인공지능에서의 언어 처리: 자연어처리

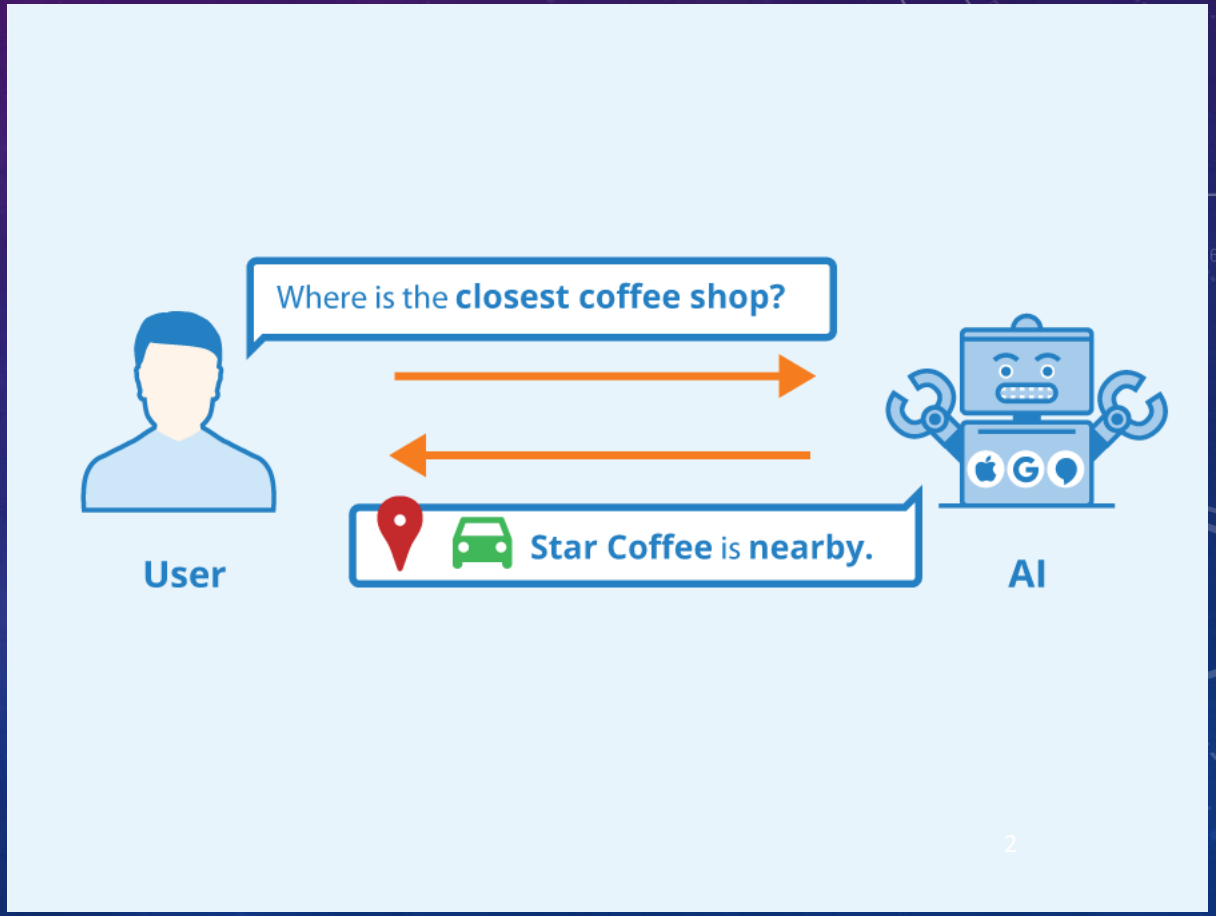
서울대학교 언어학과

신효필

HPSHIN@SNU.AC.KR

01. 자연어처리란?

- 자연어처리(Natural Language Processing) 또는컴퓨터언어학(Computational Linguistics)
  - *Natural Language Processing*, or NLP, is the sub-field of AI that is focused on enabling computers to understand and process human languages.
  - https://www.seobility.net/en/wiki/Natural_Language_Processing
- 인간의 언어와 관련되는 여러 분야-언어학, 컴퓨터학, 전기공학, 인지과학, 심리학, 통계학 등-에 걸치는 학제적인 분야



01. 자연어처리란?

- 컴퓨터는 인간의 언어를 이해할 수 있는가?
  - 인간의 언어를 이해할 수 있도록 많은 노력의 결과로 상당한 진전
  - 그러나 여전히 한계
  - 인간이 언어를 인식하는 것과 완전히 동일하지는 않지만 상당한 수준의 진전을 보임
  - 언어를 이해하기 위한 여러 Tool들이 존재함

02. 자연어처리의 응용 분야는?

- 텍스트분류
  - Spam detection
  - Sentiment Analysis
- 기계번역
- 정보검색
- 챗봇
- 문서요약/문서생성
- 질의-응답시스템

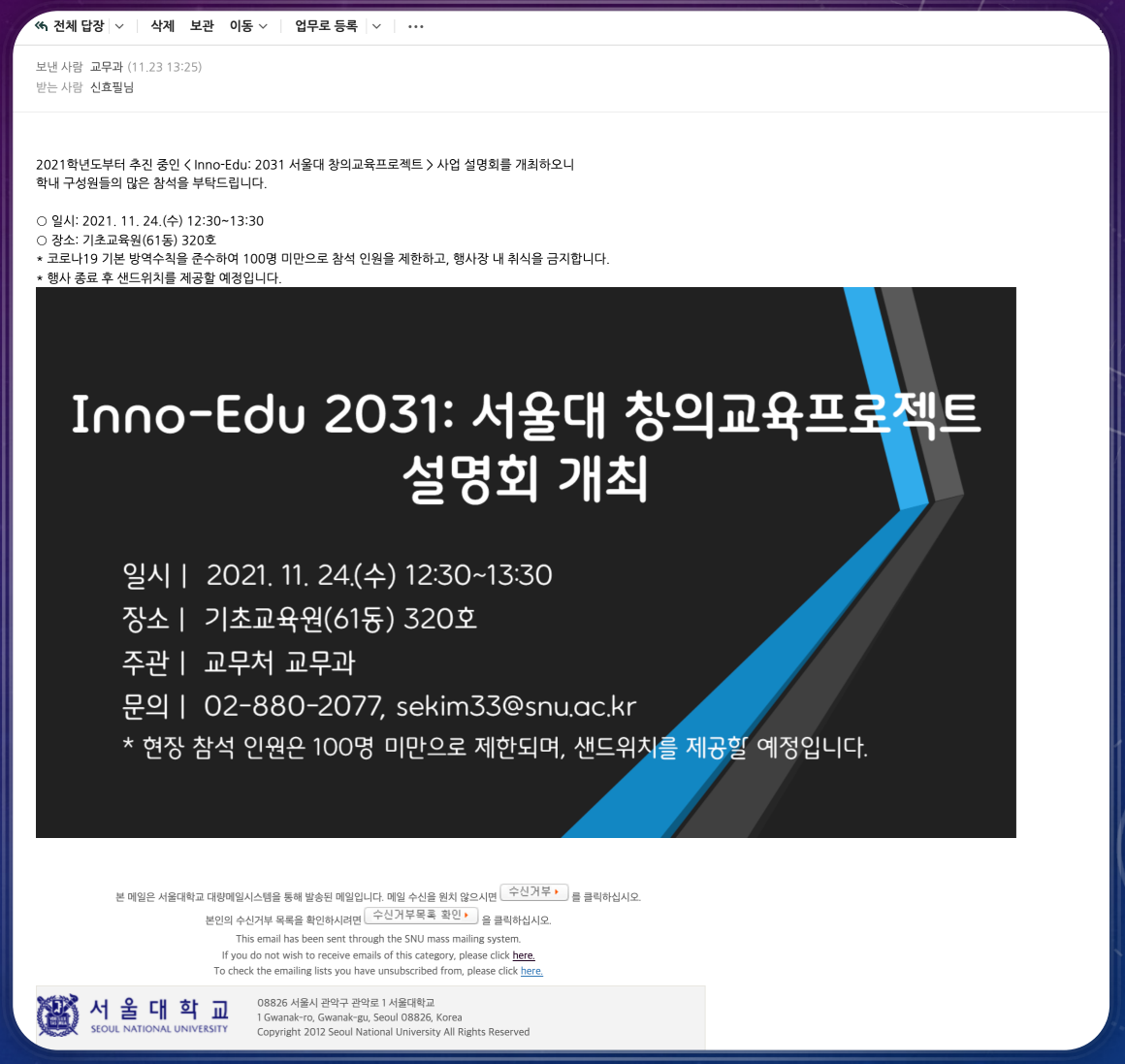
02. 자연어처리의 응용 분야

- Question Answering: IBM's Watson
  - 2011년 2월 우승

- WILLIAM WILKINSON'S "AN ACCOUNT OF THE PRINCIPALITIES OF WALLACHIA AND MOLDOVIA" INSPIRED THIS AUTHOR'S MOST FAMOUS NOVEL
- → Bram Stoker

02. 자연어처리 응용 분야

- Information Extraction
  - 행사: 창의교육프로젝트설명회
  - 날짜: 2021년 11월 24일
  - 시작: 12시 30분
  - 끝: 13:30분
  - 장소:기초교육원 320호



02. 자연어처리 응용 분야

- Sentiment Analysis
  - Attributes: 가격, 상품상태, 매직스페이스

★★★★★ 5 신세계몰 · we****** · 19.11.27.

지정일에 배송됐고 설치도 완벽

지정일에 배송됐고 설치도 완벽했습니다. 1인 가구이지만 수많은 날의 고민 끝에 이 모델로 결정했습니다. 양문형이냐 4도어냐의 고민에서부터 세미빌트인이냐 빌트인이냐까지... 결국 저의 모든 것을 만족시켜주는 건 lg 디오스 양문형 냉장고더라구요^^ 저는 1인가구라 냉동실 사용이 주를 이룰 수 밖에 없고 냉장고의 주요 사용 용도는 마실 것의 보관이였기에 냉동실 사용이 편리한 양문형으로 했고 음료 보관 및 이용이 편리한 매직스페이스가 있는 제품으로 선택하였습니다! 용량에 대한 고민도 많았는데요.. 1인 가구가 쓰기에 800리터대의 제품은 너무 낭비이지 않을까 싶었지만.. 냉장고와 티비는 거거익선이라는 말은 정말 띵언인 것 같습니다~ 작아서 불편한 것 보단 넉넉히 사용하는 게 훨씬 나은 것 같습니다~ 냉장고를 꽉꽉 채워서 쓰면 오히려 냉장고가 제 기능을 발휘하기 힘들다고 하더라구요~ 솔직히 이 정도 크기와 성능에 이름 있는 브랜드 제품을

리뷰펼치기

★★★★★ 5 11번가 · pj****** · 20.08.09.

최고예요

가격이 좋아요 2등급이고색상시 고급지고 튼튼해보이는 스탈 매직스페이스가있어 음료수 물 편하게 이용하고 에너지절감도 되고부모님께 선물했는데 아주 만족 좋아힌십니다다만 문쪽 선반이 작은거라도 맨아랫까지있음 더알찰텐데 조금아습네요냉동실 얼음얼리는곳시 크진않지만 두분이 쓰기딴좋을만큼되네요 강추~

리뷰접기

★★★★★ 5 신세계몰 · cc****** · 19.10.11.

친정엄마 냉장고가 10년 다되어

친정엄마 냉장고가 10년 다되어가는 기존 양문형 냉장고였는데 고장은 없으니 냉동실이 좁아 냉기순환이 안되는지 꽁꽁 얼지않아 선물 드렸는데 정말 마음에 들어하시네요
4도어 5도어가 대세이긴 하나 기능대비 가격 차이가 너무 많이 나서 선택했는데 대만족입니다
양문형이긴 하나 외관상으론 5도어처럼 보입니다ㅋㅋ
특히 매직스페이스가 정말 좋아요
예전 홈바보다 사이즈도 크고 냉기보존도 잘 되는듯요

리뷰펼치기

★★★★★ 5 신세계몰 · sa****** · 20.05.09.

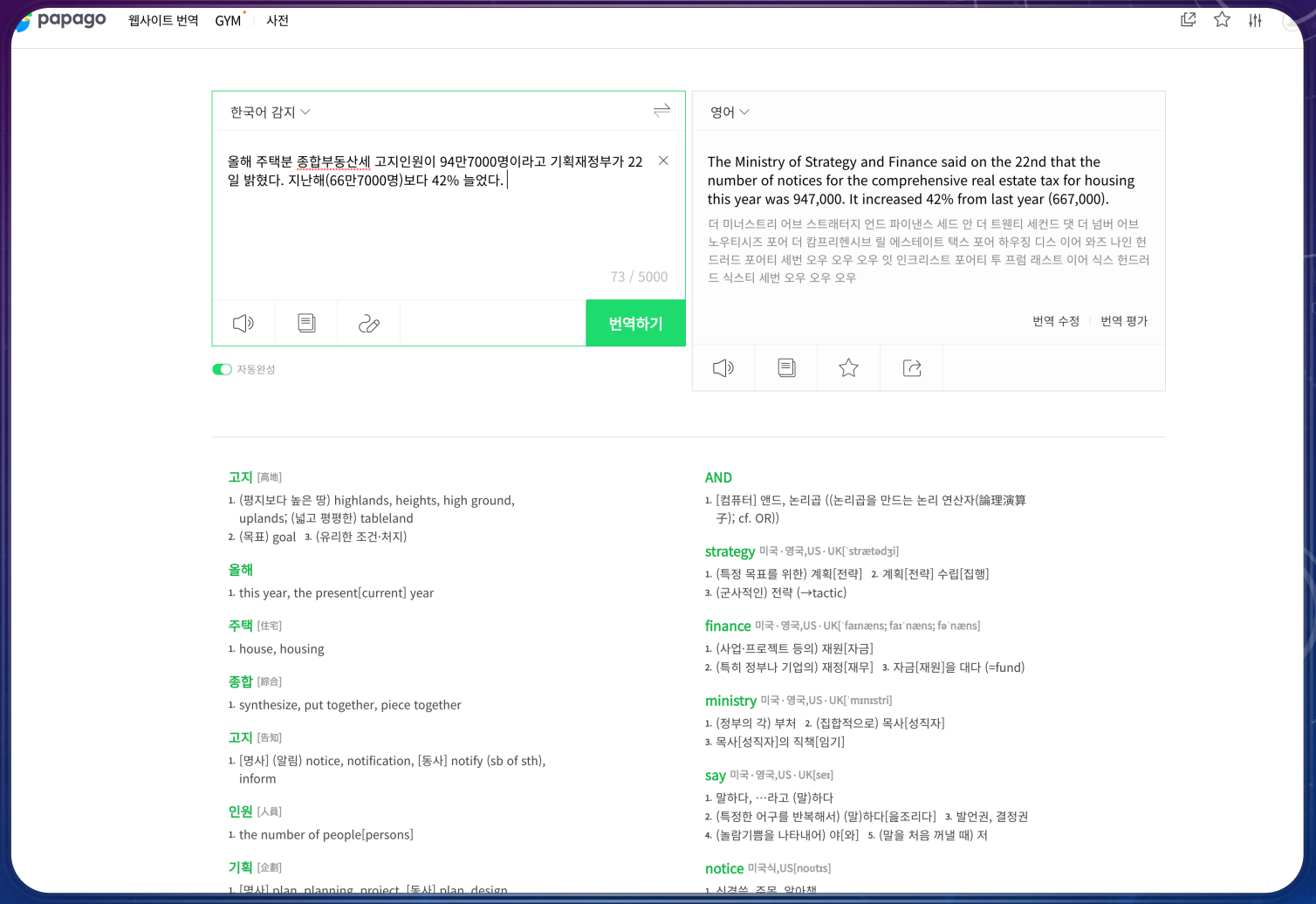
***배송 배송은 엄청 빨라요. 전국**

*배송
배송은 엄청 빨라요. 전국품절이라 오래 걸릴 줄 알았는데 5일 이내에 발송되었습니다. 배송기사님들 2명 오셔서 설치해주고 가셨어요. 폐가전수거는 하루 전이나 당일 아침에 전화오시면 이야기하시면 됩니다. 폐가전수거도 깔끔하게 해주셨어요.
*상품
상품상태는 아주 좋아요. 다른 사이트에서 살까하다가 ssg를 선택한 것은 정품, 서비스 면에서 신뢰가 갔기 때문이에요. 인증샵에서 배송된다고 해서 더욱 만족해요.

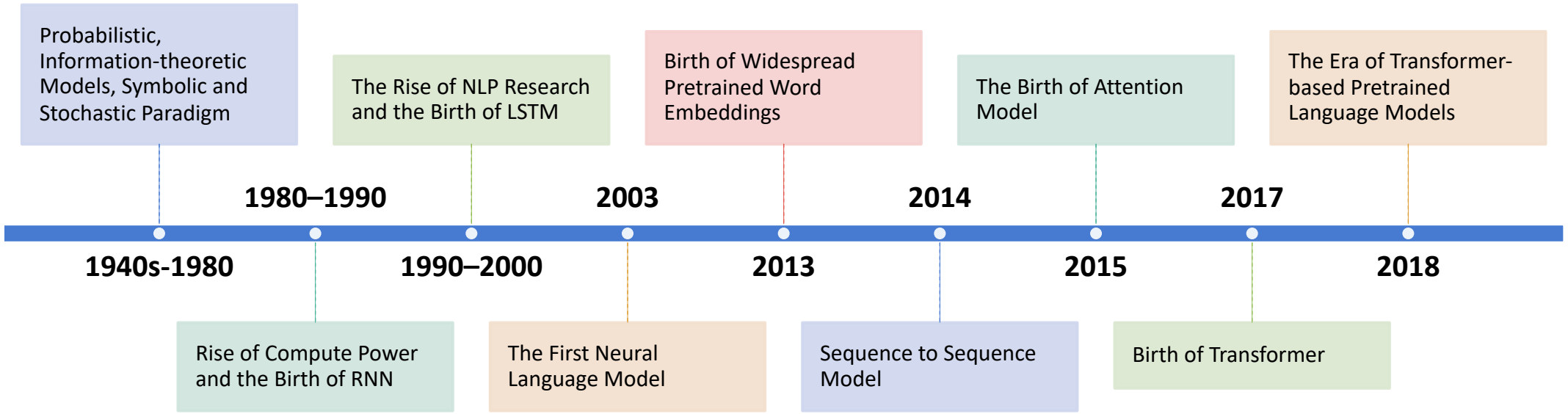
리뷰펼치기

02. 자연어처리 응용 분야

- Machine Translation

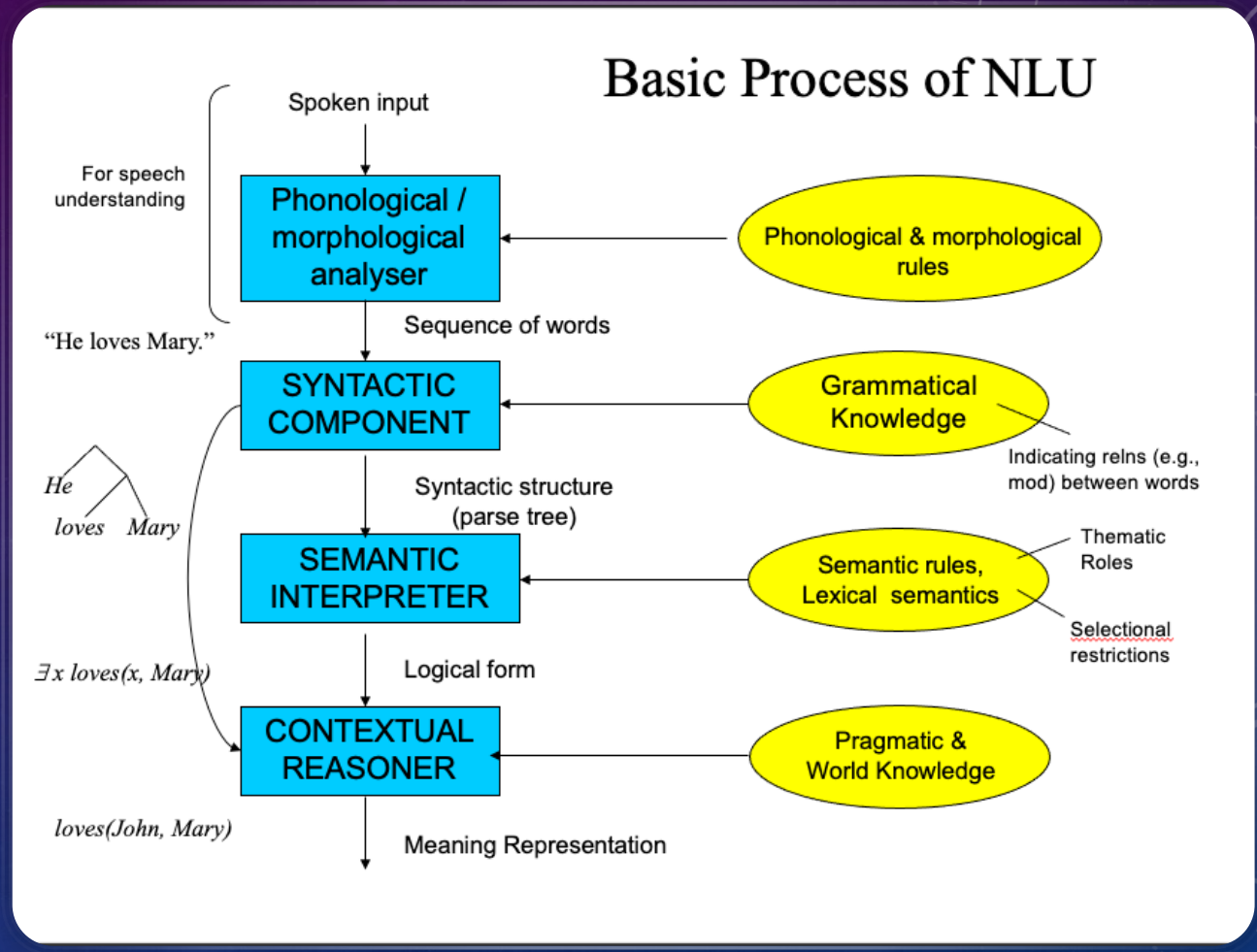


03. 자연어처리는 어떻게 발전되어 왔나?



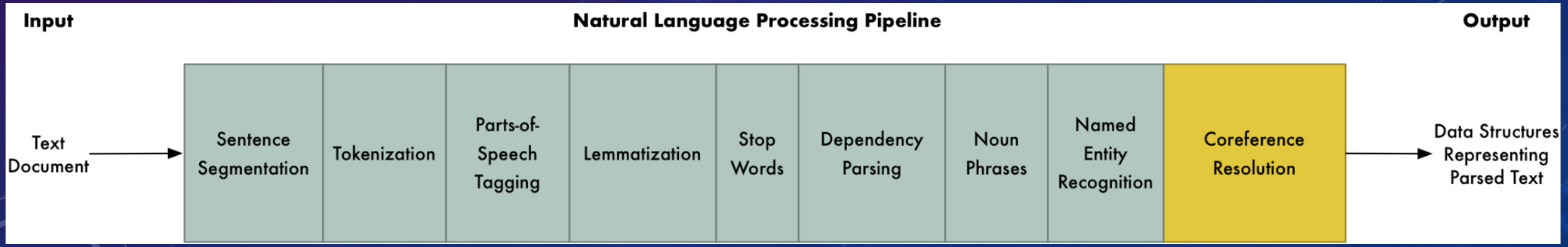
04. 자연어처리는 어떻게 이루어지나

- Top Down, Linguistic Analysis



04. 자연어처리는 어떻게 이루어지나

- Top Down, Linguistic Analysis
  - Image source (https://miro.medium.com/max/2000/1*zHLs87sp8R61ehUoXepWHA.png)



04. 자연어처리는 어떻게 이루어지나

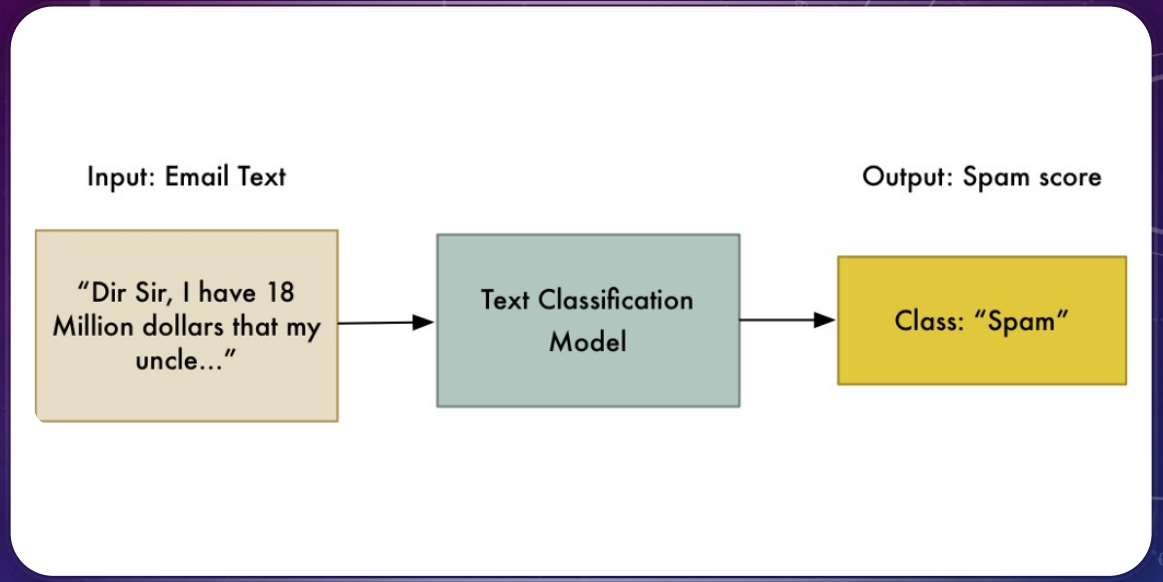
- Bottom up, Data Driven, Classification

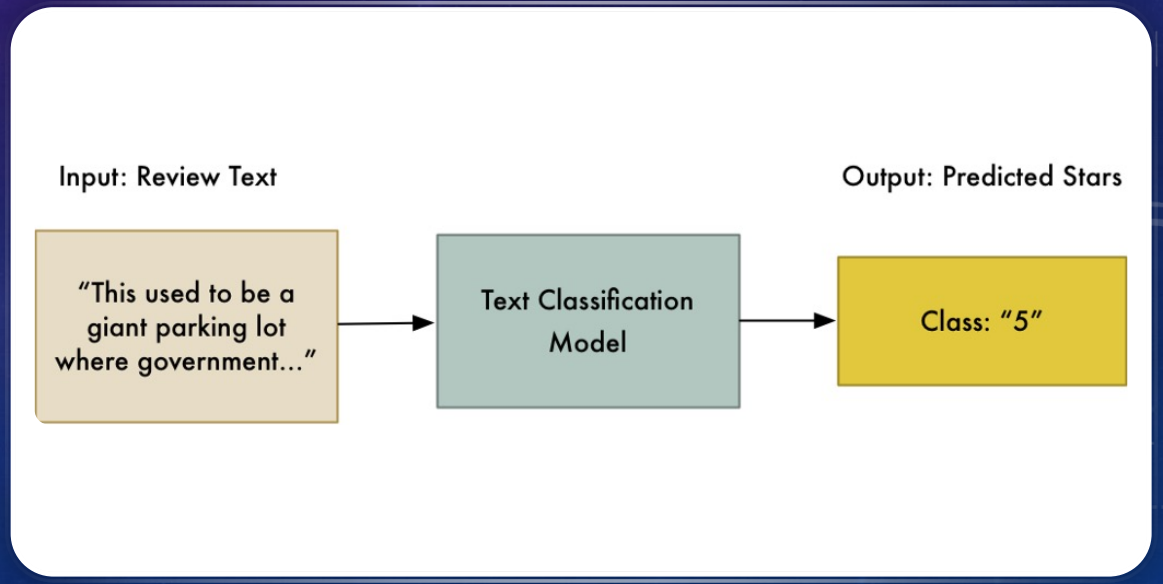
04. 자연어처리는 어떻게 이루어지나

BOTTOM UP, DATA DRIVEN, CLASSIFICATION

HTTPS://MIRO.MEDIUM.COM/MAX/1000/1*X93SFURVT_BMOMLZHGEGLW.PNG

HTTPS://MIRO.MEDIUM.COM/MAX/1000/1*HV4SHRJHK4J9ZW27QB_R-G.PNG





05. 자연어처리는 왜 어려운가

- Ambiguity (중의성): 언어의 모든 층위의 중의성
  - 어휘적 층위: '감기'?, 'Apple'?
  - 통사구조는 의미에 영향을 미친다
    - Flying planes is dangerous
    - Flying Planes are dangerous
    - Teacher Strikes Idle Kids
  - 의미와 세상지식(World Knowledge)는 통사구조에 영향을 미친다
    - *Flying insects is dangerous
    - Flying insects are dangerous
    - I saw the Grand Canyon flying to LA
    - I saw a condor flying to LA

05.
자연어처리는 왜 어려운가

비정형데이터

- accomplished! U taught us 2
- should never give up either❤️

Segmentation Issue

- The New York-New Haven Railroad

Idioms

- Dark horse
- 손이 크다

신조어

- Unfriend/Retweet/bromance/

World Knowledge

Tricky Entity Name

- Let it Be was recorded….

06. 언어모델(LANGUAGE MODEL)

- Predict Next Word
  - P(새빨간 거짓말) > P(새빨간 희망)
  - P(이 강의는 참 재미있어요)
- Assign a Probability to a sentence
  - N-gram based : Unigram, Bigram, Trigram…
  - Word Embedding
  - Transformer based Language Modeling

06. 언어모델(LANGUAGE MODEL)

- 문장의 확률을 어떻게?
  - 조건부 확률: P(B|A) = P(A, B)/P(A) → P(A, B) = P(A)p(B|A)
    - P(A, B, C, D) = P(A)P(B|A)P(C|A, B)P(D|A, B, C)
  - Chain Rule:
    - $P(X_1, X_2, X_3, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \ldots P(X_n|X_1, \ldots X_{n-1})$

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i | w_1 w_2 \ldots w_{i-1})$$

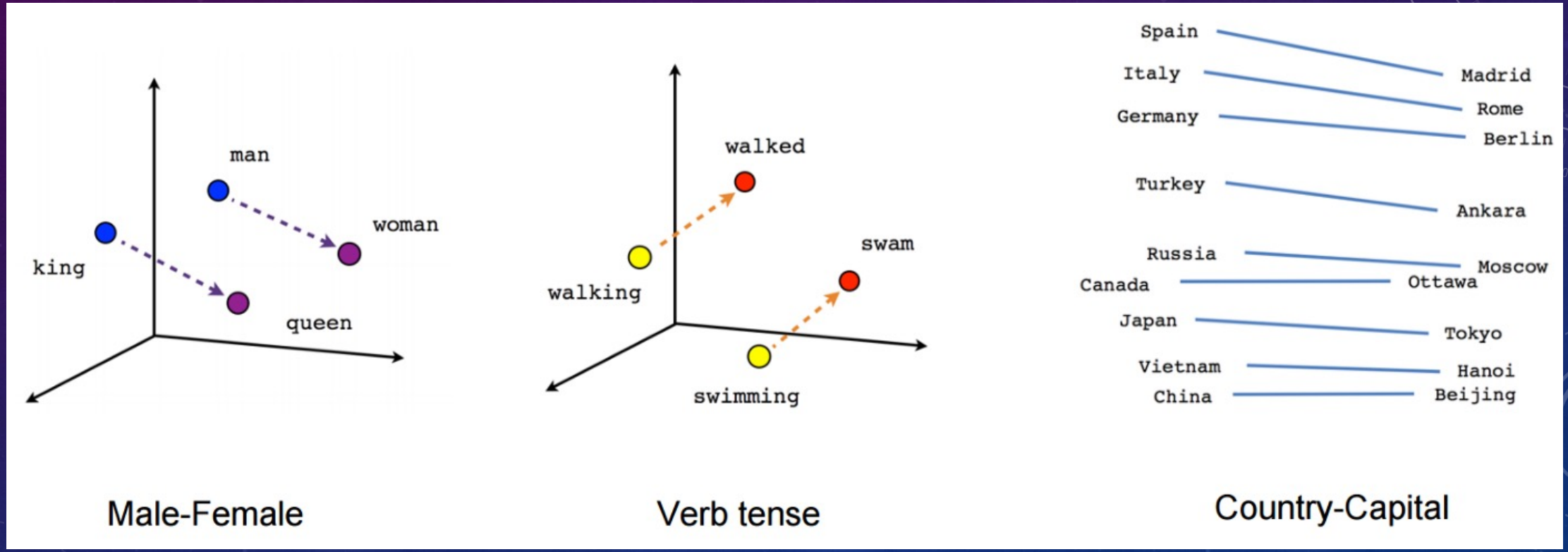
    - P("이 강의는 참 재미 있어요") = P(이) X P(강의는|이) X P(참|이 강의는) X P(재미 | 이 강의는 참) X P(있어요 | 이 강의는 참 재미)
    - P(있어요 | 이 강의는 참 재미) = Count(이 강의는 참 재미 있어요) / Count(이 강의는 참 재미)
- Markov Assumption
  - P(있어요| 이 강의는 참 재미) ≈ P(있어요 | 재미) 또는 P(있어요 | 참 재미)
  - P("이 강의는 참 재미 있어요") = P(이) X P(강의는|이) X P(참|강의는) X P(재미 | 참) X P(있어요 | 재미)

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i | w_{i-k} \ldots w_{i-1})$$

07. 단어 임베딩(WORD EMBEDDING)

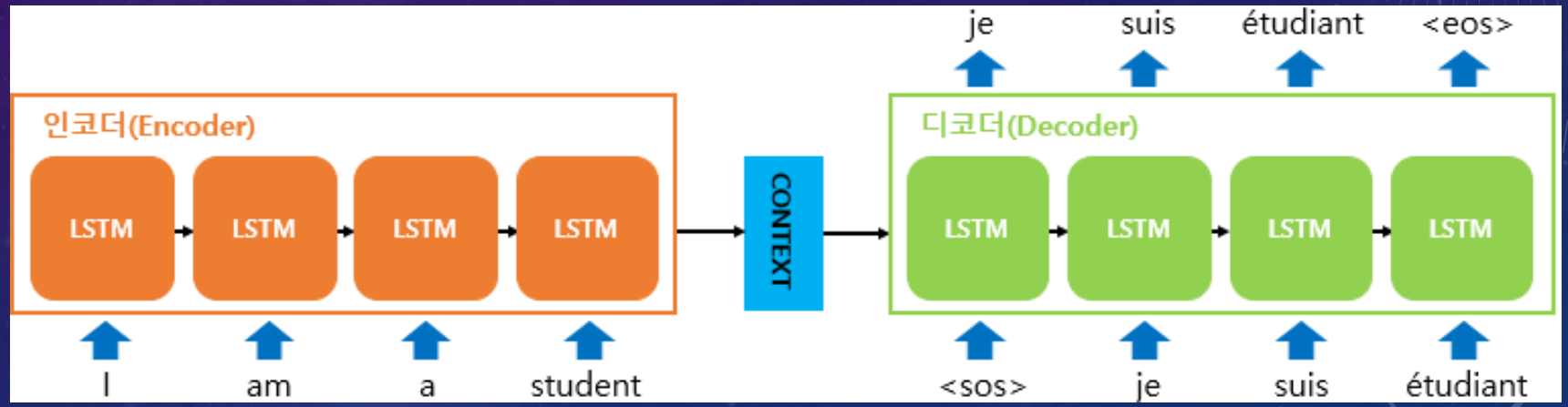
- Numericalization : 텍스트를 숫자로 바꾸는 방법
  - One-hot encoding, N-Gram 언어 모델, Bag of Words 언어 모델…
- Vector 의미론(Semantics)
  - 단어의 의미를 어떻게 표상할 수 있는가?
    - 동음이의어, 유사어, 다의어, 관련어…
- Word Embedding
  - Distributional Hypothesis
  - Word2Vec, Glove, FaxtText..
  - Static Word Embedding (다의어, 동의어 구별이 되지 않음)

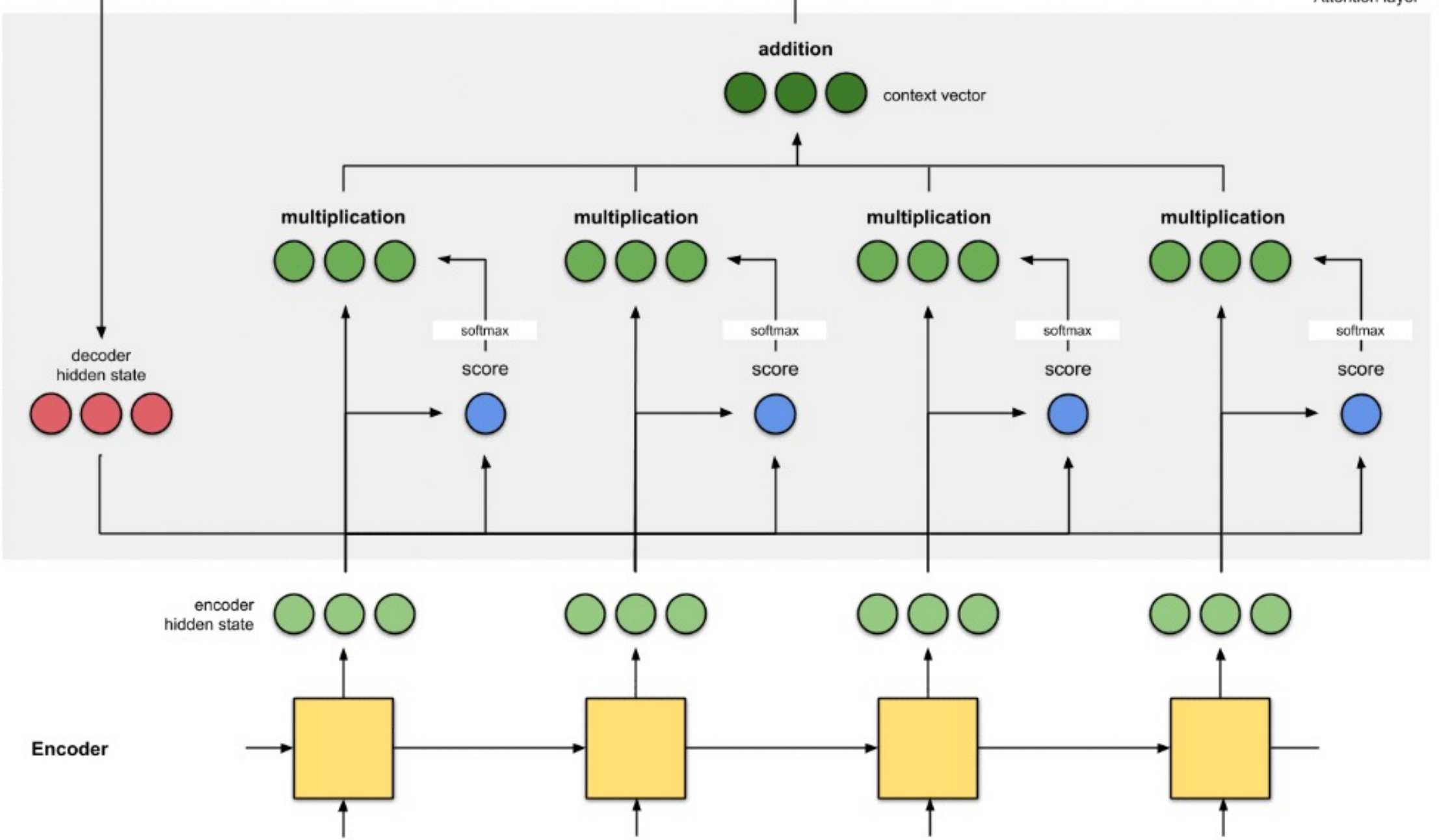
07. 단어 임베딩


man
woman
king
queen
walked
swam
walking
swimming
Spain
Madrid
Italy
Rome
Germany
Berlin
Turkey
Ankara
Russia
Moscow
Canada
Ottawa
Japan
Tokyo
Vietnam
Hanoi
China
Beijing
Male-Female
Verb tense
Country-Capital

08. 시퀀스투시퀀스 /어텐션(SEQUENCE TO SEQUENCE/ATTENTION)

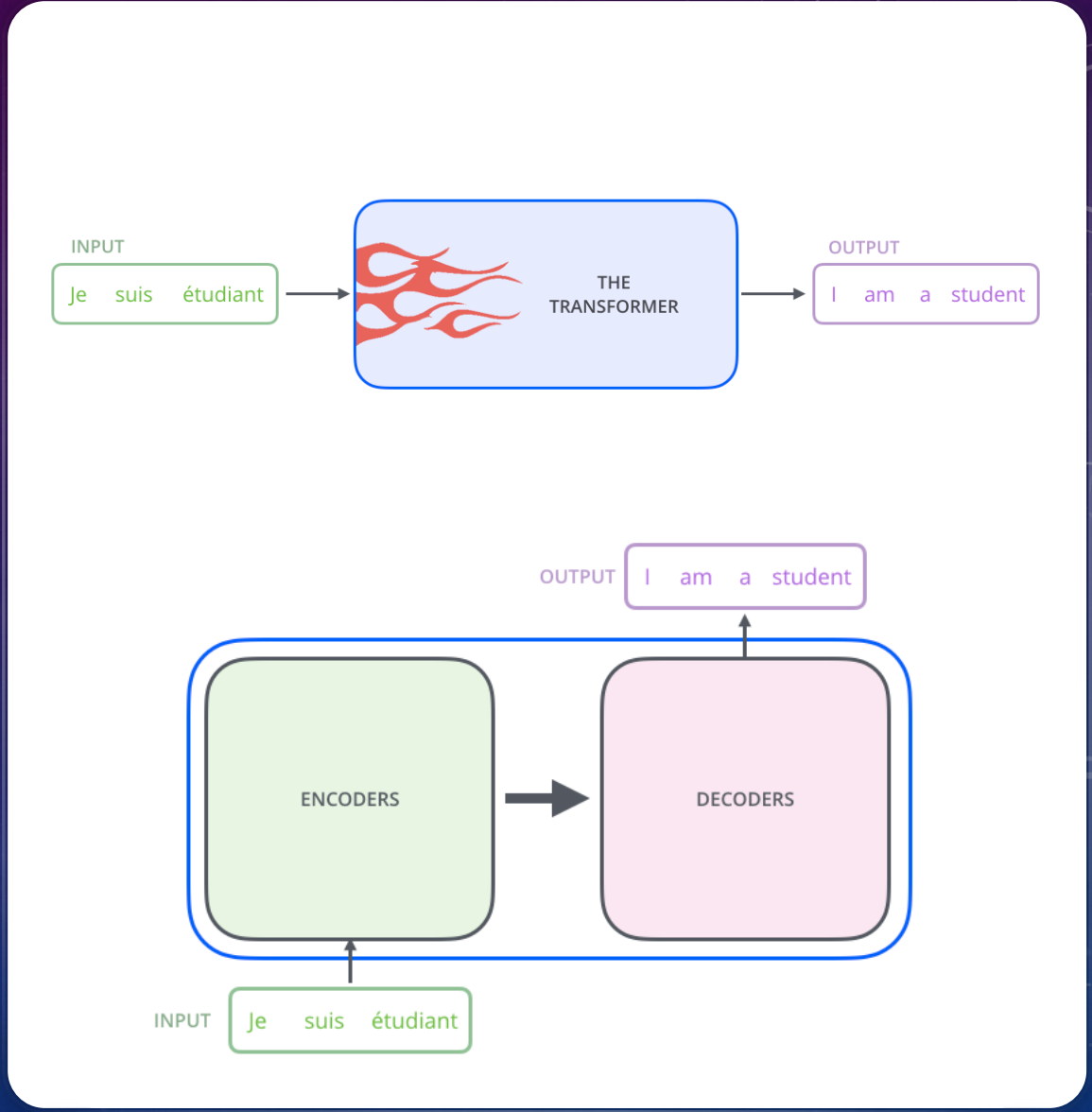
- Sequence to sequence Model
- Encoder-Decoder




addition
context vector
multiplication
multiplication
multiplication
multiplication
softmax
softmax
softmax
softmax
score
score
score
score
decoder
hidden state
encoder
hidden state
Encoder

09. 트랜스포머

- https://jalammar.github.io/images/t/the_transformer_3.png
- https://jalammar.github.io/images/t/The_transformer_encoders_decoders.png



09. 트랜스포머

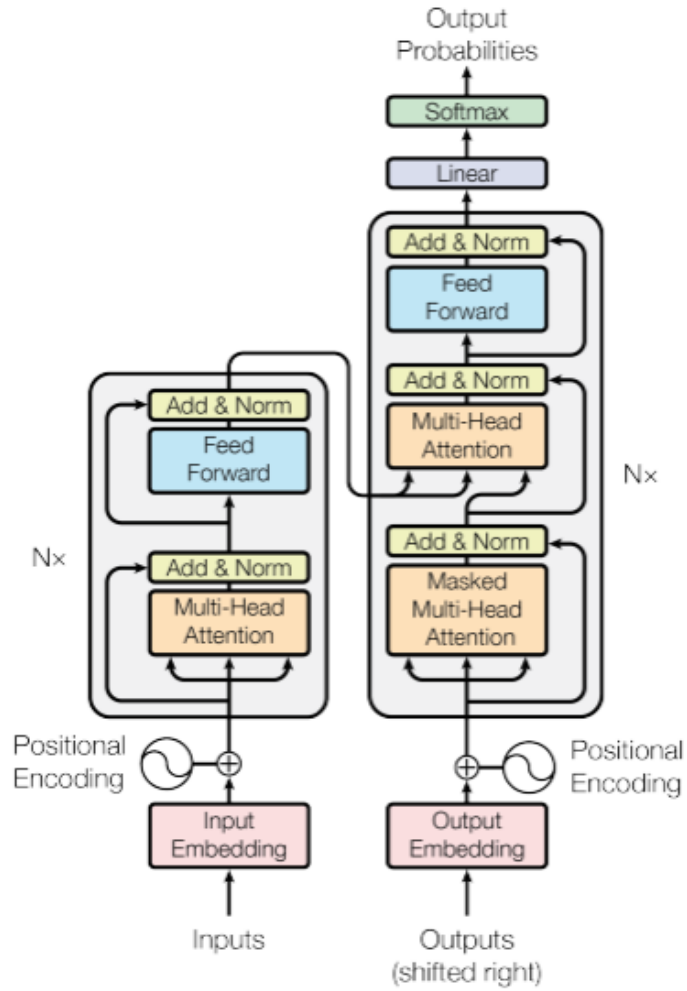
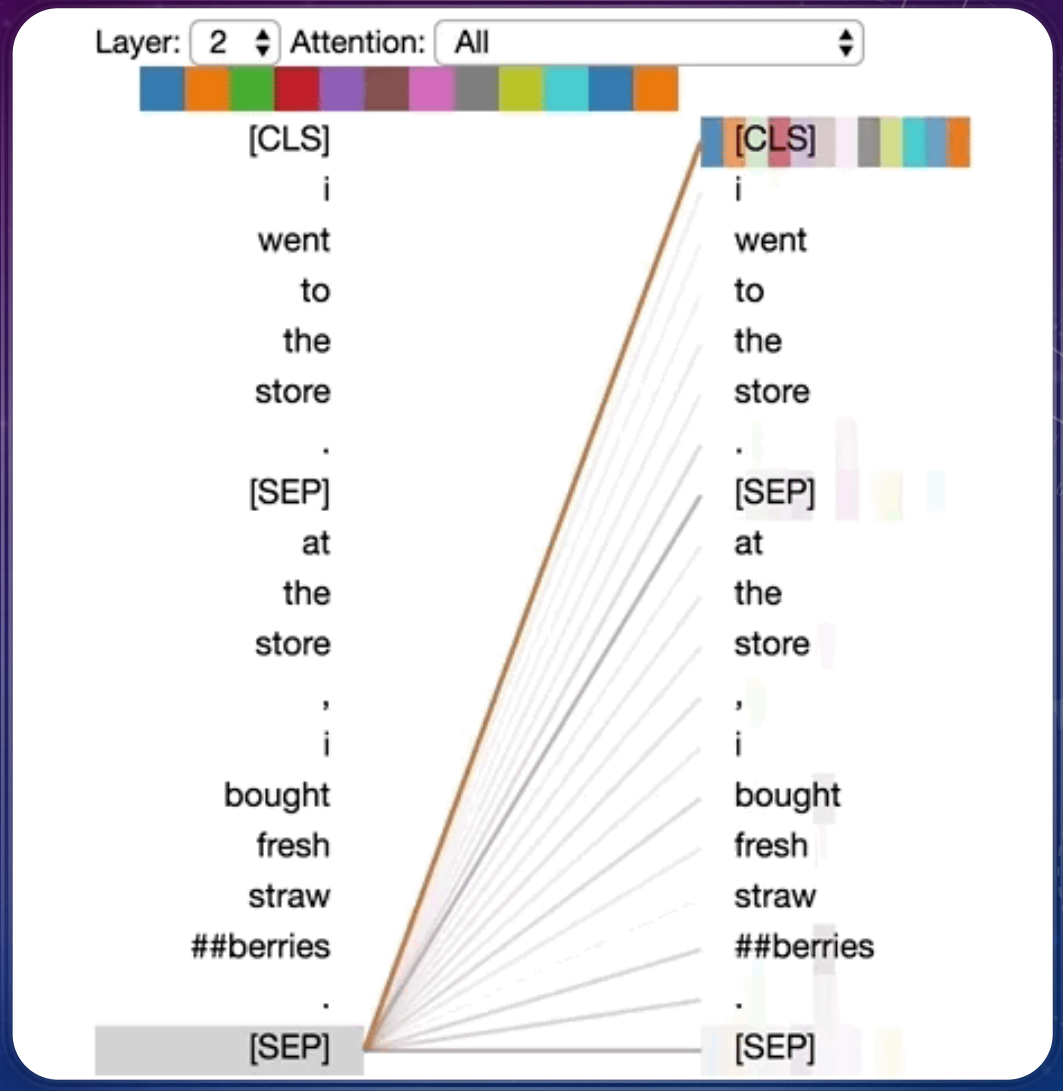


Figure 1: The Transformer - model architecture.

09. 트랜스포머

- Self Attention
  - https://towardsdatascience.com/deconstructing-bert-distilling-6-patterns-from-100-million-parameters-b49113672f77

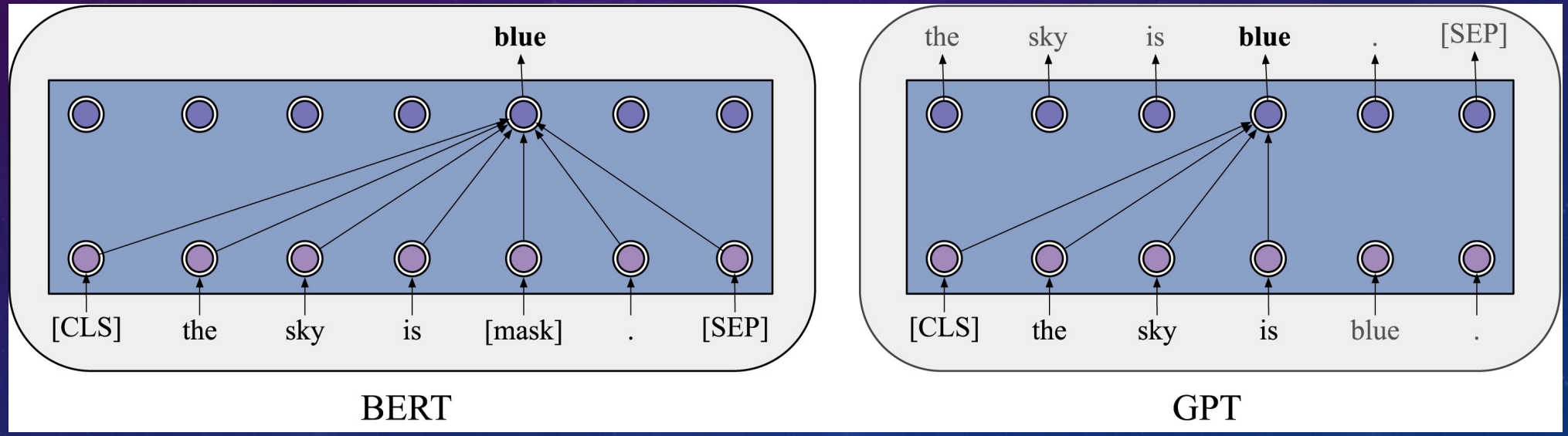


10. 사전학습모델

- Decoders or autoregressive models:
  - GPT, GPT-2, GPT3. CTRL(Conditional Transformer Language Model for Controllable Generation, Transformer-XL, Reformer, XLNet
- Encoders or autoencoding models
  - BERT, ALBERT, RoBERTa, DistilBERT, ConvBERT, XLM, XLM-RoBERTa, FlauBERT, ELECTRA, Funnel Transformer, Longformer
- Sequence-to-sequence models
  - BART, Pegasus, MarianMT, T5, MT5, Mbart, ProphetNet, XLM-ProphetNet
- Multimodal models
  - MMBT
- Retrieval-based models
  - DPR, RAG

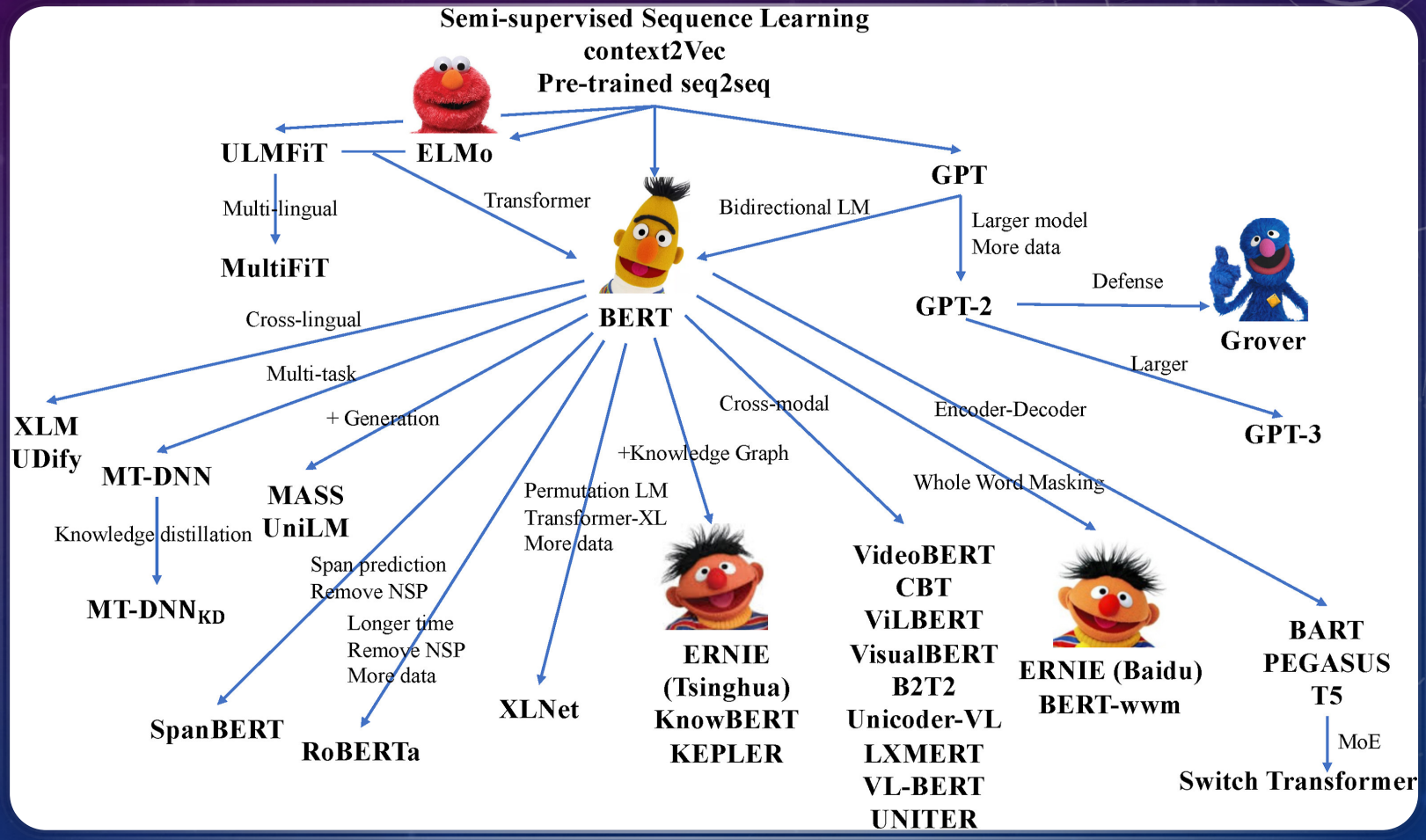
10. 사전학습모델

- BERT and GPT (Generative Pre-Training)
- Auto Encoder Model vs. Auto Regressive Model



10. 사전학습모델

- https://ars.els-cdn.com/content/image/1-s2.0-S2666651021000231-gr8.jpg



거대언어모델(LARGE LANGUAGE MODEL)

사전학습모델(Pre-trained Language Model)의 모델크기나 데이터 크기를 확장하면 downstream task에서 모델 용량이 향상됨

- LLM은 기존 언어모델에서 볼 수 없던 새로운 능력을 보여줌
- LLM은 인간이 AI를 개발하고 사용하는 방식에 변화를 가져옴
  - LLM에 접근하는 방식은 prompt를 통해서 주로 이루어짐
  - 따라서 LLM이 이해할 수 있는 방식으로 지시문(instruction)을 만들어야 함
- chatGPT와 GPT-4의 출현으로 Artificial General Intelligence의 가능성 모색됨

그럼에도 불구하고 LLM의 기본원리는 잘 밝혀지지 않음

- 작은 PLM이 아닌 LLM에서 새로운 기능이 발생하는지 의문
- LLM을 구축하기 위해서는 엄청난 양의 데이터와 고성능의 GPU가 대량으로 필요하기 때문에 훈련 비용이 많이 듬
  - 몇 거대 기업들만 훈련 가능한 상황
  - 구축된 모델 및 사용된 데이터셋의 비공개

LLM의 배경

- 일반적으로 거대언어모델은 트랜스포머(Transformer) 아키텍쳐를 기반으로 수천억 또는 그 이상의 parameter를 포함하는 언어모델을 지칭
  - Scaling: 거대언어모델의 능력향상은 모델의 크기에 따라 성능이 대략적으로 증가하는 scaling 법칙으로 부분적으로 설명가능.
  - Emergent Ability: “작은 모델에는 존재하지 않지만 큰 모델에서 발생하는 능력”. 규모가 일정수준에 도달하면 성능이 무작위보다 크게 상승함
    1. In-Context Learning(ICL): GPT-3에서 처음으로 도입. 언어모델에 자연어 지시 또는 여러 작업의 데모가 제공되었다면, 추가 학습이나 gradient 업데이트 없이 입력 테스트의 순서를 완성하여 예상 출력을 생성함
    2. Instruction-following: 자연어 설명을 통해 포맷된 다중 작업 데이터셋을 혼합하여 미세조정(instruction tuning)하면, LLM은 명령어 형태로 설명되는 보이지 않는 작업에서도 잘 수행됨
    3. Step-by-step Reasoning: 수학 연산, 논리적 추론 등을 위해 사고의 연쇄(Chain-of-Thoughts)를 사용하여, 최종 답을 도출하기 위한 중간 추론 단계가 포함된prompt를 활용하여 적용

LLM의 주요 기술(KEY TECHNICS)

- Scaling:
  - 더 큰 모델/데이터/학습이 모델의 성능을 향상. GPT-3 175B, PaLM 540B Parameter 사용
- Training:
  - 모델크기가 엄청나기 때문에 성공적인 LLM을 잘 학습하기가 쉽지 않음.
  - 분산훈련, 병렬 훈련(DeepSpeed, Megatron-LM) 등의 프레임 개발.
  - 최적화(optimization) 방법론 개발
- Ability Eliciting: 범용 작업능력을 끌어내기 위한 방법.
  - CoT(Chain of Thoughts), Instruction tuning…
- Alignment Tuning : 학습된 데이터에 따라 유해하거나 편향된 내용을 생성할 수 있음
  - InstructGPT: 인간의 피드백을 통한 강화학습(RLHF)을 적용하여 LLM에 예상되는 지침을 따르도록 함
- Tools manipulation: LLM은 방대한 일반 텍스트 말뭉치에 대해서 텍스트 생성기를 위해 훈련되었기 때문에 텍스트 형태로 잘 표현되지 않는 작업 (숫자 계산 등)에서는 성능이 떨어짐
  - 이를 해결하기 위해 외부도구를 활용해 LLM의 결함을 보완
  - 외부 계산기나, 검색 엔진을 사용, 외부 플러그인 결합
  - 이외에도 많은 다른 요인 (하드웨어 업그레이드)

LIST OF LARGE LANGAUGE MODELS (HTTPS://EN.WIKIPEDIA.ORG/WIKI/LARGE_LANGUAGE_MODEL)

Name	Release date[a]	Developer	Number of parameters[b]	Corpus size	License[c]	Notes
BERT	2018	Google	340 million[55]	3.3 billion words[55]	Apache 2.0[56]	An early and influential language model,[2] but encoder-only and thus not built to be prompted or generative[57]
XLNet	2019	Google	~340 million[58]	33 billion words		An alternative to BERT; designed as encoder-only[59][60]
GPT-2	2019	OpenAI	1.5 billion[61]	40GB[62] (~10 billion tokens)[63]	MIT[64]	general-purpose model based on transformer architecture
GPT-3	2020	OpenAI	175 billion[24]	300 billion tokens[63]	public web API	A fine-tuned variant of GPT-3, termed GPT-3.5, was made available to the public through a web interface called ChatGPT in 2022.[65]
GPT-Neo	March 2021	EleutherAI	2.7 billion[66]	825 GiB[67]	MIT[68]	The first of a series of free GPT-3 alternatives released by EleutherAI. GPT-Neo outperformed an equivalent-size GPT-3 model on some benchmarks, but was significantly worse than the largest GPT-3.[68]
GPT-J	June 2021	EleutherAI	6 billion[69]	825 GiB[67]	Apache 2.0	GPT-3-style language model
Megatron-Turing NLG	October 2021[70]	Microsoft and Nvidia	530 billion[71]	338.6 billion tokens[71]	Restricted web access	Standard architecture but trained on a supercomputing cluster.
Ernie 3.0 Titan	December 2021	Baidu	260 billion[72]	4 Tb	Proprietary	Chinese-language LLM. Ernie Bot is based on this model.
Claude[73]	December 2021	Anthropic	52 billion[74]	400 billion tokens[74]	Closed beta	Fine-tuned for desirable behavior in conversations.[75]
GLaM (Generalist Language Model)	December 2021	Google	1.2 trillion[76]	1.6 trillion tokens[76]	Proprietary	Sparse mixture-of-experts model, making it more expensive to train but cheaper to run inference compared to GPT-3.
Gopher	December 2021	DeepMind	280 billion[77]	300 billion tokens[78]	Proprietary	
LaMDA (Language Models for Dialog Applications)	January 2022	Google	137 billion[79]	1.56T words,[79] 168 billion tokens[78]	Proprietary	Specialized for response generation in conversations.
GPT-NeoX	February 2022	EleutherAI	20 billion[80]	825 GiB[67]	Apache 2.0	based on the Megatron architecture
Chinchilla	March 2022	DeepMind	70 billion[81]	1.4 trillion tokens[81][78]	Proprietary	Reduced-parameter model trained on more data. Used in the Sparrow bot.
PaLM (Pathways Language Model)	April 2022	Google	540 billion[82]	768 billion tokens[81]	Proprietary	aimed to reach the practical limits of model scale
OPT (Open Pretrained Transformer)	May 2022	Meta	175 billion[83]	180 billion tokens[84]	Non-commercial research[d]	GPT-3 architecture with some adaptations from Megatron
YaLM 100B	June 2022	Yandex	100 billion[85]	1.7TB[85]	Apache 2.0	English-Russian model based on Microsoft's Megatron-LM.

LIST OF LARGE LANGUAGE MODELS

Minerva	June 2022	Google	540 billion[86]	38.5B tokens from webpages filtered for mathematical content and from papers submitted to the arXiv preprint server[86]	Proprietary	LLM trained for solving "mathematical and scientific questions using step-by-step reasoning".[87] Minerva is based on PaLM model, further trained on mathematical and scientific data.
BLOOM	July 2022	Large collaboration led by Hugging Face	175 billion[88]	350 billion tokens (1.6TB)[89]	Responsible AI	Essentially GPT-3 but trained on a multi-lingual corpus (30% English excluding programming languages)
Galactica	November 2022	Meta	120 billion	106 billion tokens[90]	CC-BY-NC-4.0	Trained on scientific text and modalities.
AlexaTM (Teacher Models)	November 2022	Amazon	20 billion[91]	1.3 trillion[92]	public web API[93]	bidirectional sequence-to-sequence architecture
LLaMA (Large Language Model Meta AI)	February 2023	Meta	65 billion[94]	1.4 trillion[94]	Non-commercial research[e]	Trained on a large 20-language corpus to aim for better performance with fewer parameters.[94] Researchers from Stanford University trained a fine-tuned model based on LLaMA weights, called Alpaca.[95]
GPT-4	March 2023	OpenAI	Exact number unknown, approximately 1 trillion [f]	Unknown	public web API	Available for ChatGPT Plus users and used in several products.
Cerebras-GPT	March 2023	Cerebras	13 billion[97]		Apache 2.0	Trained with Chinchilla formula.
Falcon	March 2023	Technology Innovation Institute	40 billion[98]	1 Trillion tokens (1TB)[98]	Apache 2.0[99]	The model is claimed to use only 75% of GPT-3's training compute, 40% of Chinchilla's, and 80% of PaLM-62B's.
BloombergGPT	March 2023	Bloomberg L.P.	50 billion	363 billion token dataset based on Bloomberg's data sources, plus 345 billion tokens from general purpose datasets[100]	Proprietary	LLM trained on financial data from proprietary sources, that "outperforms existing models on financial tasks by significant margins without sacrificing performance on general LLM benchmarks"
PanGu-Σ	March 2023	Huawei	1.085 trillion	329 billion tokens[101]	Proprietary	
OpenAssistant[102]	March 2023	LAION	17 billion	1.5 trillion tokens	Apache 2.0	Trained on crowdsourced open data
PaLM 2 (Pathways Language Model 2)	May 2023	Google	340 billion[103]	3.6 trillion tokens[103]	Proprietary	Used in Bard chatbot.[104]

상업적으로 사용 가능한 공개 언어 모델

	License	Data	Architecture	Weights	Size	Checkpoints	Language
Meta Llama2	Llama license	Open	Open	Open	7, 13, 70	Yes	English / Multilingual
EleutherAI Pythia	Apache 2.0	Open	Open	Open	7, 12	Yes	English
EleutherAI Polyglot	GPL-2.0	Open	Open	Open		Yes	English / Multilingual
GPT-J	MIT	Open	Open	Open	6	Yes	English
Databricks Dolly 2	Apache 2.0	Open	Open	Open	7, 12	Yes	English
Cerebras-GPT	Apache 2.0	Open	Open	Open	7, 13	Yes	English / Multilingual
StableLM	CC BY-SA-4.0	Open	Open	Open	3, 7, (15, 30, 65, 175)	Yes	English
Mosaic MPT	Apache 2.0	Open	Open	Open	7, 30	Yes	English
Falcon GPT	Apache 2.0	Open	Open	Open	7, 40	Yes	English

TECHNICAL EVOLUTION OF GPT-SERIES MODELS

A timeline of Existing Large Language Models(Zhao et al. (2023))