

AWS SageMaker 사용 전 배경지식 조사

000부서 000

- 목차 -

1. 문서 작성 배경
2. 파운데이션 모델 연관 지식
3. 기계번역과 인공지능
4. Sequence-to-Sequence와 어텐션(Attention)
5. 트랜스포머 모델(Transformer Model)
6. 파운데이션 모델(Foundation Model)의 정의
7. 파운데이션 모델의 특징
8. 대규모 언어 모델(Large Language Model)
9. 인코더-디코더 모델 vs 디코더 모델
10. 기계번역 현황과 전망
11. 고찰
12. 참고문헌

문서 작성 배경

이후 문서에서 다룰 AWS SageMaker의 사용 이전에 관련한 배경지식을 조사한 문서입니다. AWS SageMaker는 기계학습의 훈련, 운영 및 배포를 도와주는 도구인 MLOps입니다. SageMaker가 가진 수 많은 기능 중 가장 핵심적으로 사용할 기능은 AWS SageMaker의 파운데이션 모델(Foundation Model)을 사용할 수 있는 기능인 JumpStart입니다.

JumpStart를 사용하기에 앞서 파운데이션 모델의 정의와 기계번역의 현황에 대해서 알아보도록 하겠습니다.

파운데이션 모델 연관 지식

AWS SageMaker JumpStart는 파운데이션 모델(Foundation Model, FM) 간단한 클릭 몇 번으로 다루게 해주는 기능을 제공합니다. 파운데이션 모델은 트랜스포머 모델(Transformer Model)과 대규모 언어 모델과 같이 방대한 데이터로 학습된 동시에 다양한 가능성을 가지고 새로운 발전의 기반이 될 수 있는 모델들의 범주를 정의하는 용어입니다. 파운데이션 모델을 이해하기 위해서는 트랜스포머 모델에 대해서도 어느정도 이해할 필요가 있습니다. 트랜스포머 모델을 이해하기 위해서는 어텐션(Attention)과 셀프어텐션(Self-Attention)을 추가적으로 알아야 합니다. 이렇게 꼬리에 꼬리를 무는 형태의 용어들의 정의를 하나씩 순차적으로 짚어 보고자 합니다.

기계번역과 인공지능

우선 짧게 기계 번역의 역사를 살펴보겠습니다. 최초의 기계번역의 개념 등장 이후, 단순 단어의 전환에서부터 규칙 기반, 통계 기반, 신경망 기반 기계번역으로 발전해왔습니다. 통계 기반 기계번역 단계에서는 과거의 데이터 속 의미를 찾는 것이었다면, 현재 단계인 신경망 기반 기계번역은 번역 내용을 '예측'하는 것입니다. 이는 기존 통계분석에서 인공지능으로 진화한 과정과 일맥상통합니다.

Sequence-to-Sequence와 어텐션(Attention)

기존 통계 기반 기계번역의 한계를 극복하고자 딥러닝과 기계번역이 결합되어 **신경망 기반 기계번역(Neural Machine Translation, NMT)**이 태어났습니다. 가장 먼저 등장한 신경망 기반 기계번역은 **Sequence to Sequence** 모델입니다. 해당 모델은 seq2seq라고도 불리며 Recurrent Neural Network(이하 RNN) 기반의 인코더-디코더(Encoder-Decoder)모델의 형태를 하고 있습니다. RNN은 정보를 순환시키는 모델이기 때문에 데이터의 순차를 인식하고 패턴을 수색합니다. 이 때문에 순차성 데이터를 사용하는 텍스트 관련 인공지능 분야에서도 초기에는 RNN을 자주 활용하였습니다.

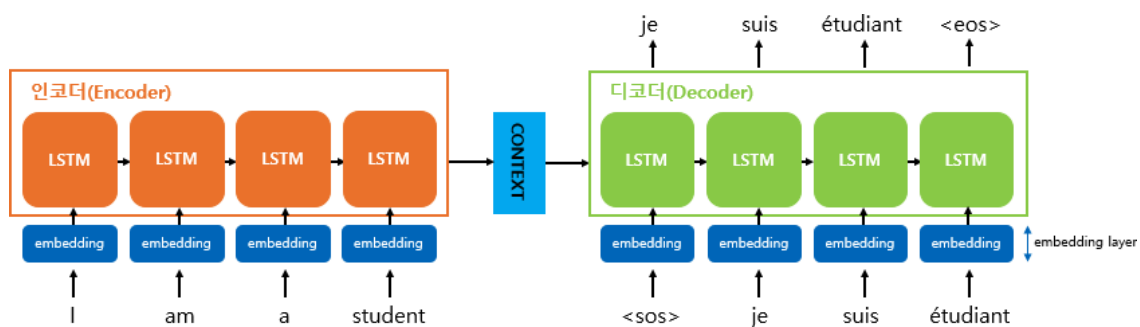


그림 출처 : 위키독스 - 『딥 러닝을 이용한 자연어 처리 입문』<https://wikidocs.net/book/2155>

Seq2seq 모델에는 두 가지의 큰 문제점이 있습니다. 첫 번째 문제는, 단일의 고정된 용량의 벡터에 모든 정보를 압축하는 메커니즘으로 인해 **정보 손실**이 발생한다는 것입니다. 두 번째로는 RNN모델의 고질적인 문제인 **기울기 소실**(vanishing gradient)의 발생입니다.(기울기 소실이 발생하면 모델의 학습이 제대로 진행되지 않습니다.)

해당 문제는 번역품질 저하로 이어지는데 이를 극복하고자 한 생겨난 기법이 **어텐션(Attention)**입니다. 어텐션은 디코더에서 단어를 출력할 때 시점(step)마다, 인코더에 입력된 전체 문장을 다시 참고하게 되는데, 출력(예측)할 단어와 연관성이 있는 입력 단어를 집중(attention)해서 참고합니다.

트랜스포머 모델(Transformer Model)

트랜스포머 모델(Transformer Model)은 앞선 seq2seq모델의 구조(인코더-디코더)를 유지하면서 RNN을 사용하지 않고 어텐션을 적극활용한 모델입니다. 어텐션은 본래 RNN을 보정하기 위한 용도로 사용되는 것이 일반적이었지만, 트랜스포머

모델은 어텐션만을 활용하여 인코더와 디코더를 구현하였습니다. 트랜스포머 모델은 이렇게 어텐션을 활용하여 순차적 데이터 내의 관계를 추적하고 의미를 학습하는 신경망이라고 할 수 있습니다.

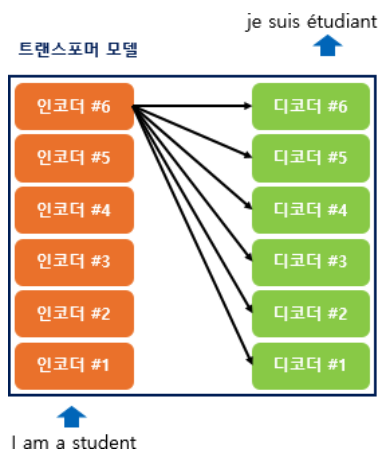


그림 출처: 위키독스 - 『딥 러닝을 이용한 자연어 처리 입문』<https://wikidocs.net/book/2155>

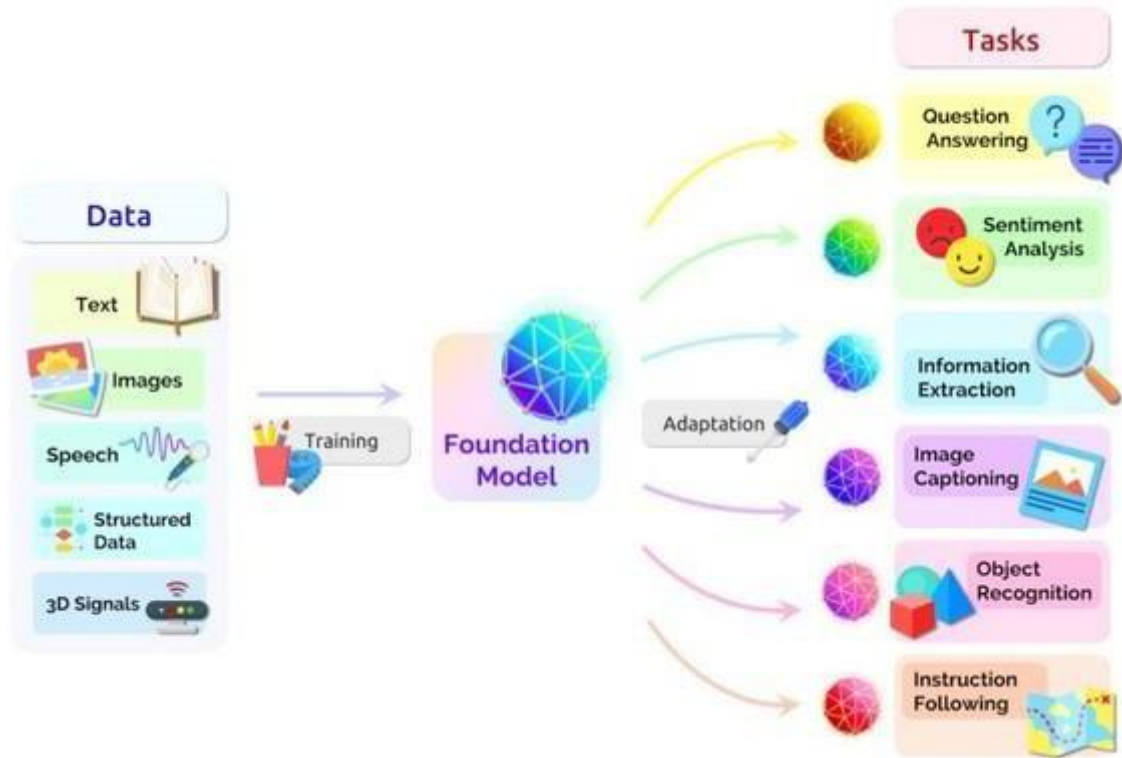
트랜스포머 모델은 **셀프어텐션(Self-Attention)**이라는 메커니즘을 사용하는데 자기 자신에게 어텐션을 취하는 개념인데 이를 통해 입력된 단어들의 유사도를 구합니다. 쉽게 말해 입력된 단어들의 관계성을 찾는 하나의 메커니즘이라고 할 수 있습니다. 트랜스포머 모델은 데이터 사이의 패턴을 수학적으로 찾아내기 때문에 라벨링이 없는 데이터셋으로 훈련이 가능합니다. 라벨링 없는 데이터로 학습하는 것은 곧 자기지도로 발전하는 것을 의미합니다. 이러한 이유로 트랜스포머 모델은 끊임없이 진화하는 수학적 기법이 적용되었다고도 합니다.

파운데이션 모델(Foundation Model)의 정의

파운데이션 모델의 정의는 트랜스포머 모델과 상당히 연관이 깊습니다. 파운데이션 모델에 해당하는 상당수의 모델이 트랜스포머 기반으로 구축된 모델이기 때문입니다. 파운데이션 모델의 대표적 예시인 BERT는 트랜스포머의 인코더를, GPT는 트랜스포머의 디코더를 분리해 각각 독자적으로 발전시킨 모델입니다.

파운데이션 모델은 그 이름처럼 여러 산업 혹은 인공지능 분야 발전의 기반(foundation)이 될 모델들의 범주형 명칭입니다. “파운데이션 모델은 방대한 양의 폭넓은 데이터를 사용하여 자기 지도학습을 통해 방대한 내부 파라미터를 지닌 모델을 학습시킨 후, 아직 명확하게 수행해야 할 작업이 특정되지 않은 상태로

배포된 것”이라고 한국지능정보사회진흥원에서 정의하고 있습니다. 파운데이션 모델을 이용하게 되는 개인 혹은 조직은 해당 모델을 자신들의 이용목적에 맞게 미세조정(fine-tuning)을 하여 원하는 방식으로 사용할 수 있습니다. 이러한 이유로 파운데이션 모델은 ‘**광활한 가능성**’을 지니고 있다고 칭해집니다.



파운데이션 모델 개념도 (출처 : On the Opportunities & Risks of Foundation Models. 2021)

파운데이션 모델을 더욱 간단하게 지칭하자면 “게임 체인저(Game Changer)”라고 할 수 있을 것 같습니다. 파운데이션 모델은 이후 인공지능 분야의 판도를 바꿀 만한 모델들을 일컫습니다. 등장만으로 기존 인기 모델이었던 CNN과 RNN의 성능을 뛰어넘은 트랜스포머 모델을 포함하는 것은 물론, 생성형 적대 네트워크 (GAN), 변량 인코더 등과 같은 생성형 인공지능 기반으로 한 모델들을 말합니다. 파운데이션 모델은 생성형 AI의 한 형태라고도 할 수 있습니다.

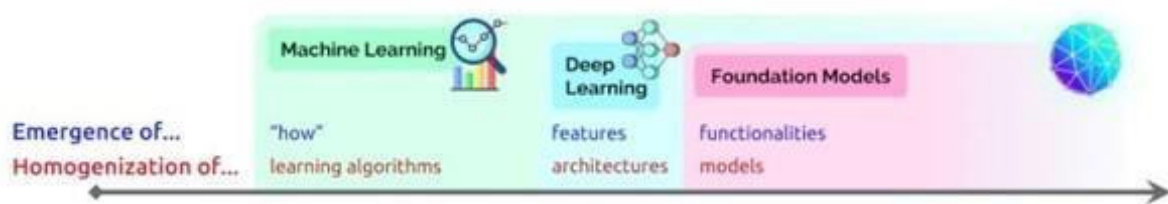
파운데이션 모델의 특징

파운데이션 모델의 대표적인 예시로는 구글의 자연어처리 언어 모델 BERT, Open AI의 대화형생성 인공지능 GPT 시리즈 등이 있습니다. 이러한 모델들의 특징은 ‘수행할 역할(downstream)’이 정해지지 않은 채 배포되어 사용자가 원하는

목표를 모델에 부여할 수 있습니다. 미완성이기에 범용적으로 사용할 수 있는 모델이고 응용 방법이 무궁무진하며 분야와 산업에 대한 적응성이 뛰어나다고 볼 수 있습니다. 다운스트림 작업은 보통 파인튜닝의 과정을 거치지만, GPT의 경우 미세조정(fine-tuning)을 거치지 않고 프롬프트 러닝을 통해 사용자가 원하는 결과물을 출력할 수 있게 유도하기도 합니다.

파운데이션 모델은 레이블링이 없는 데이터에서 학습할 수 있는 자기지도 학습을 사용합니다. 이러한 이유로 파운데이션 모델은 라벨링이 되지 않은 방대한 규모의 원시 데이터로 학습을 진행합니다.

파운데이션 모델의 개념 등장 이후 머신러닝 내지는 딥러닝은 더 이상 모델을 처음부터 제작하지 않고 파운데이션 모델을 시작지점 삼아 기계학습 모델을 제작하게 됩니다.



머신러닝과 딥러닝을 기반으로 '창발성'과 '균일화'를 추구해면서 진화해온 인공지능의 흐름을 표현한 이미지
(출처 : On the Opportunities & Risks of Foundation Models. 2021)

¹파운데이션 모델의 특징을 두 단어로 요약하면 '**출현**(혹은 창발성, Emergence)'과 '**균일화**(혹은 동질화, Homogenization)'입니다. ²**발성성**은 모델이 스스로 어떠한 문제를 해결하기 위한 **지식을 스스로 도출하는 능력**을 말합니다. 예를 들어 대규모 언어 모델은 다국어 문장 데이터를 학습하여 문장 간의 의미를 이해하고 번역할 수 있게 됩니다. 이러한 모델은 언어 간의 번역 기능도 생겨나는 데 이것이 학습 데이터에서 명시적이지 않은 기능이 자동으로 "출현"하는 예시입니다.

³**균일화**는 모델이 점차 일반화된 지식을 도출해낼 수 있게 됨에 따라, 하나의 뛰어난 **모델이 적용될 수 있는 범위가 점차 확대**되며 더욱 보편적이고 범용적으로 활용되는 현상을 의미합니다. 일례로 트랜스포머 모델 등장 이전의 컨볼루션

¹ Center for Research on Foundation Models (2021), On the Opportunities & Risks of Foundation Models

² 구자환, 김도형 (한국지능정보사회진흥원(NIA)), 2023, 파운데이션 모델의 이해와 미래 전망, Digital Insight 2023-1

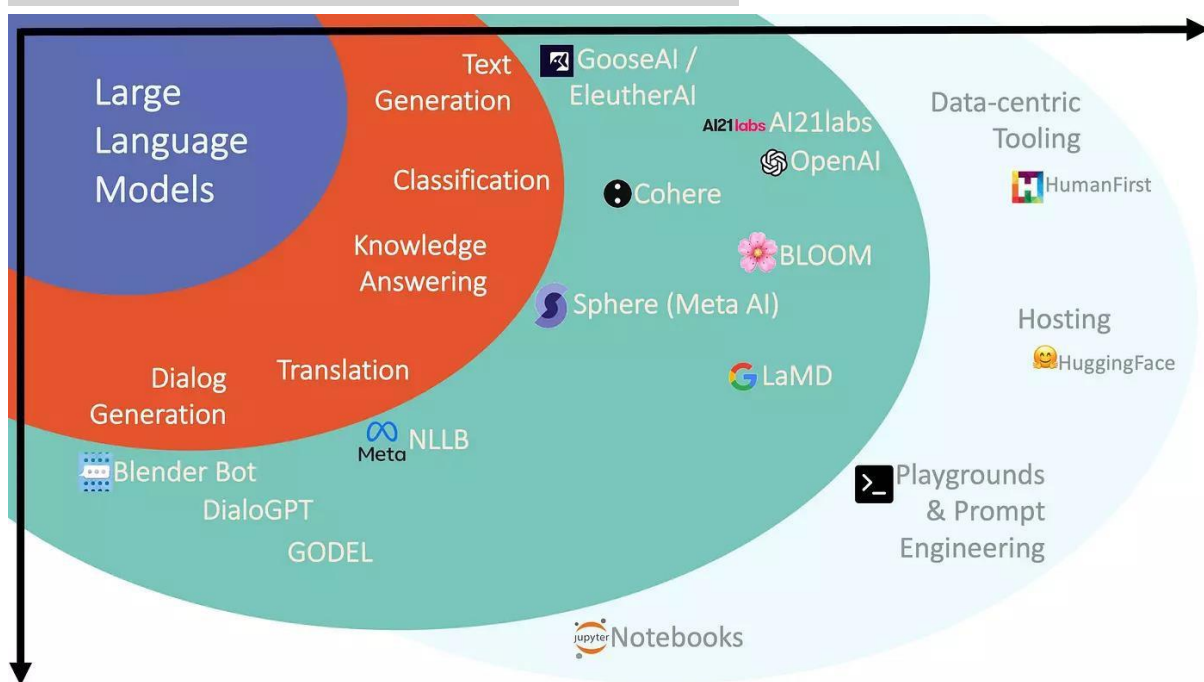
³ 구자환, 김도형 (한국지능정보사회진흥원(NIA)), 2023, 파운데이션 모델의 이해와 미래 전망, Digital Insight 2023-1

신경망(CNN)은 컴퓨터 비전 작업뿐만 아니라 음성 및 자연어 처리 분야에서도 활발하게 사용되는 표준 모델이었습니다.

확장성이 높고 지속적인 발전이 가능한 파운데이션 모델이지만, 장점만 있는 것은 아닙니다. 창발성과 균일화의 특성은 파운데이션 모델이 가지는 위험성도 설명합니다. 창발성은 모델에서 **예기치 못한 부정적인 결과**가 생겨날 수 있음을 그 의미에 내포하고 있습니다. 균일화에서는 같은 모델이 많은 작업에 쓰이는 것은 해당 모델에 결함이 발견될 시, 관련된 작업들과 **연계되어 더 커다란 결함으로 연결될 수** 있음을 암시합니다.

이러한 단점들 외에도 파운데이션 모델이 가장 조심해야 할 리스크는 바로 **편향성**입니다. 모델이 훈련 데이터 세트에서 나타나는 부정적인 편견이나 편향성을 찾아내어 학습할 수도 있기 때문입니다. 이를 방지하려면 개발자는 훈련 데이터를 신중하게 필터링하고 특정 규범을 모델에 인코딩해야 할 것입니다.

대규모 언어 모델(Large Language Model)



정의

대규모 언어 모델(LLM)은 그 범위를 어디까지 보는가에 따라 특성이 약간 다를 수 있습니다. 구글의 BERT를 포함하는 경우와 파라미터가 조 단위의 생성형 인공지능 모델들만 포함하는 경우로 정의하는 등 의견마다 미세한 차이가 존재함

니다. 추가적으로 LLM의 한국어 번역문인 '초거대 언어 모델'을 같은 LLM을 지칭하는 '대규모 언어 모델' 보다 더 거대한 언어 모델의 명칭으로 보는 시각도 존재합니다. 명확한 정의와 근거는 존재하지 않아서 대규모 언어 모델에 대한 정의는 약간씩의 차이가 존재하지만 공통적으로는 이름 그 자체에서 알 수 있듯이 방대한 규모의 언어 데이터로 비지도학습을 진행한 모델입니다. 이외에도 트랜스포머 모델 기반으로 자기지도 학습을 통해 다양한 언어적 기능을 지원하고, 시퀀스 병렬처리를 하며, 최소 억 단위 이상의 파라미터를 보유하고 있는 특성이 있습니다.

대규모 언어 모델은 **일정 수준 이상의 파라미터를 보유하게 되면 성능이 아주 높아집니다**. 현재는 파라미터의 수가 모델의 성능과 정비례하는 지표라고 볼 수 있습니다.

사용처

자연어 처리는 파운데이션 모델에 가장 큰 영향을 받은 분야입니다. 특히 파운데이션 모델의 특성인 창발성이 가장 두드러지게 나타나는 분야이기도 합니다. 대형 언어 모델(LLM)은 방대한 양의 텍스트 데이터를 기반으로 사전 학습된 초대형 딥 러닝 모델입니다. 가장 유명한 사례는 Open AI의 GPT 시리즈가 있지만, 초거대 언어 모델은 단순히 대화형 텍스트 생성에만 그 기능이 국한되지 않습니다. 이외에도 텍스트 분류, 요약, 질의응답, 번역, 감정분석 등 자연어 이해(Natural Language Understanding)와 자연어 생성(Natural Language Generation)을 아우르는 광범위한 자연어 처리 작업에 사용됩니다.

대규모 언어 모델은 **퍼블릭(Public)**과 **프라이빗(Private)**으로 나뉩니다. 퍼블릭은 파운데이션 모델처럼 사용처를 정하지 않고 **오픈 소스**로 배포된 것을 말합니다. 프라이빗은 주로 기업이 자신들의 목적에 맞게 사용하는 모델을 말합니다. 프라이빗 LLM의 경우에는 보통 퍼블릭 LLM보다 학습되는 데이터의 규모는 작습니다. 하지만 보다 전문성 있는 양질의 데이터와 파인튜닝으로 인해 특정 산업 혹은 분야에서 만큼은 퍼블릭 LLM보다 빠른 속도와 뛰어난 결과물을 출력합니다.

프라이빗 LLM의 장점은 보안성이 높고 비용이 낮다는 점입니다. 프라이빗 LLM은 **적은 데이터 규모로도 효과적으로 학습**할 수 있기 때문에 상대적으로 적은 비용이 듭니다. 추가적으로 인공지능이 잘못된 정보를 전달하는 할루시네이션

(Hallucination)도 감소할 수 있습니다. 프라이빗 LLM 모델은 주로 기업의 내부 직원들을 지원하는 도구로 사용하거나, 고객 경험을 증대시켜주는 고객 접대용으로 사용합니다. 퍼블릭 LLM과 비교해서 상대적으로 제한된 인원이 접근을 가지고 권한체계를 임의로 부여할 수 있기 때문에 보안성이 더 높습니다.

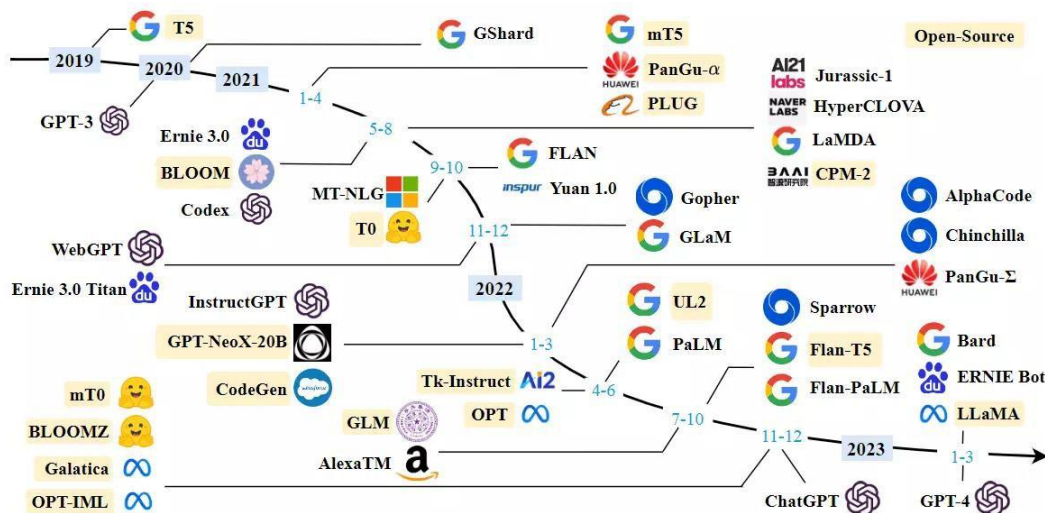


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

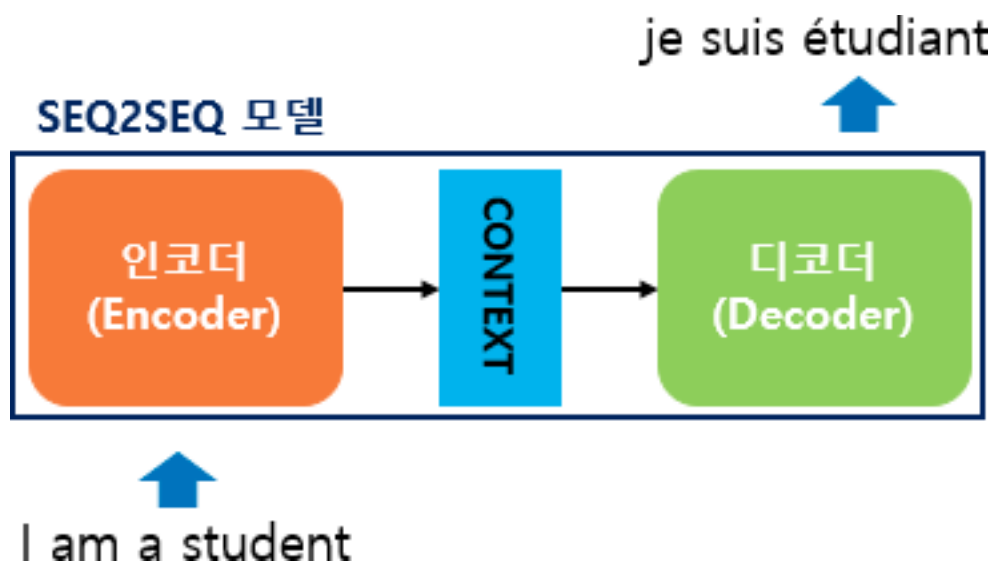
초거대 생성형 AI

LLM을 프롬프트로 입력문을 받는 최근의 초거대 생성형 인공지능 모델로 정의를 내리는 시각에서는 BERT 같은 대규모이지만 생성형이 아닌 모델 혹은 데이터 학습 규모가 상대적으로 적은 경우에는 Pre-trained Language Model(PLM)로 부르며 구분 짓습니다. 이 경우 LLM은 프롬프트를 통해서도 결과물의 품질을 조정할 수 있습니다. 이에 관련해서는 프롬프트로 사례를 입력하는 원샷(one-shot), 퓨샷(few-shot)과 사례 없이 원하는 결과물을 출력하는 제로샷(zero-shot) 등이 연구되었고 이외에도 프롬프트 엔지니어링이라는 전문 분야가 새로 생겼습니다.

GPT와 같은 생성형 인공지능 모델이 앞으로의 자연어처리 분야의 메인스트림이 될 것이 자명하기에, "LLM = 초거대 생성형 언어 모델"의 명제 또한 충분한 설득력이 있습니다. 다만, GPT-2는 15억개의 파라미터를 가졌지만 그 다음 버전인 GPT-3는 1750억개의 파라미터를 보유하고 있고 최근 모델은 파라미터의 개수가 조 단위입니다. 시간이 지날수록 언어 모델의 규모는 점점 더 커져 갈 것이기에 향후 대규모 언어 모델에 대한 용어나 범위가 변화해가는 추이를 지켜볼 필요가 있습니다.

현재 생성형 언어 모델에는 어느정도 한계점이 있습니다. 우선 인공지능에 대한 교정은 사람이 하기 때문에 결국 출력물의 최신성, 완성도, 신뢰성 등도 조정하는 이의 능력이 영향을 미칩니다. 또한 인공지능이 생성해낸 출력물의 신뢰도의 정도를 사용자가 알 수 없습니다. 전혀 엉뚱한 대답이어도 해당 대답이 어떻게 나왔는지, 얼마나 신뢰 가능한지를 사용자가 알 수 없습니다. 추가적으로 윤리 적, 법적, 정치적 등과 관련한 내용에 대한 완벽한 통제가 아직은 불가능한 한계 점도 있습니다.

인코더-디코더 모델 vs 디코더 모델

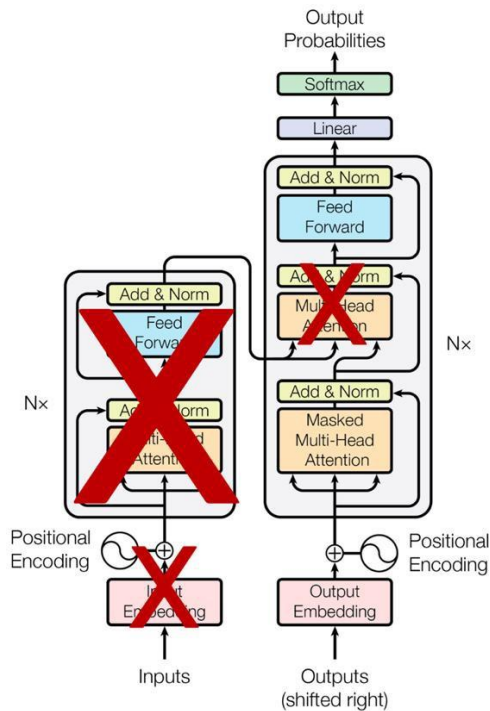


인코더-디코더 모델

자연어처리 중 번역에 한해서는 인코더-디코더(seq2seq) 모델이 가장 뛰어난 성능을 보입니다. 이는 '번역'이라는 작업이 한 언어의 문장(sequence)을 입력 받아(to) 해당 문장을 다른 언어 문장(sequence)으로 변환하는 것이 때문입니다. 이 때문에 자연어처리 분야에서는 seq2seq라고도 불리는 인코더-디코더 모델이 가장 번역 기능에 적합할 수 밖에 없습니다.

인코더(Encoder)는 입력된 자연어를 숫자인 벡터로 치환해주는 임무를 맡습니다. 반대로 디코더(Decoder)는 입력받은 인코딩 벡터를 자연어로 다시 재생성을 합니다. 때문에 인코더가 존재하는 모델은 그렇지 않은 모델보다 입력문을 더 잘

이해하며, 보다 충실히 이행할 수 있습니다.



디코더 모델 : GPT

인코더-디코더 모델은 입력 받은 문장 혹은 단어의 의미를 가장 확률이 높은 해석 혹은 여러가지 해석은 확률이 높은 순서로 사용자에게 제시합니다. 단어나 문자 그대로의 의미(literal meaning)에 충실한 번역을 진행합니다. 반면에 디코더만 존재하는 모델인 GPT는 조금 다른 모습을 보입니다.

인코더가 없는 GPT 시리즈 모델들은 **디코더만을 활용**합니다. GPT는 텍스트 **생성**에 특출난 성능을 발휘합니다. GPT는 입력된 자연어를 통해 추론을 하여 그 다음에 올 내용을 출력합니다. 셀프어텐션 기법을 통해 프롬프트에 입력된 내용을 바탕으로 텍스트를 생성해냅니다. 이러한 원리를 통해 GPT와 사용자는 대화, 질의응답 등 다양한 상호작용이 가능합니다.

GPT는 인코더가 없기 때문에 미래의 텍스트가 아닌 이전의 텍스트(프롬프트)에만 반응을 합니다. 또한 GPT를 포함한 생성형 거대 언어 모델은 **자동회귀(auto-regressive) 디코더**를 가지고 있습니다. 자동회귀는 이전에 출력한 결과물을 포함하여 다시 다음 출력을 예측하는 재귀함수의 형태를 띤 모델입니다. 자동회귀 모델을 통해 이전 단어들로 다음 단어를 유추하는 것이 반복됩니다. 이러한 메커니

즘으로 양질의 결과물을 출력할 수 있는 이유는, 디코더 모델이 방대한 양의 텍스트 데이터를 학습하면서 각 단어와 문장이 어떻게 사용되는지에 대한 패턴을 학습하기 때문입니다.

디코더 모델의 특징

GPT와 같이 디코더만 있는 모델의 특징은 **발의성**이 뛰어나다는 점입니다. GPT는 방대한 규모로 학습된 데이터를 바탕으로 입력된 프롬프트에 대한 창의적인 대답을 내놓습니다. 인코더가 없기 때문에 입력문에 대한 이해가 떨어져 부적절한 결과물을 출력할 수도 있지만 동시에 보다 창의적인 결과물을 내놓아 사용자에게 다양한 대안을 제시할 수 있습니다.

이러한 특징은 GPT를 활용한 번역에도 마찬가지로 적용됩니다. 특정 언어나 콘텐츠를 현지화를 하는 것은 단순히 문자나 단어의 의미를 초월하여 문화나 사회적인 맥락 전체를 고려해야하는 경우가 많습니다. 이에GPT와 같은 생성형 인공지능은 기존의 신경망 기계번역에 비해 문화적 해석이나 창의적인 해석능력이 탁월합니다.

GPT에 대응되는 생성형 AI : 인코더-디코더 모델의 Bard

특이하게도 구글의 Bard 대화형 인공지능은 GPT와 같은 기능을 구현하면서 인코더-디코더 모델의 형태를 띄고 있습니다. 이론적으로는 Bard는 인코더가 있기에 보다 프롬프트를 이해하고 관련성이 높은 대답을 할 수 있습니다. 반대로 GPT의 경우는 보다 인간에 가깝게 창의적이고 다양한 대답을 생성해낼 수 있는 장점이 있습니다. 다만, 창의적인 만큼 결과물의 변동성 또한 크고, 사용자가 입력한 프롬프트와 관련성이 떨어질 수 있는 단점이 동시에 존재합니다. 창의성이나 언어의 유창함 측면으로 보면 인코더-디코더 모델은 입력 문장의 의미를 왜곡하지 않는 선에서 번역을 수행하기 때문에, 다소 딱딱하고 단조로운 표현의 번역 결과를 생성할 수 있습니다.

	Bard	GPT
용도	다양	적음
비용	비쌈	상대적으로 싼
모델 구조	복잡	상대적으로 단순

데이터 훈련(학습)	상대적으로 더 어려움	상대적으로 더 쉬움
비고	장문의 입력문에서 거리가 먼 문자간의 관계나 전체적인 맥락을 이해	창의적인 결과물을 생성

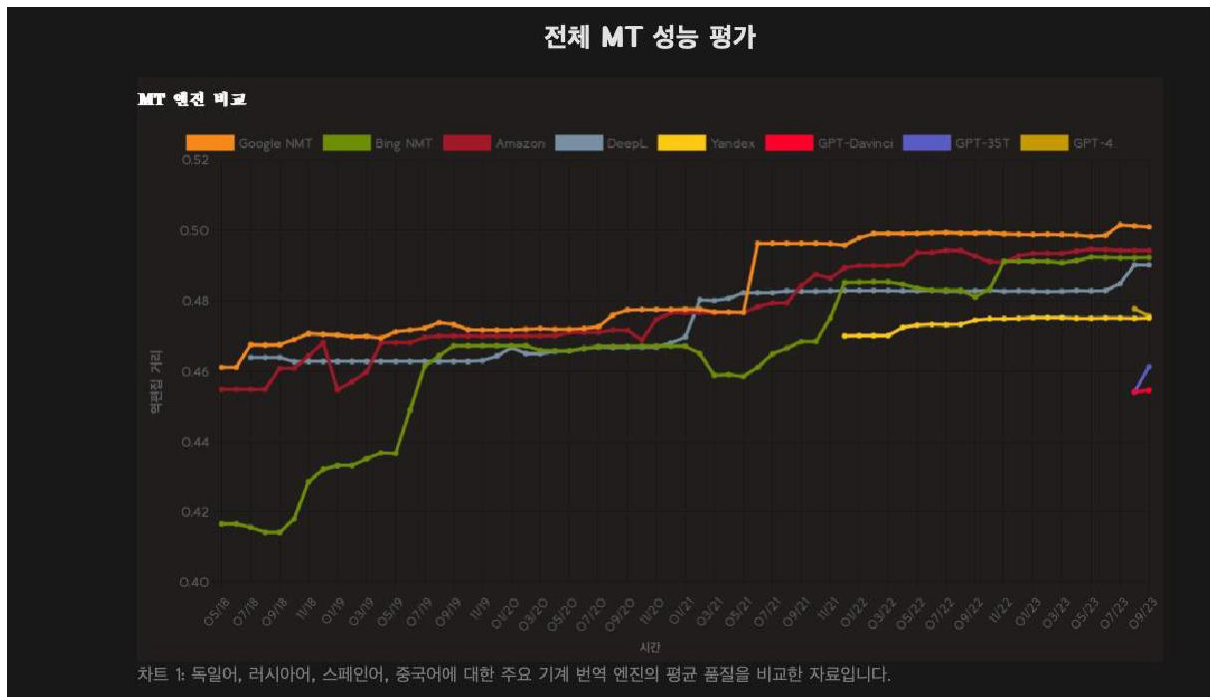
디코더 모델은 입력 문장의 순차적인 정보를 바탕으로 의미를 이해하기 때문에, 입력 문장의 구조나 의미의 복잡성이 높을 경우 정확한 번역을 수행하기 어렵습니다. 반면, 인코더-디코더 모델은 입력 문장의 전체적인 의미를 이해할 수 있기 때문에, 구조나 의미가 복잡한 문장도 보다 정확하게 번역할 수 있습니다.

경험적인 관측이지만 이전에 작성한 보고서로 미루어 보면 확실히 Bard가 프롬프트를 보다 상세하게 이해한 결과물을 내놓았고 ChatGPT-3.5는 장문의 프롬프트 내용 일부를 무시하는 경우가 잦았습니다.

GPT와 Bard의 번역 성능 비교 난항

디코더 모델인 동시에 생성형 인공지능인 GPT와 인코더-디코더 모델인 동시에 생성형 인공지능인 Bard의 번역 성능을 비교하려 했지만 관련 연구자료가 없어서 이 둘 간의 명확한 번역 성능비교는 어렵습니다. 명확한 비교가 어려운 이유로는 Bard가 발표된 지 그리 오랜 시간이 지나지 않은 점, 짧은 주기로 경쟁력 있는 새로운 LLM이 발표되거나 기존의 모델이 개선되어서 새로 발표(GPT-4 turbo)되고 있는 점, GPT와 Bard는 학습된 데이터에 차이도 존재하기에 결과물의 차이가 다른 변수의 개입없이 단순히 인코더의 유무로 인한 것인지 판별하기 어려운 점 등이 있습니다.

기계번역 현황과 전망



라이온브리지에서 주관한 기계번역 성능 평가 차트 (출처 : lionbridge.com)

ChatGPT를 위시한 거대 생성형 인공지능의 등장 이후로 기계번역 변화가 생기기 시작했습니다. 생성형 인공지능이 기계번역에 대한 새로운 시사점을 보였기 때문입니다.

⁴현재 번역 성능이 가장 뛰어나다고 평해지는 것은 인코더-디코더 모델의 신경망 기계번역(NMT)입니다. 실제로 사용되고 있는 번역기 중 상당수가 트랜스포머 모델 기반의 인코더-디코더 형태를 취하고 있는 신경망 기계번역 모델입니다.

하지만 GPT와 Bard 같은 생성형 인공지능은 프롬프트로 번역 품질 조정 가능한 **잠재성**이 있습니다. 이 부분도 아직은 연구되고 있는 분야이지만 프롬프트 조정에 따라 개인 맞춤화 문체번역을 보급하거나 보다 상세한 디테일의 현지화가 가능할 수 있습니다.

⁵추가적으로 현재도 생성형 인공지능은 신경망 기반 기계번역보다 문화적인 해석이나 특정 언어 번역(독일어->영어) 등 번역성능이 우세인 분야가 꽤 있습니다.

⁴ 라이온브리지 (2023), 한 차례의 기계번역 평가에서 NMT 엔진의 성능을 능가한 생성형 AI 모델

⁵ Jungha Son, Boyoung Kim (2023), Translation Performance from the User's Perspective of Large Language Models and Neural Machine Translation Systems

신경망 기반 기계번역은 상대적으로 성숙기에 들어선 반면, 생성형 LLM의 분야는 이제 성장이 한참남은 분야입니다. 때문에 생성형 인공지능을 통한 번역은 그 잠재성이 높기에 그 변화 추이를 지켜봐야 하겠습니다.

고찰

현재 업무의 목표는 AWS의 SageMaker의 파운데이션 모델을 문체번역의 용도로 활용해보고, 이를 GPT와 Bard와 비교해보는 것입니다. 하지만 실제로 SageMaker의 FM을 사용해본 결과 해당 모델들을 GPT나 Bard처럼 바로 문체번역에 사용할 수는 없었습니다.

영어기반으로 문장완성의 기능을 가진 모델, 신경망 기반 기계번역과 기능이 차이가 없는 모델만 존재했습니다. 해당 모델들에 '문체 번역'이라는 다운스트림을 지정하고 이에 대한 파인튜닝을 진행해야할 것으로 생각됩니다. 이에 앞서 파운데이션 모델과 기계번역 관련된 배경지식을 높이고자 해당 문서에 해당하는 항목들을 조사한 후 정리했습니다.

기본적인 개념정리를 통해 상승된 시각으로 SageMaker의 파운데이션 모델을 효과적으로 연구해보겠습니다.

추가사항

<https://www.aitimes.com/news/articleView.html?idxno=154975>

이전에 작성했던 Wisetranslate, 프롬프트 엔지니어링과 관련된 기사를 발견해서 본 내용을 추가하였습니다.

11월 6일 Open Ai가 GPT-4 터보(Turbo)와 GPT스토어를 발표했습니다. ChatGPT의 사용자 정의 도구 'GPT'를 도입한다고 전했습니다. GPT와 GPT를 만드는 GPT빌더는 개인 맞춤형 챗봇(GPT)을 구축할 수 있는 도구이고, 이로 인해 기업은 코드 작업 없이 내부 전용 GPT를 설계할 수 있게 되었다고 합니다.

이외에도 다음과 같은 기능이 탑재된다고 합니다.

- 기존 챗봇을 가져와 테스트하고 수정 가능
- 챗봇 추가 기능 정의
- 챗봇이 참고할 파일 추가
- 웹 탐색이나 이미지 생성 등 기능 추가
- 챗봇 사용 데이터 분석
- 생성한 챗봇의 라이브 테스트

커스텀 GPT를 생성하고 이를 통해 다른 사용자와 교류할 수 있는 커뮤니티가 GPT스토어를 통해 구현되었습니다. 이를 통해 프롬프트 엔지니어링이 더욱 활성화되어서 다양한 템플릿이 나오고, GPT를 통해 다양한 작업을 할 수 있을 것으로 생각됩니다.







GPT스토어를 간단하게 살펴 본 결과 다양한 유저들이 만든 다양한 GPT를 사용할 수 있는 형태였습니다. ChatGPT에 대한 유료 계정이 필요한 것 같아 각 템플릿을 시험운용해 보지는 못하였습니다.

GPTStore.ai
GPTs List of ChatGPT | GPTStore.AI
gptstore.ai
GPTs
Creators
GPT Plugins
CharClub
Submit GPTs

GPTs Collection

We have found 8797 GPTs

[Submit your amazing GPTs](#)

 <p>Architect Assistant</p> <p>Humorous architect in sustainable, modern design</p>	 <p>Argentina Balotage 2023</p> <p>Analista político objetivo para las elecciones argentinas, con enfoque en las plataformas de Unión por la Patria y La Libertad Avanza.</p>	 <p>Art Mystic</p> <p>Your Guide to AI Artistry</p>
 <p>Art to NFT</p> <p>Guide for Artists on NFT Creation & Web3 Intecration</p>	 <p>ArticleGPT</p> <p>Expert in SEO-optimized article writing</p>	 <p>Artista Noticiero</p> <p>Crea cómics en español sobre noticias actuales.</p>

Home > GPTs > 韩国语 翻译 Pro



韩国语 翻译

Daily updated m

Author

SONG CHENGWEN [View his/her GPTs](#)

Voice

-

GPT link in ChatGPT

[Use 韩国语 翻译 Pro in ChatGPT](#)

Welcome message

환영합니다! 매일 최신의 데이터베이스를 통해 정확한 번역을 제공합니다.

CharClub

Play, Chat, and Create Memorable Moments with unfiltered virtual characters.

Get started

Explore characters

Search

NSFW

Top



Raiden Shogun and



Gojo Satoru



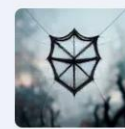
Yae Miko



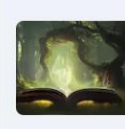
SM64 Mario



Psychologist



Text Adventure



Isekai narrator

번역 템플릿이 존재하는 것은 물론, "CharClub"이라는 GPT에 캐릭터를 입력해 해당 캐릭터와 소통하는 목적으로 만든 템플릿을 공유하는 커뮤니티도 내부에 별도로 존재했습니다. 이러한 GPT스토어를 통해 문체 번역이나 특정 캐릭터 어투를 유지한채 번역하는 것도 추후에 크게 활성화 될 가능성이 있어 보입니다.

참고문헌

- 1) NVIDIA KOREA(2023.04.04), 파운데이션 모델이란 무엇인가?, NVIDIA 블로그, <https://blogs.nvidia.co.kr/2023/04/04/what-are-foundation-models/>
- 2) NVIDIA KOREA(2022.04.01), 트랜스포머 모델이란 무엇인가? (1), NVIDIA 블로그, <https://blogs.nvidia.co.kr/2022/04/01/what-is-a-transformer-model/>
- 3) NVIDIA KOREA(2022.04.01), 트랜스포머 모델이란 무엇인가? (2), NVIDIA 블로그, <https://blogs.nvidia.co.kr/2022/04/01/what-is-a-transformer-model-2/>
- 4) Google (2017), Attention Is All You Need, <https://arxiv.org/abs/1706.03762>
- 5) Center for Research on Foundation Models (CRFM), Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University (2021), On the Opportunities and Risks of Foundation Models, <https://arxiv.org/abs/2108.07258v3>
- 6) 유원준, 안상준 (2023), 딥 러닝을 이용한 자연어 처리 입문 (E-book), <https://wikidocs.net/book/2155>
- 7) Felix Stahlberg(2019), Neural Machine Translation: A Review and Survey, <https://arxiv.org/abs/1912.02047>
- 8) 전윤미 기자(2023.04.13), 초대형 생성AI의 대중화 이끄는 '파운데이션 모델', 애플경제, <https://www.apple-economy.com/news/articleView.html?idxno=71203>
- 9) AWS, 파운데이션 모델이란?, <https://aws.amazon.com/ko/what-is/foundation-models/>
- 10) AWS, 대규모 언어 모델이란?, <https://aws.amazon.com/ko/what-is/language-model/>
- 11) 구자환, 김도형(한국지능정보사회진흥원(NIA)) , 2023, 파운데이션 모델의 이해와 미래 전망, Digital Insight 2023-1
- 12) 최형광(한국정보통신기술협회), 20223, AI 파운데이션 모델 구축 이슈와 국내 업계 동향, TTA 저널 207호, 05/06월호

- 13) Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, Ji-Rong Wen (2023), A Survey of Large Language Models, <https://arxiv.org/abs/2303.18223>
- 14) 안성원, 유재흥, 조원영, 노재원, 손효현 (2023), 초거대언어모델의 부상과 주요이슈 - ChatGPT의 기술적 특징과 사회적·산업적 시사점, SPRI 이슈리포트, <https://spri.kr/posts/view/23561>
- 15) 임수종 , 초거대 인공지능 언어모델 동향 분석
- 16)Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, Xia Hu (2023), Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond, <https://arxiv.org/abs/2304.13712>
- 17) Jungha Son, Boyoung Kim (2023), Translation Performance from the User's Perspective of Large Language Models and Neural Machine Translation Systems, <https://www.mdpi.com/2078-2489/14/10/574>
- 18) 라이온브리지 (2023.10), 라이온브리지 기계 번역 추적 도구, <https://www.lionbridge.com/ko/machine-translation/mt-tracker/>
- 19) 라이온브리지 (2023.05), 한 차례의 기계번역 평가에서 NMT 엔진의 성능을 능가한 생성형 AI 모델, <https://www.lionbridge.com/ko/blog/translation-localization/machine-translation-a-generative-ai-model-outperformed-a-neural-machine-translation-engine/>
- 20) 라이온브리지 (2023.02), 챗GPT의 번역 성능 및 기계번역의 미래에 미칠 영향, <https://www.lionbridge.com/ko/blog/translation-localization/chatgpts-translation-performance-and-what-it-tells-us-about-the-future-of-localization/>
- 21) Binwei Yao, Ming Jiang, Diyi Yang, Junjie Hu (2023), Empowering LLM-

based Machine Translation with Cultural Awareness,
<https://arxiv.org/abs/2305.14328>

- 22) Minghan Wang, Jinming Zhao, Thuy-Trang Vu, Fatemeh Shiri, Ehsan Shareghi, Gholamreza Haffari (2023), Simultaneous Machine Translation with Large Language Models, <https://arxiv.org/abs/2309.06706>
- 23) Vikas Raunak, Arul Menezes, Matt Post, Hany Hassan Awadalla (2023), Do GPTs Produce Less Literal Translations? , <https://arxiv.org/abs/2305.16806>
- 24) Chunlan Jiang (2023), Investigation on the Application of Artificial Intelligence Large Language Model in Translation Tasks, Proceedings of the 2023 7th International Seminar on Education, Management and Social Sciences (ISEMSS 2023)