

인공지능에서의 언어 처리: 컴퓨터언어학(자연어처리)

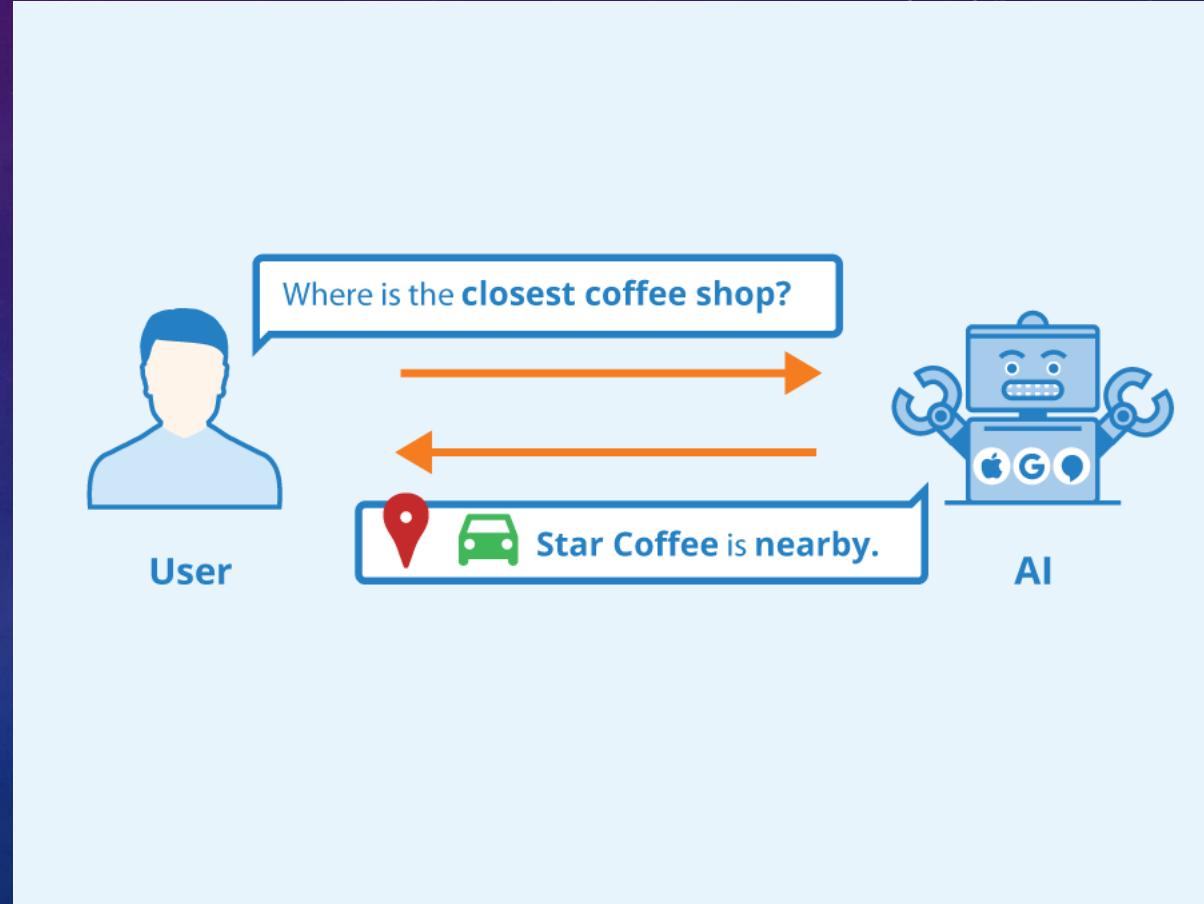
서울대학교 언어학과

신효필

HPSHIN@SNU.AC.KR

01. 컴퓨터언어학(자연어처리)이란?

- 컴퓨터언어학(Computational Linguistics)
또는 자연어처리(Natural Language Processing)
 - 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. 컴퓨터언어학(자연어처리)이란?

- 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. 컴퓨터언어학 응용 분야는?

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

02. 컴퓨터언어학 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. 컴퓨터언어학 응용 분야

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

전체 담장 | 삭제 보관 이동 | 업무로 등록 | ...

보낸 사람 교무과 (11.23 13:25)
받는 사람 신효필님

2021학년도부터 추진 중인 <Inno-Edu: 2031 서울대 창의교육프로젝트> 사업 설명회를 개최하오니 학내 구성원들의 많은 참석을 부탁드립니다.

○ 일시: 2021. 11. 24.(수) 12:30~13:30
○ 장소: 기초교육원(61동) 320호
* 코로나19 기본 방역수칙을 준수하여 100명 미만으로 참석 인원을 제한하고, 행사장 내 취식을 금지합니다.
* 행사 종료 후 샌드위치를 제공할 예정입니다.

Inno-Edu 2031: 서울대 창의교육프로젝트 설명회 개최

일시 | 2021. 11. 24.(수) 12:30~13:30
장소 | 기초교육원(61동) 320호
주관 | 교무처 교무과
문의 | 02-880-2077, sekim33@snu.ac.kr
* 현장 참석 인원은 100명 미만으로 제한되며, 샌드위치를 제공할 예정입니다.

본 메일은 서울대학교 대량메일시스템을 통해 발송된 메일입니다. 메일 수신을 원치 않으시면 수신거부 를 클릭하십시오.
본인의 수신거부 목록을 확인하시려면 수신거부목록 확인 을 클릭하십시오.
This email has been sent through the SNU mass mailing system.
If you do not wish to receive emails of this category, please click [here](#).
To check the emailing lists you have unsubscribed from, please click [here](#).

서울대학교
SEOUL NATIONAL UNIVERSITY

08826 서울시 관악구 관악로 1 서울대학교
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Copyright 2012 Seoul National University All Rights Reserved

02. 컴퓨터언어학 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. 양문형이나 4도어나의 고민에서부터 세미빌트인이나 빌트인이나 까지... 결국 저의 모든 것을 만족시켜주는 건 Ig 디오스 양문형 냉장고더라고요^^ 저는 1인가구라 냉동실 사용이 주를 이를 수 밖에 없고 냉장고의 주요 사용 용도는 마실 것의 보관이었기에 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 스텀 매직스페이스가있어 음료수 물 편하게 이용하고 에너지절감도 되고부모님께 선물했는데 아주 만족 좋아하십니다다만 문쪽 선반이 작은거라도 맨아랫까지있음 더알찰텐데 조금이아슬네요냉동실 얼음얼리는곳시 크진않지만 두분이 쓰기단촐을만큼되네요 강추~

리뷰접기 ^

★★★★★ 5 신세계몰 · cc***** · 19.10.11.

친정엄마 냉장고가 10년 다되어◆

친정엄마 냉장고가 10년 다되어가는 기준 양문형 냉장고였는데 고장은 없으니 냉동실이 좀아 냉기순환이 안되는지 공공 열지않아 선물 드렸는데 정말 마음에 들어하시네요 4도어 5도어가 대세이긴 하나 기능대비 가격 차이가 너무 많이 나서 선택했는데 대만족입니다 양문형이긴 하나 외관상으로 5도어처럼 보입니다ㅋㅋ 특히 매직스페이스가 정말 좋아요 예전 흠바보다 사이즈도 크고 냉기보존도 잘 되는듯요

리뷰펼치기 ▼

★★★★★ 5 신세계몰 · sa***** · 20.05.09.

*배송 배송은 엄청 빨라요. 전국

*배송

배송은 엄청 빨라요. 전국품질이라 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. 컴퓨터언어학 응용 분야

- Machine Translation

The screenshot shows the Papago machine translation interface. The input text in Korean is: "올해 주택분 종합부동산세 고지인원이 94만7000명이라고 기획재정부가 22일 밝혔다. 지난해(66만7000명)보다 42% 늘었다." The output text in English is: "The Ministry of Strategy and Finance said on the 22nd that the number of notices for the comprehensive real estate tax for housing this year was 947,000. It increased 42% from last year (667,000)." Below the main text, there are several definitions for Korean words like 고지, 주택, 종합, 기획, and 인원, each with its English meaning and example sentences.

한국어 감지 ▾

웹사이트 번역 GYM 사전

영어 ▾

번역 수정 번역 평가

고지 [高地]
1. (평지보다 높은 땅) highlands, heights, high ground, uplands; (**넓고 평평한**) tableland
2. (목표) goal 3. (유리한 조건·처지)

올해
1. this year, the present[current] year

주택 [住宅]
1. house, housing

종합 [綜合]
1. synthesize, put together, piece together

고지 [告知]
1. [명사] (알림) notice, notification, [동사] notify (sb of sth), inform

인원 [人員]
1. the number of people[persons]

기획 [企劃]
1. [명사] plan, planning, project, [동사] plan, design

AND

strategy 미국·영국·US·UK[ˈstrætədʒi]
1. (특정 목표를 위한) 계획[전략] 2. 계획[전략] 수립[집행]
3. (군사적인) 전략 (→tactic)

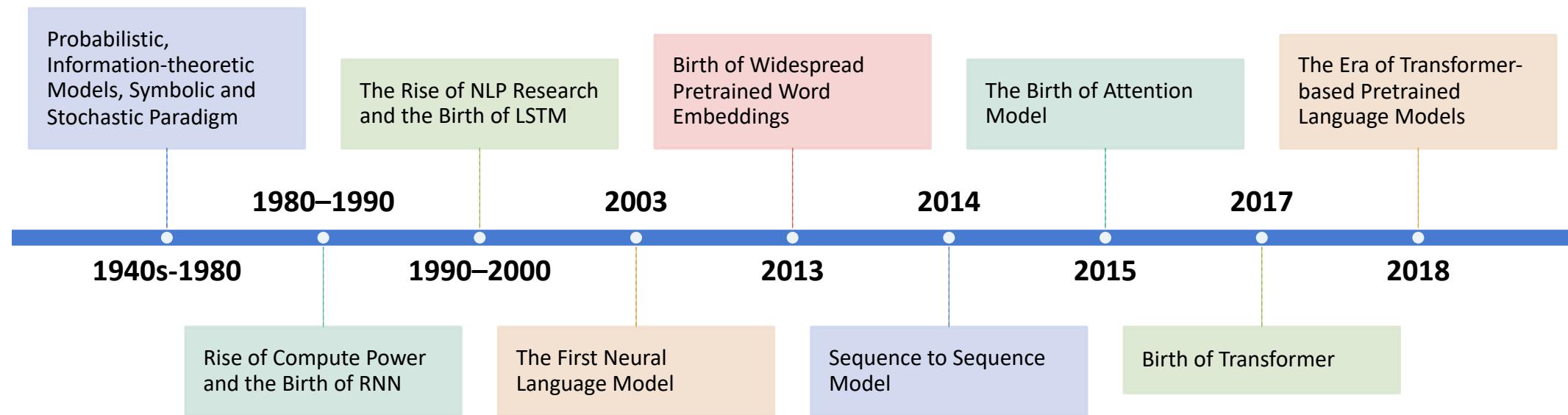
finance 미국·영국·US·UK[ˈfaɪnæns; fərˈnæns; fərˈnæns]
1. (사업·프로젝트 등의) 재원[자금]
2. (특히 정부나 기업의) 재정[재무] 3. 자금[재원]을 대다 (=fund)

ministry 미국·영국·US·UK[ˈmɪnɪstri]
1. (정부의 각) 부처 2. (집합적으로) 목사[성직자]
3. 목사[성직자]의 직책[임기]

say 미국·영국·US·UK[seɪ]
1. 말하다, …라고 (말)하다
2. (특정한 어구를 반복해서) (말)하다[읊조리다] 3. 발언권, 결정권
4. (놀랄 기쁨을 나타내어) 야[와] 5. (말을 처음 꺼낼 때) 저

notice 미국식·US[noʊtɪs]
1. 시경설, 주문, 알리새

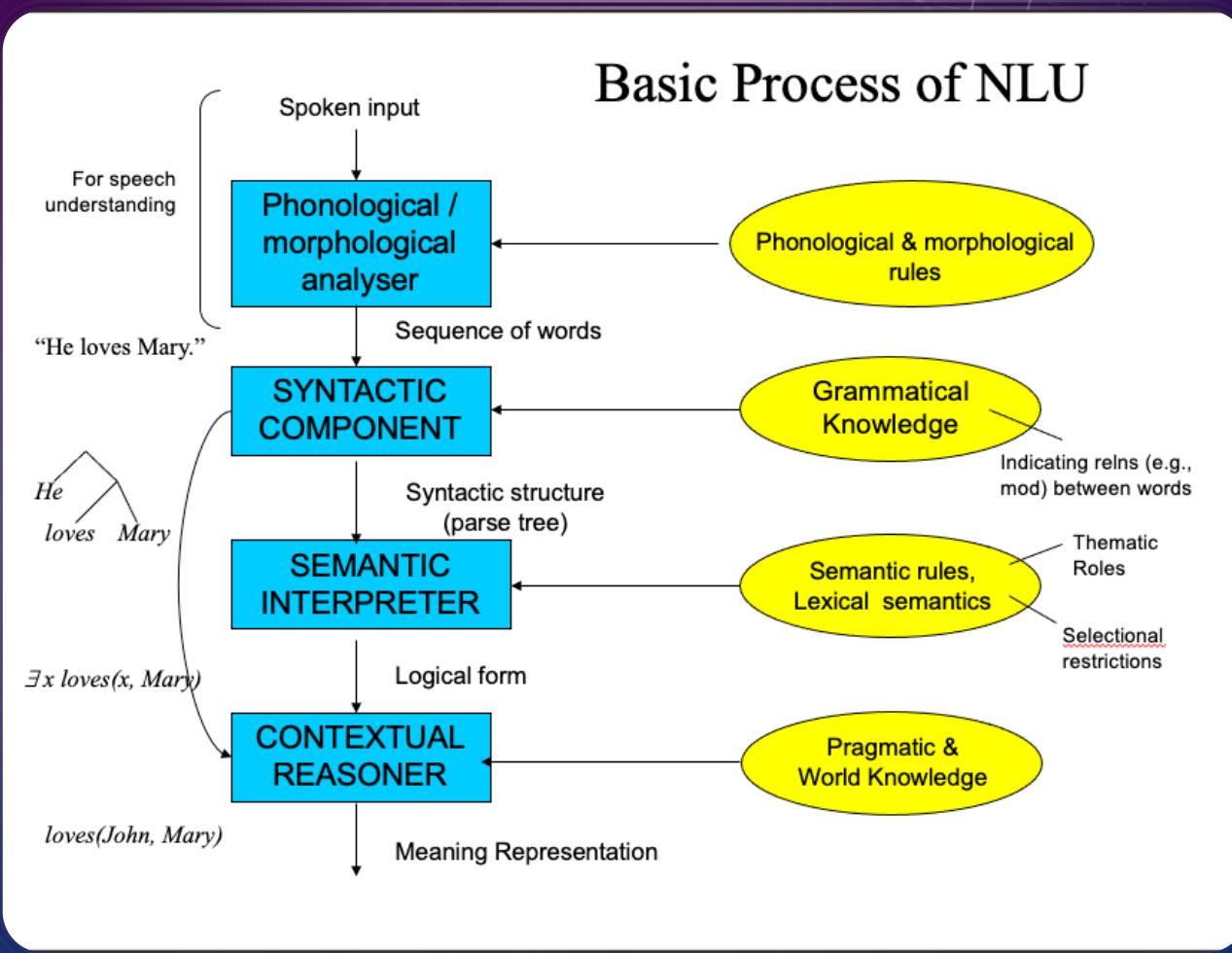
03. 컴퓨터언어학은 어떻게 발전되어 왔나?



04. 컴퓨터언어학은 어떻게 이루어지나

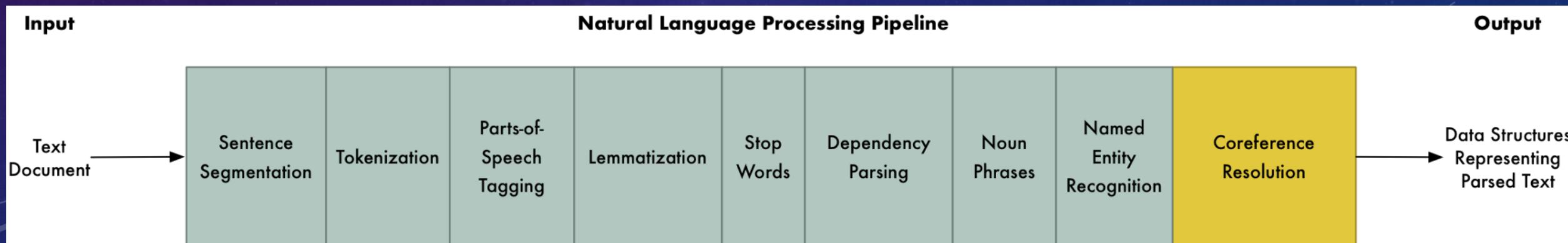
- Top Down, Linguistic Analysis

Basic Process of NLU



04. 컴퓨터언어학은 어떻게 이루어지나

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



04. 컴퓨터언어학은 어떻게 이루어지나

- Bottom up, Data Driven, Classification

The screenshot shows the NAVER 영화 (NAVER Movie) website interface. The left sidebar includes links for 영화홈, 상영작·예정작, 영화랭킹, 평점·리뷰 (selected), 네이버 평점, 네이버 리뷰, 다운로드, and 인디극장. The main content area displays a search bar for '네이버·관람객 평점' and a dropdown menu for '현재 상영작'. Below this is a '전체 리스트' section with 13,761,953 reviews. Each review entry includes the review ID, title, rating (e.g., ★★★★★ 10), user name, and a brief summary. To the right, there's a sidebar for '영화 인기검색어' (Popular movie search terms) with links to 유제이탈자, 장르만 로맨스, 이태년스, 연애 빠진 로맨스, and 드라마, along with their respective counts (1, 1, 0, 3, 1). A date range from 2021.11.26 to 2021.11.26 is shown. Other sections include '네이버 최고 평점' (Highest-rated reviews) for 코다 (9.24), and a '가장 많이 추천된 리뷰' (Most recommended reviews) section.

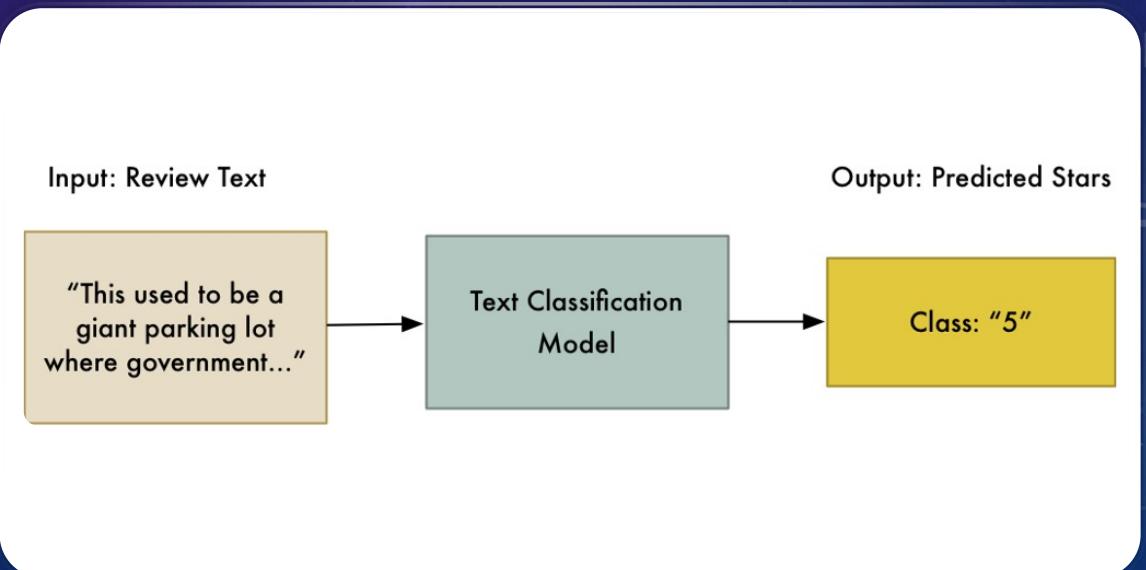
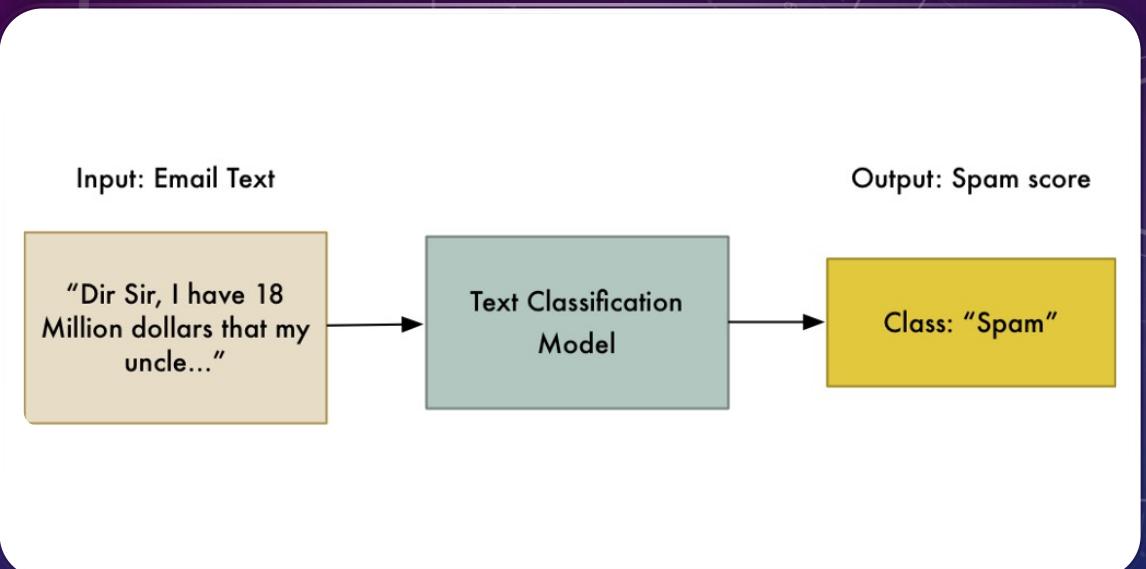
제목	평점	작성자	날짜
유제이탈자	★★★★★ 10	ik05****	21.11.27
장르만 로맨스	★★★★★ 8	alsq****	21.11.27
이태년스	★★★★★ 8	sson****	21.11.27
연애 빠진 로맨스	★★★★★ 10	mih****	21.11.27
드라마	★★★★★ 7	yisy****	21.11.27
한팅 인 살렘 : 악령의 미움	★★★★★ 1	dool****	21.11.27
서터	★★★★★ 8	thdw****	21.11.27
자산아보	★★★★★ 10	wow****	21.11.27
베놈 2: 헛 데어 비 카니지	★★★★★ 10	coco****	21.11.27

04. 컴퓨터언어학은 어떻게 이루어지나

BOTTOM UP, DATA DRIVEN, CLASSIFICATION

[HTTPS://MIRO.MEDIUM.COM/MAX/1000/1*x93sfurvt_bmomlzhgeglw.png](https://miro.medium.com/max/1000/1*x93sfurvt_bmomlzhgeglw.png)

[HTTPS://MIRO.MEDIUM.COM/MAX/1000/1*HV4SHRJHK4j9ZW27OB_R-G.PNG](https://miro.medium.com/max/1000/1*HV4SHRJHK4j9ZW27OB_R-G.png)



05. 컴퓨터언어학은 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. 컴퓨터언어학은 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
 - $P(\text{새빨간 거짓말}) > P(\text{새빨간 희망})$
 - $P(\text{이 강의는 참 재미있어요})$
- Assign a Probability to a sentence
 - N-gram based : Unigram, Bigram, Trigram...
 - Word Embedding
 - Transformer based Language Modeling

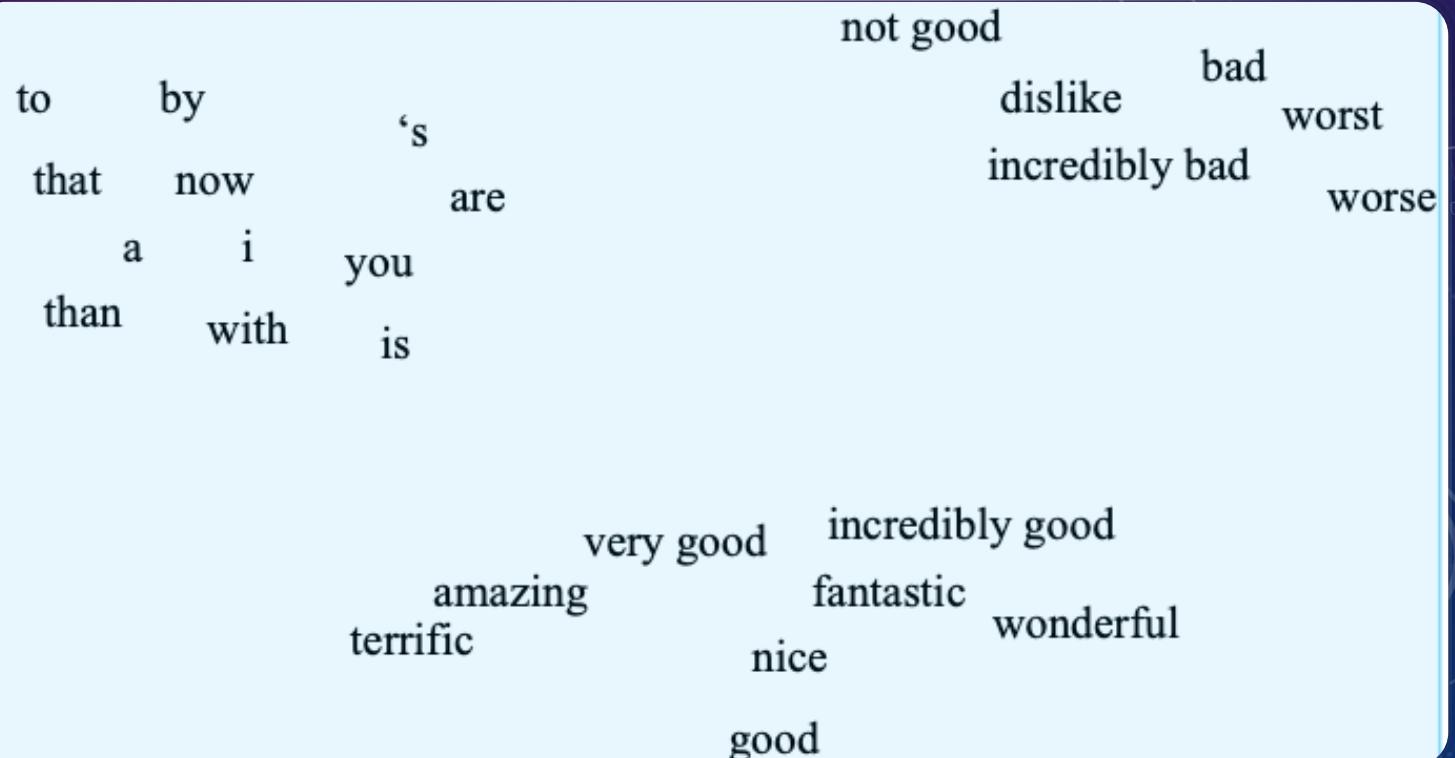
06. 언어모델(LANGUAGE MODEL)

- 문장의 확률을 어떻게?
 - 조건부 확률: $P(B|A) = P(A, B)/P(A) \rightarrow P(A, B) = P(A)p(B|A)$
 - $P(A, B, C, D) = P(A)P(B|A)P(C|A, B)P(D|A, B, C)$
 - Chain Rule:
 - $P(X_1, X_2, X_3, \dots, X_n) = P(X_1)P(X_2 | X_1)P(X_3 | X_1, X_2) \dots P(X_n | X_1, \dots, X_{n-1})$
- $$P(w_1 w_2 \dots w_n) = \prod_i P(w_i | w_1 w_2 \dots w_{i-1})$$
- $P(\text{"이 강의는 참 재미 있어요"}) = P(0|) \times P(\text{강의는}|0) \times P(\text{참}|\text{이 강의는}) \times P(\text{재미} | \text{이 강의는 참}) \times P(\text{있어요} | \text{이 강의는 참 재미})$
 - $P(\text{있어요} | \text{이 강의는 참 재미}) = \text{Count}(\text{이 강의는 참 재미 있어요}) / \text{Count}(\text{이 강의는 참 재미})$
- Markov Assumption
 - $P(\text{있어요} | \text{이 강의는 참 재미}) \approx P(\text{있어요} | \text{재미})$ 또는 $P(\text{있어요} | \text{참 재미})$
 - $P(\text{"이 강의는 참 재미 있어요"}) = P(0|) \times P(\text{강의는}|0) \times P(\text{참}|\text{강의는}) \times P(\text{재미} | \text{참}) \times P(\text{있어요} | \text{재미})$

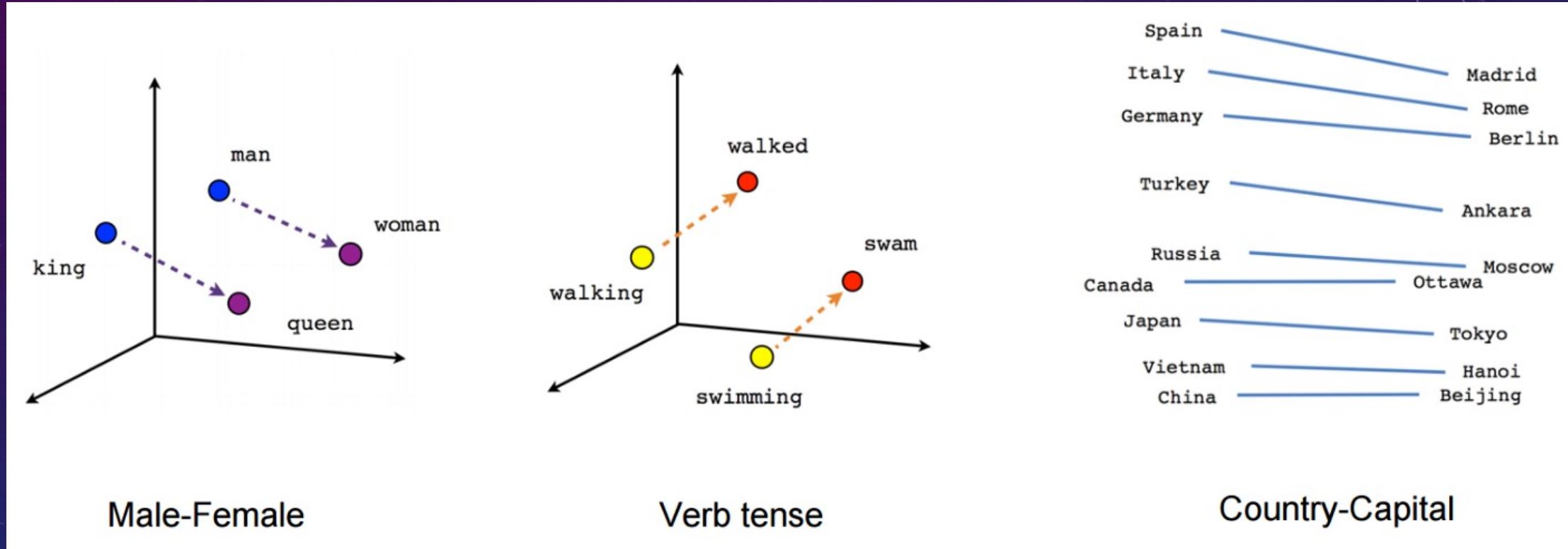
$$P(w_1 w_2 \dots w_n) \approx \prod_i P(w_i | w_{i-k} \dots w_{i-1})$$

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?
 - 동의어, 유사어, 다의어, 관련어...
- Word Embedding
 - Distributional Hypothesis
 - Word2Vec, Glove, FastText..
 - Static Word Embedding (다의어, 동의어 구별이 되지 않음)

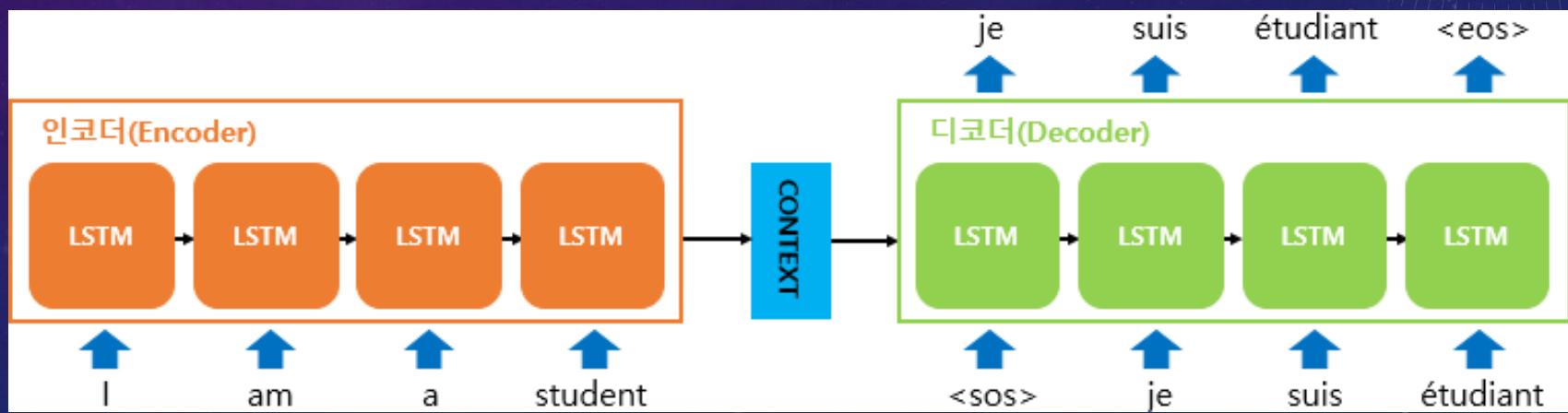


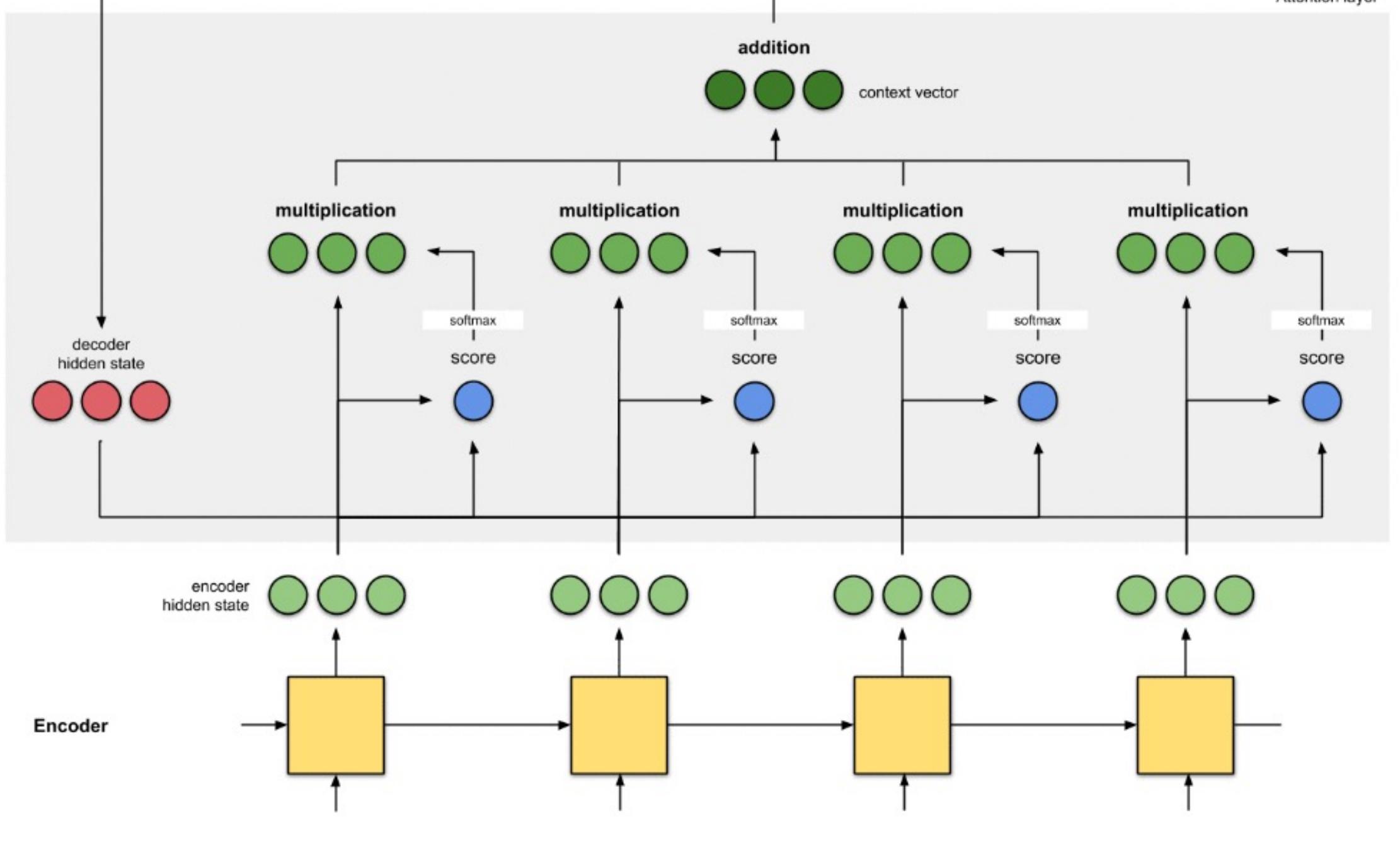
07. 단어 임베딩



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

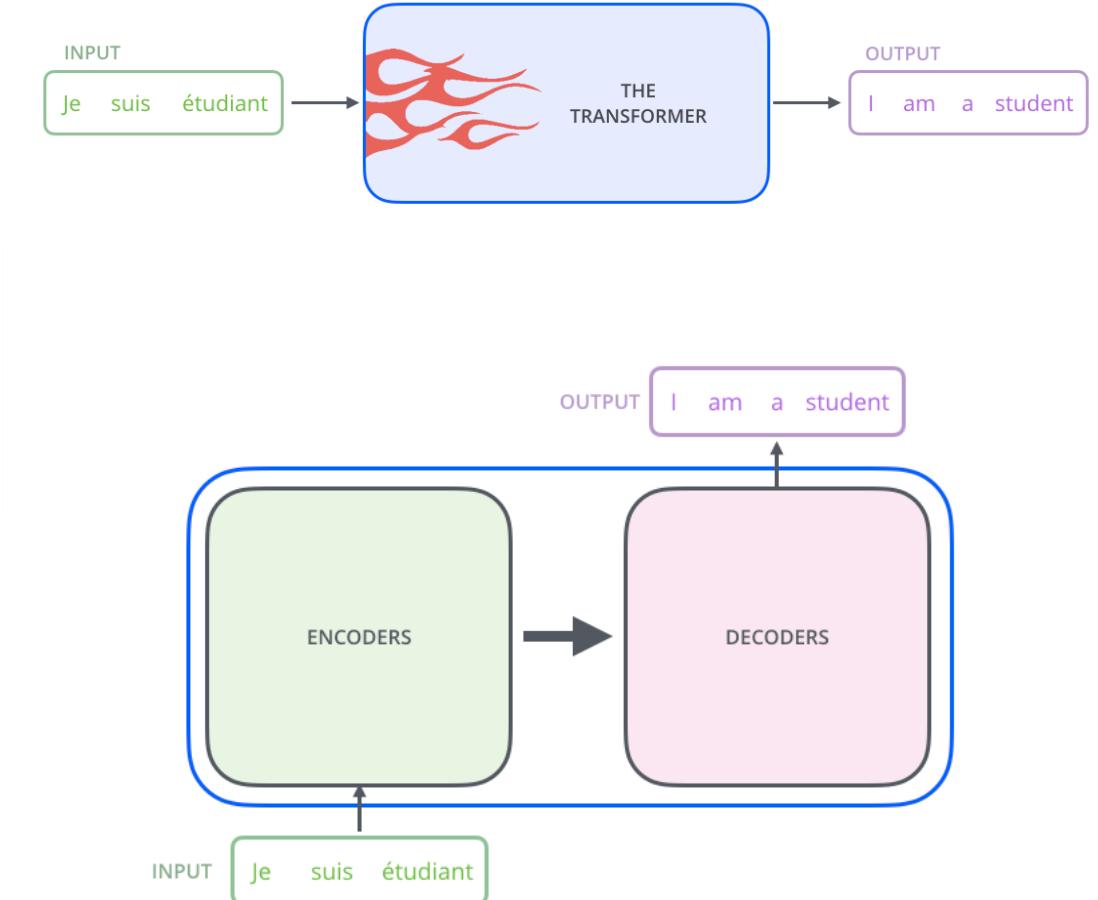
- Sequence to sequence Model
- Encoder-Decoder





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

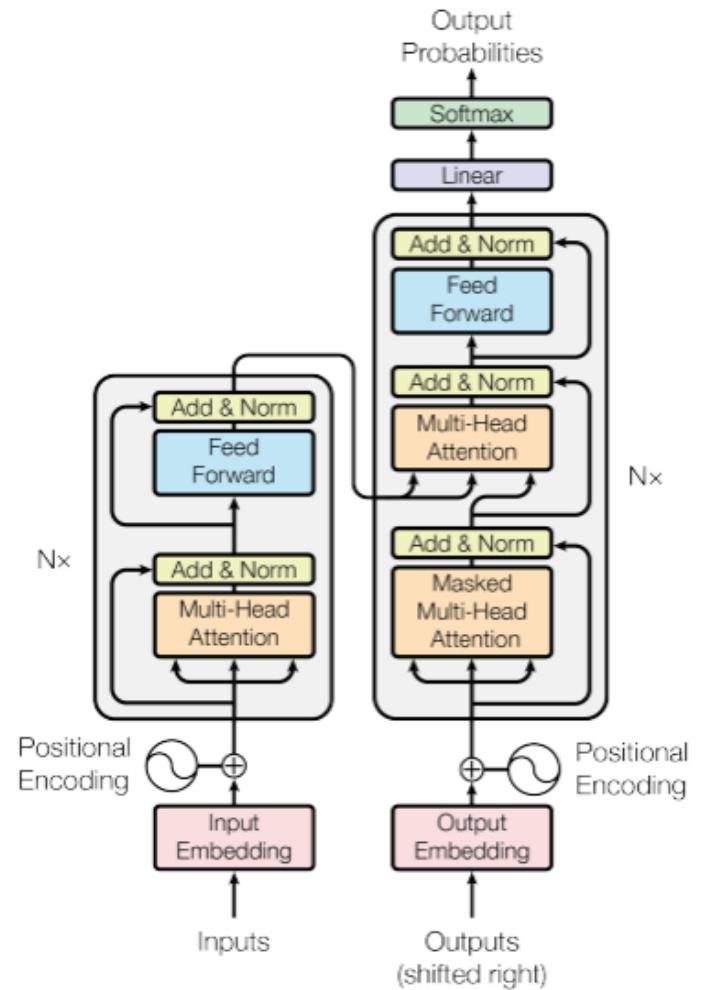
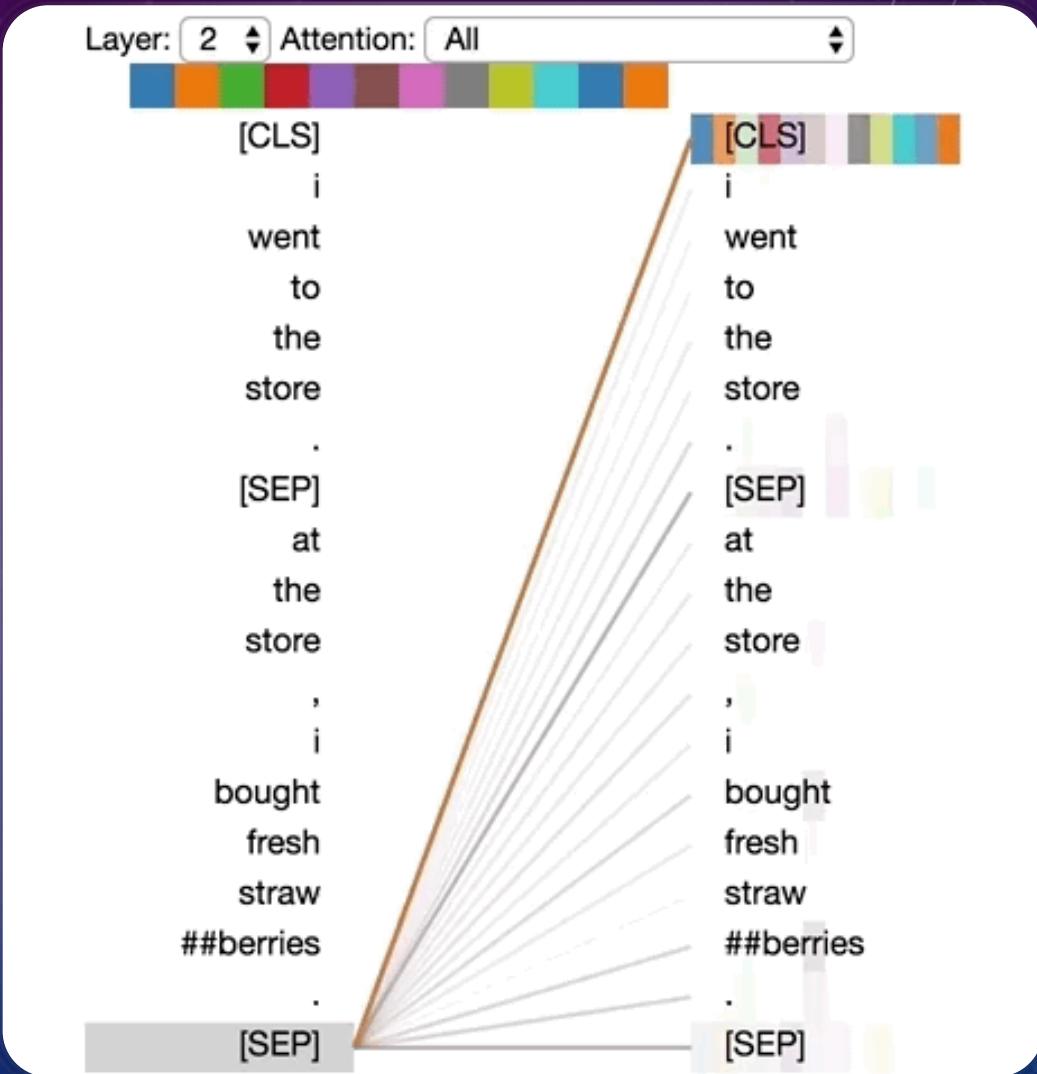


Figure 1: The Transformer - model architecture.

09. 트랜스포머

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

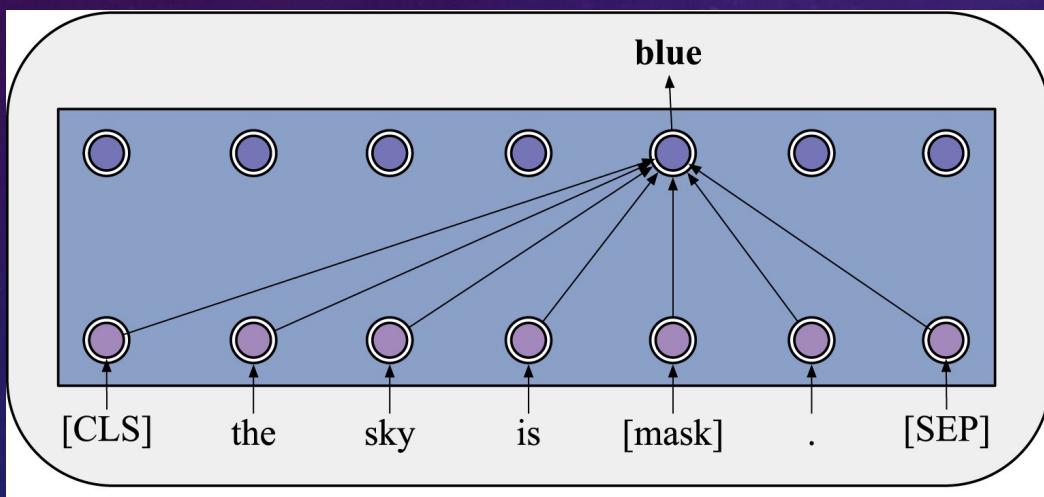


10. 사전학습모델

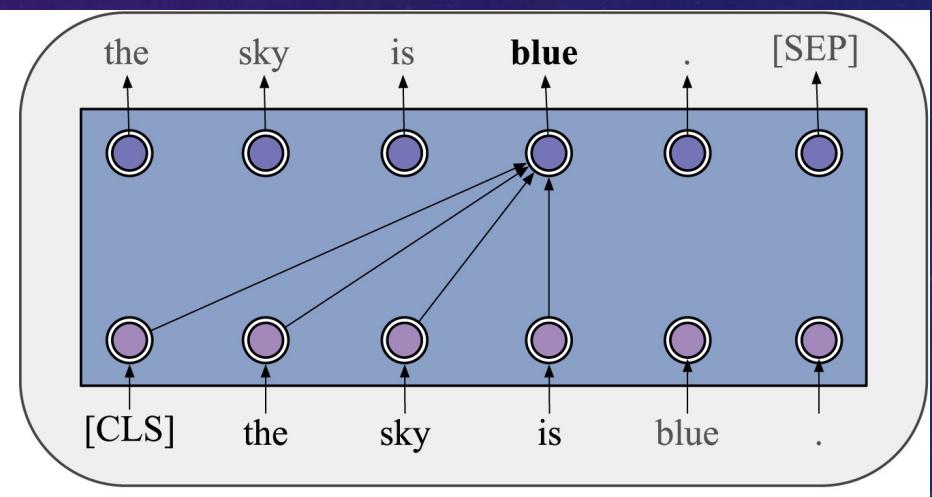
- Decoders or autoregressive models:
 - GPT, GPT-2, GPT3, CTRL(Conditional Transformer Language Model for Controllable Generation), Transformer-XL, Reformer, XLNet
- Encoders or autoencoding models
 - BERT, ALBERT, RoBERTa, DistilBERT, ConvBERT, XLM, XLM-RoBERTa, FlauBERT, ELECTRA, Funnel Transformer, 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



BERT



GPT

10. 사전학습모델

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

