려운 수준의 글쓰기 성능을 보이며, 상식 추론 테스트(TriviaQA, PhysicalQA) 등에서 기존 모델 대비 우수한 성능을 보인다.

퓨샷 학습은 n개의 단어로 n+1 번째 단어를 예측하는 기존 생성 모델에 다음 식과 같이 특정 태스크 $task_{name}$을 추가 조건으로 하는 조건부 언어모델(conditional language model, CLM)을 사용하면 가능해진다.

$$L_1(u) = \sum_i \log P(u_i | task_{name}, u_{i-k}, ..., u_{i-1}; \Theta)$$

사전 학습 단계에서 기존 생성 모델처럼 일반적인 자연어 현상만을 학습하는 것이 아니라, 각 태스크 이름과 같은 조건을 통해서 태스크 고유의 자연어 현상도 함께 학습하는데, 이를 문맥 내 학습(in-context learning)이라고 한다. 문맥 내 학습을 하지 않는 태스크에도 적절한 태스크 명이 붙거나 태스크에 대한 설명이 입력될 경우, 기존에 학습된 태스크의 자연어 현상에서 유추하여 퓨샷 학습이 가능하다.

그러나 태양에 몇 개의 눈이 있는가 하는 질문에 한 개의 눈이 있다고 답변하는 상식이

7) 285,000CPU cores, 10,000 GPUs, 400Gbps network 사용

결여된 텍스트 생성과 미국 프로야구(MLB) 월드시리즈 우승팀이 결정되지 않은 시점에서 우승팀을 묻는 질문에 최다 우승팀인 뉴욕 양키스라고 대답하는 닫힌 지식은 언어 생성 모델(Language Generation Model)의 한계점이다. 학습모델은 가짜 뉴스 범죄에 악용될 것을 우려하여 미공개하였으나 2020년 9월 Microsoft와 독점 라이선스를 체결하고 유료 API 서비스를 시작하였다.

실제 추론으로 퓨샷 학습을 해결하였는지 사전 학습 시의 패턴 인식 결과를 활용하였는지 불확실해서 들어가는 비용에 비해 예측되는 한계를 알 수 없다. 또한 데이터가 제한된 한국어 모델 학습에서도 영어 수준의 성능이 가능할지 고려해야 한다. 초대형 컴퓨팅과 대규모 데이터로 기술 한계를 극복하려고 하지만 1,750억 개의 매개변수를 한 번 훈련하는 데 약 50억 원의 비용이 든다고 하니, 비용 문제가 기술 확산 및 보급에 걸림돌이 될 것으로 예상된다.

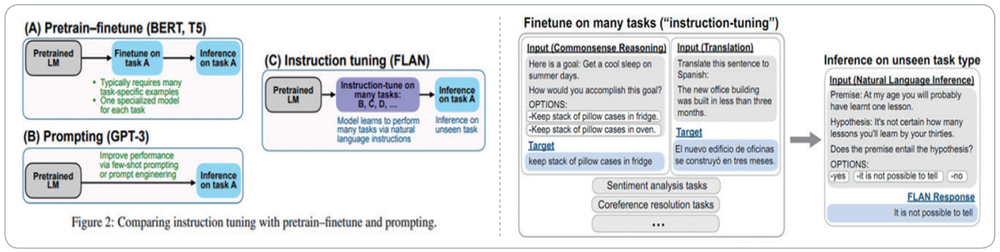
2 MT-NLG(Megatron-Turing Natural Language Generation)

Microsoft 와 NVIDIA가 공동으로 초거대 언어모델을 개발하였다. 매개변수 개수는 5,300억 개이고 기존 동일 유형 모델보다 3배 크며 레이어 수는 105개이다. 이 모델은 MRC와 상식 추론에서도 좋은 성능을 보인다. 대규모 AI 모델 훈련의 효율성을 높이는 방법을 제시하였다. 또한 각 기업이 보유한 슈퍼컴퓨터 인프라에 분산 학습(distributed learning) SW stack을 통합해 사용한다. 이 모델은 left-to-right transformer로 전형적인 생성 모델 구조이다. 이 컴퓨팅 장치들의 잠재력을 최대한 끌어내기 위해 MS Megatron 언어모델과 NVIDIA deepspeed SW를 접목하였다. 연구 그룹 엘레우테르AI(EleutherAI)가 오픈소스로 제공하는 22개 소규모 데이터세트로 이루어진 총 835GB 규모의 더파일(The Pile)과 인터넷상 크롤링 데이터를 결합해 2,700억 개 토큰(token)으로 구성된 훈련 데이터로 학습하였다.

3 FLAN(Finetuned LAnguage Net)

구글이 2021년 10월 GPT-3의 문맥 내 학습을 이용한 퓨샷 학습에 대응하고자 명령어 조정 학습(instruction learning)이라는 개념을 제시하였다. GPT-3의 문맥 내 학습이 사전 학습 시 태스크에 대한 문맥을 같이 학습하는 개념이라면, 구글의 명령어 조정 학습은 사전 학습 종료 후 각 태스크에 맞는 명령어를 통해 추가로 학습을 하는 방법이고, 추가 학습을 통해 특정 태스크에 최적화된 언어 모델을 구축하는 방법과 다르게 단일 모델에 다른 종류의 태스크를 수행할 수 있다.

[그림 7] 구글 FLAN 모델



자료: Jason W.(2021), https://arxiv.org/pdf/2110.05679.pdf

FLAN은 GPT-3에 비해 적은 양인 137B의 매개변수를 이용하여 학습을 하였으나 20~25개의 자연어 태스크를 문맥 내 학습한 GPT-3보다는 많은 60개의 태스크를 명령어 조정 학습하여 퓨샷 학습에 더 가까이 접근하였다. 이를 증명하고자 논문은 60개의 명령어 조정 학습에 포함되지 않은 자연어 태스크에서 더 나은 성능을 보였다고 주장한다.

4 AI 패러다임 변화

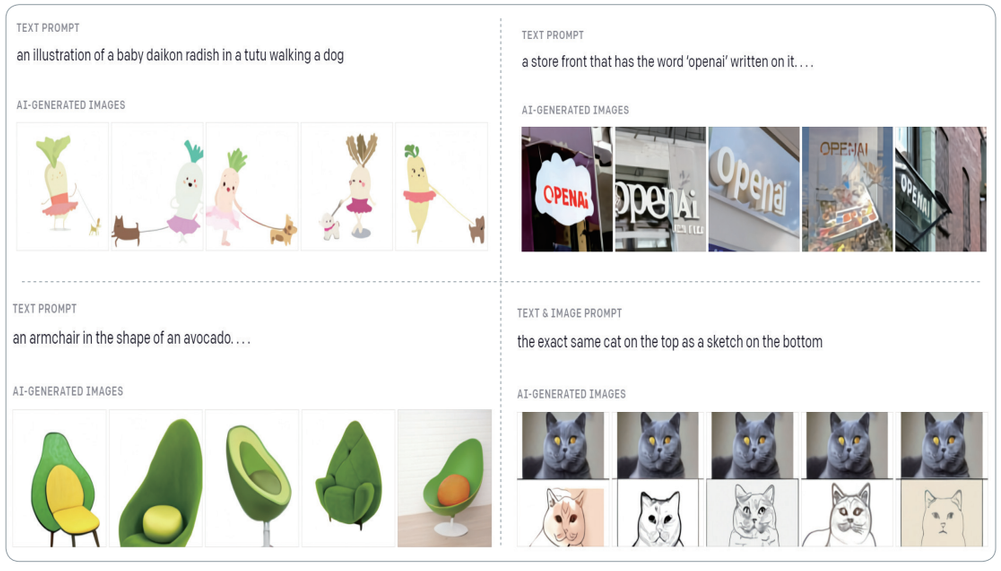
딥러닝 기반 자연어처리의 성능 향상을 위해 기존 언어모델에 딥러닝 방식을 적용한 인공지능 언어모델은 현재 구글, 페이스북 같은 거대 글로벌기업뿐만 아니라 네이버, 카카오와 같은 국내 유수 기업도 중요한 기술로 인식하고 전사적인 리소스를 투입하여 개발에 박차를 가하고 있다. 인공지능 모델은 우리 생활과 산업에 활용되며 기술 혁신을 가속화하는데, 그 결과도 인간의 지능과 비슷하여 일반인도 인공지능에 관심을 가지고 흥미로운 결과를 내며 경쟁하고 있다.

초거대 언어모델은 두 가지 방향으로 새로운 AI 패러다임을 만들어 나간다. 첫째는 특정 태스크를 위해 학습 데이터가 거의 필요 없는 few-shot, one-shot, 학습 데이터가 전혀 필요

없는 zero-shot 학습이 가능하다는 점이다. GPT-3는 제한된 태스크에서만 이 학습 방법이 우수한 성능을 보이지만, 학습하지 않은 미지의 태스크까지 확장이 가능한 다음 버전은 범용 인공지능(Artificial General Intelligence, AGI)의 시작이 될 것이다. 현재 인공지능 기술은 대부분 특정한 문제를 해결하는 데 집중된 반면 AGI는 어떤 문제를 사고하고 해결할 수 있는 인간 수준 혹은 그 이상의 범용적인 지적 능력을 갖춘 인공지능이다. 이러한 초거대 언어모델은 zero-shot 학습으로 AGI의 실현 가능성을 보여 준다.

둘째는 언어모델이 단순히 텍스트를 대상으로 하는 자연어처리 영역에서만 효과를 보이는 것이 아니라 음성, 이미지, 촉각, 통각, 센서 데이터 등의 데이터와 융합하여 인공지능 적용 대상을 무한대로 확장할 수 있다는 점이다. 초거대 언어모델은 언어를 생성하는 태스크에 사람이 보기에도 자연스러운 문장 생성이 가능하고 최근에는 언어와 이미지를 동시에 사전 학습하여 텍스트에서 이미지 생성이 가능해졌다. 자연어 문장인 “openai라고 쓰인 가게의 정면”이라고 입력하면 출력은 그림과 같이 기존 방법으로 자연어 문장에 해당하는 이미지를 검색해서 보여 주는 것이 아닌 해당하는 이미지를 직접 ‘생성’한다.

[그림 8] OpenAI DALL-E



자료: https://openai.com/blog/dall-e/

Ⅳ. 결론

지도 학습에 필요한 학습 데이터양을 줄이기 위해 시작된 딥러닝 기반 인공지능 언어모델은 초기 버전의 Word2vec부터 진화를 거듭하였다. 현재는 추가적인 학습 데이터 없이 태스크 수행이 가능한 zero-shot과 텍스트, 이미지까지 포함한 다양한 데이터와 융합하여 사실상 기술의 도메인 경계를 무너뜨리는 비전을 보여 주며 범용 인공지능의 지평을 열고 있다. 딥러닝 시대 이전에는 분야별 데이터 특성이 있어 높은 칸막이로 막혀 있었다. 이제는 특정 분야에 특화된 기술이 점점 허물어지고 있기 때문에 이러한 초거대 언어모델뿐만 아니라 AGI를 향해 가는 기술 동향도 분석하여 주로 숫자 데이터로 구성된 국가 통계 분야의 높은 칸막이를 없애는 것이 필요하다. 국가의 텍스트, 이미지를 포함한 모든 데이터를 통계의 영역으로 적극적으로 가져온다면 한 단계 발전한 국가 통계 서비스를 구축할 수 있을 것이다.

참고문헌

임수종 · 김현기(2019). 「자연어처리를 위한 딥러닝 사전 학습 현황 및 한국어 적용 방안: 구글 BERT 사례를 중심으로」, 「디지털문화아카이브지」, 2(2), 111-118.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu(2019), Exploring the limits of transfer learning with a unified text-to-text transformer, arXiv preprint arXiv:1910.10683, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.(2018), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, ArXive-prints: 1810.04805. 2018.10.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V. Le(2021), Finetuned language models are zero-shot learners, arXiv preprint arXiv:2109.01652, 2021.