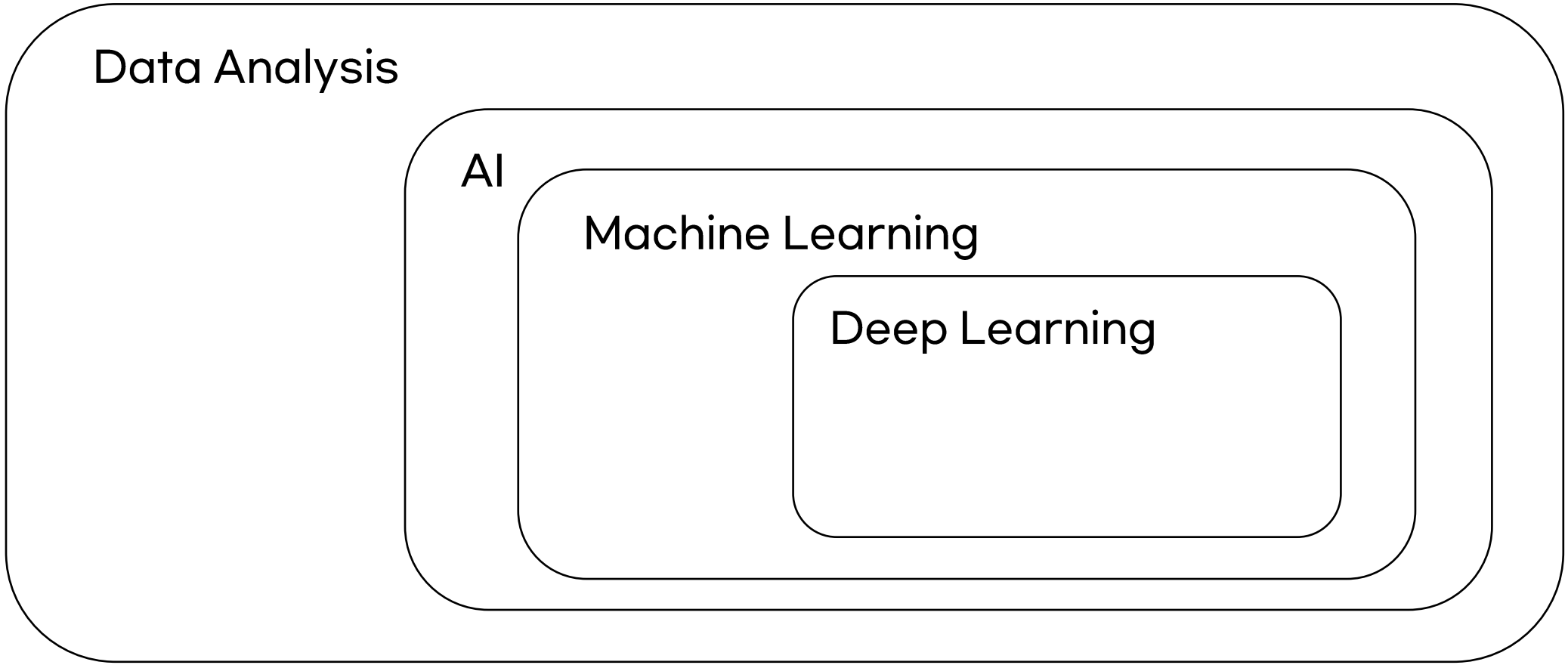
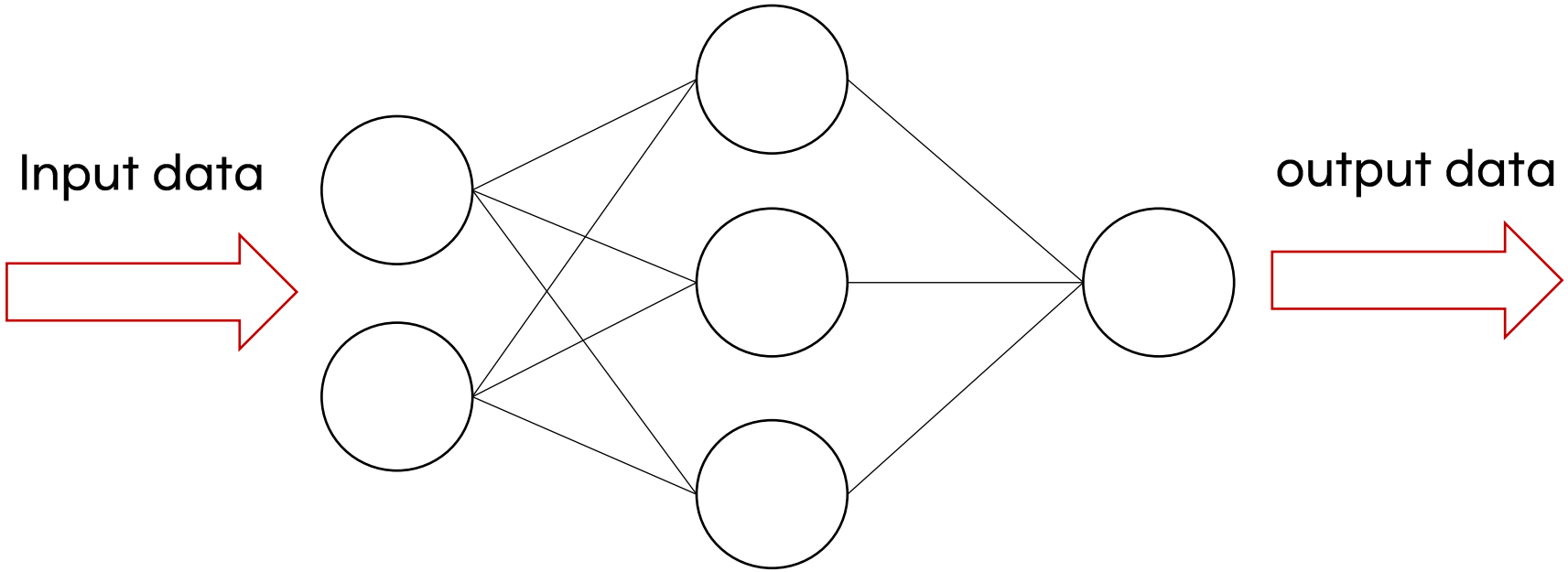
What is NLP?

KHUDA 이상민

데이터분석과 AI

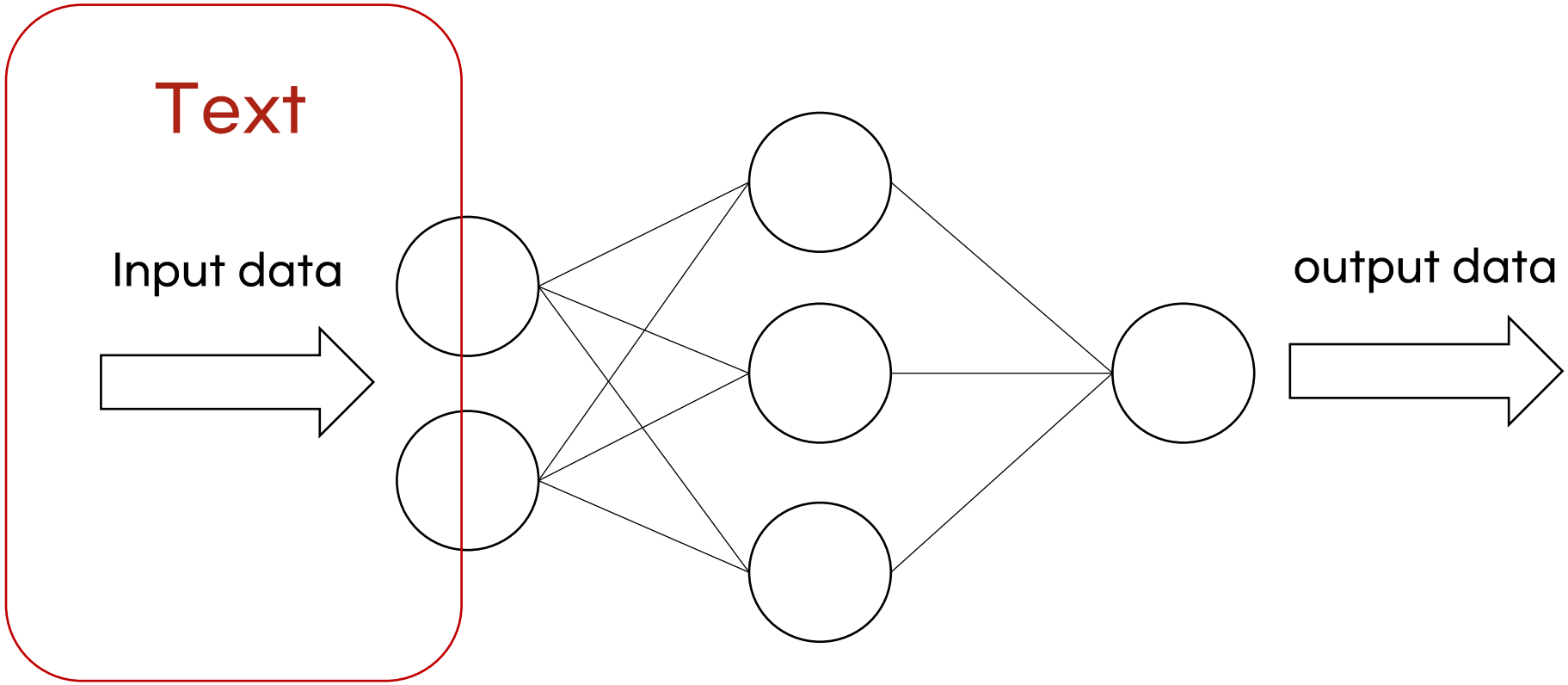


Deep Learning?



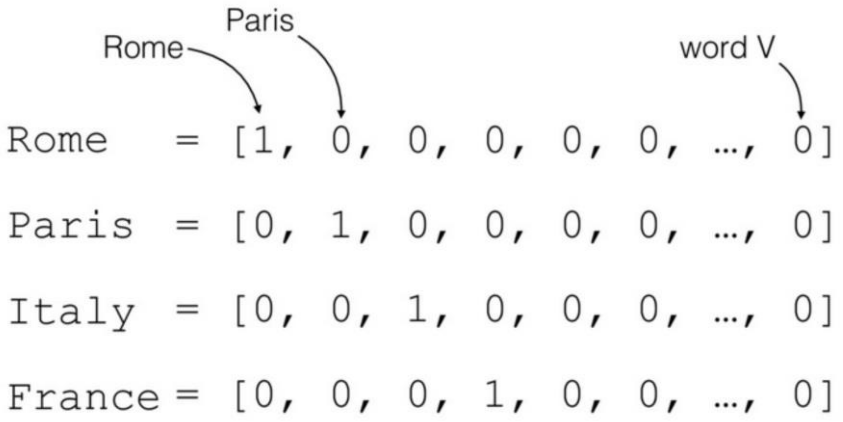
인간의 뉴런과 비슷한 인공신경망 방식으로 정보를 처리

NLP in Deep Learning

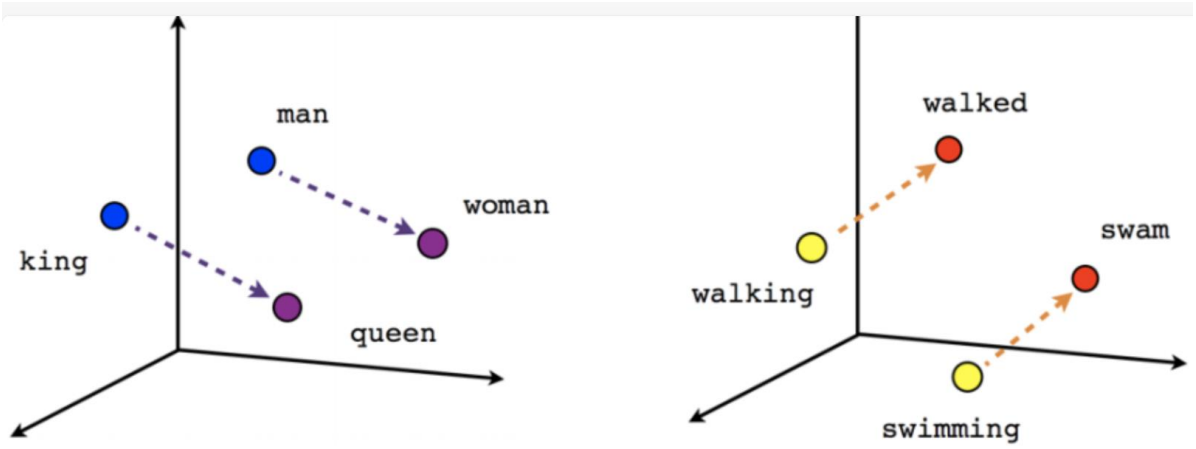


쉽게 말해 기존 딥러닝 구조에서 입력값이 텍스트로 바뀐 것

Word Embeddings



One hot Vector Representation

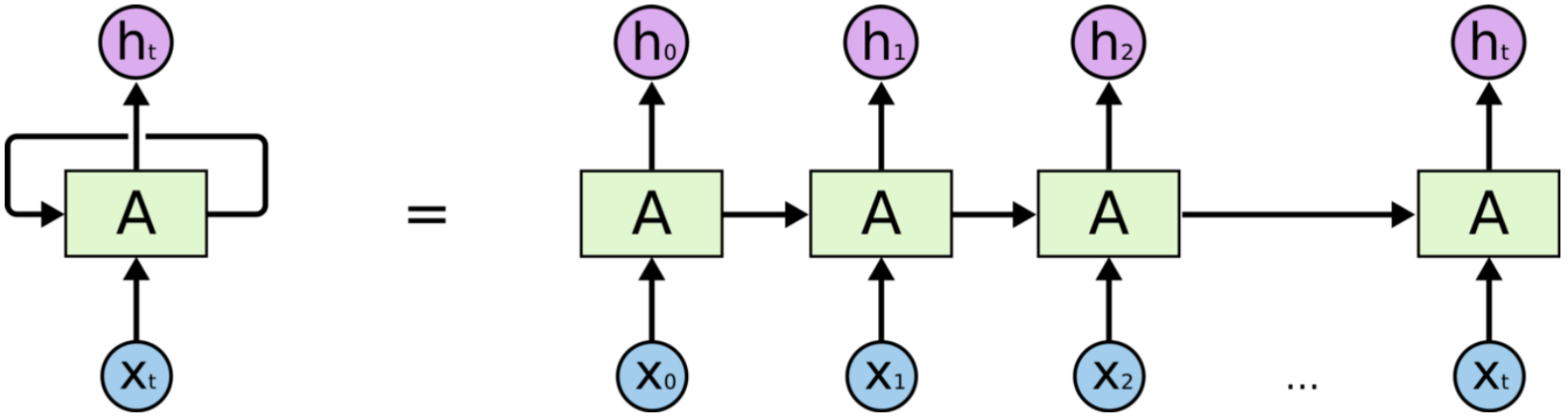


word2vec

텍스트를 의미 있는 벡터로 변환하는 것

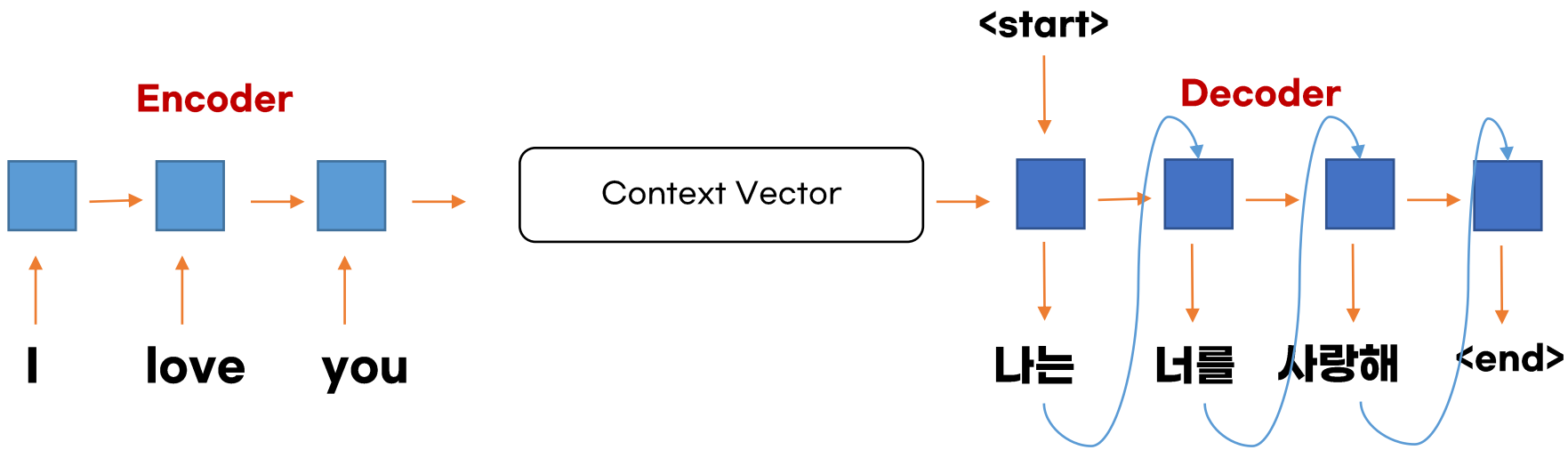
Efficient Estimation of Word Representations in Vector Space (ICLR, 2013)

RNN



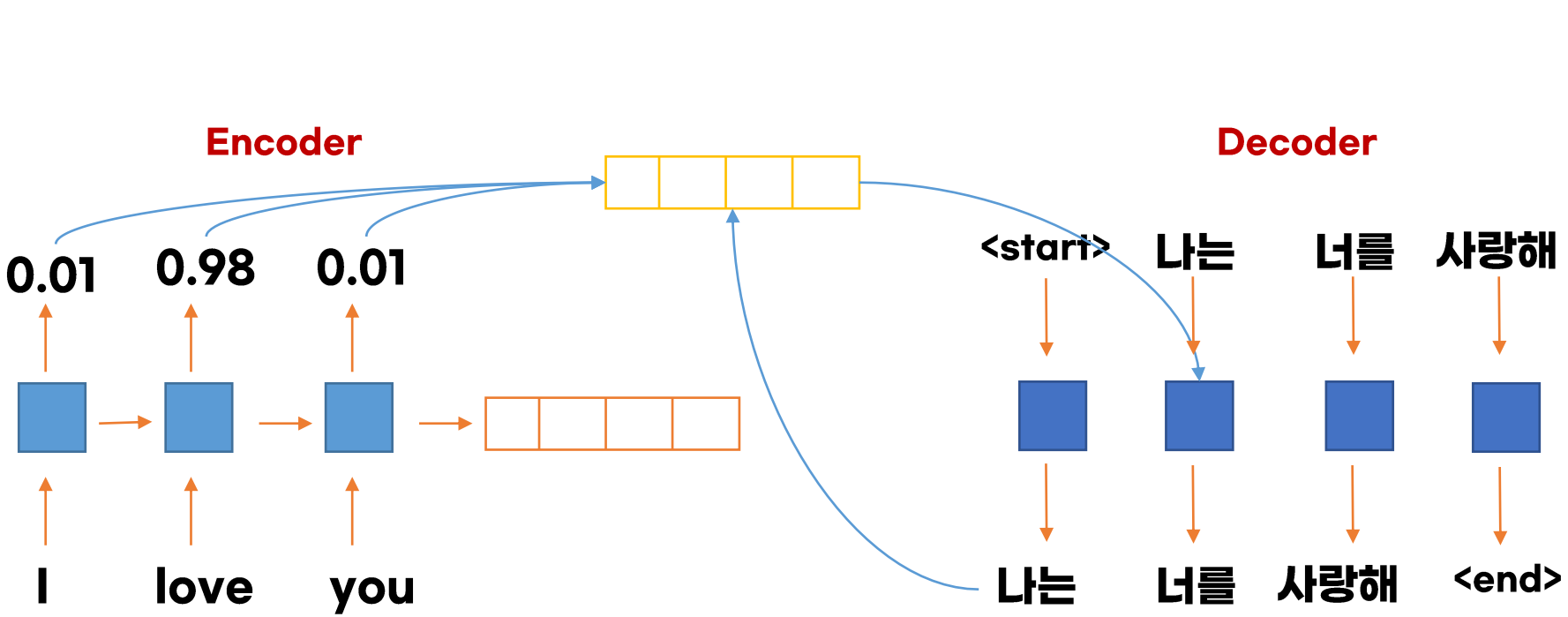
이전 단어들을 이해해야 텍스트 전체의 맥락을 이해할 수 있다.
-> 텍스트를 Sequence data로 보기 시작

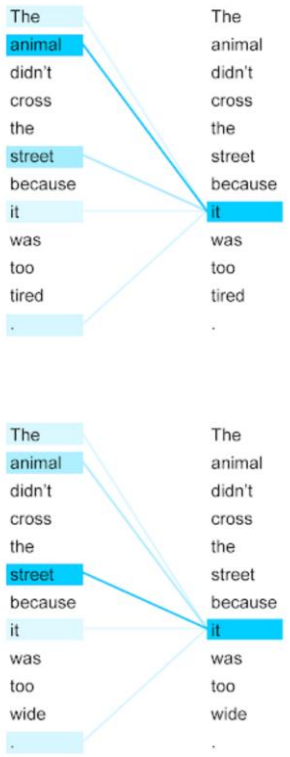
RNN based encoder decoder



순차적으로 학습하면서 하나의 벡터에 모든 정보를 저장해
병목 현상이 일어나 속도가 느리고 성능이 저하됨

Attention





디코더에서 출력 단어를 예측하는 매 시점(time step)마다
인코더에서의 전체 입력 문장을 다시 참고

단, 해당 시점에서 예측해야 할 단어와 연관이 있는 단어 부분을 좀 더 집중(Attention)해서 보는 구조

Neural Machine Translation by Jointly Learning to Align and Translate (ICLR, 2015)

Transformer

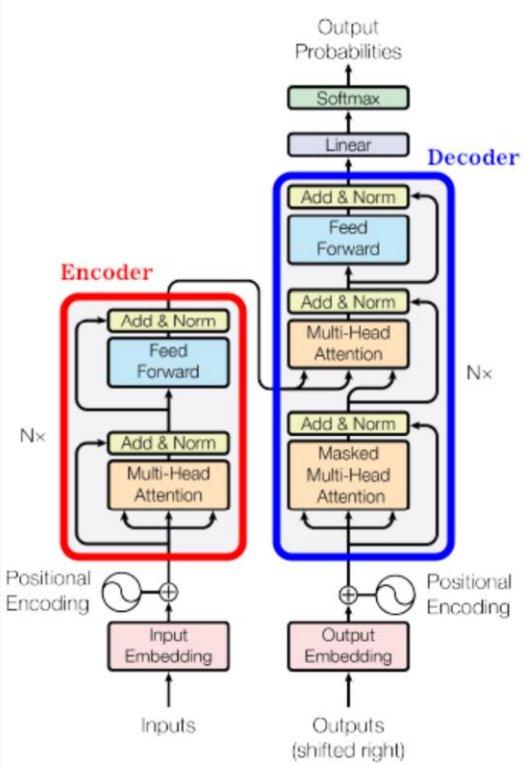
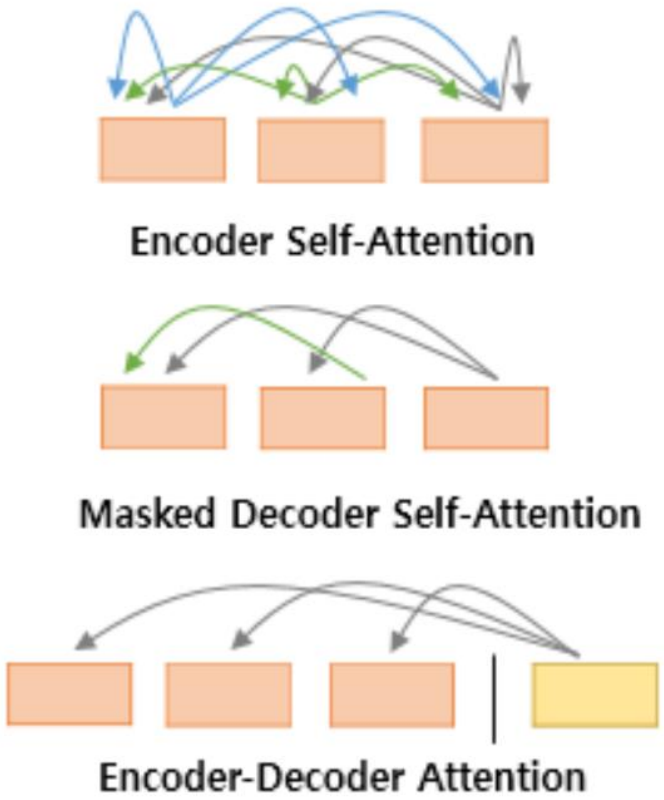


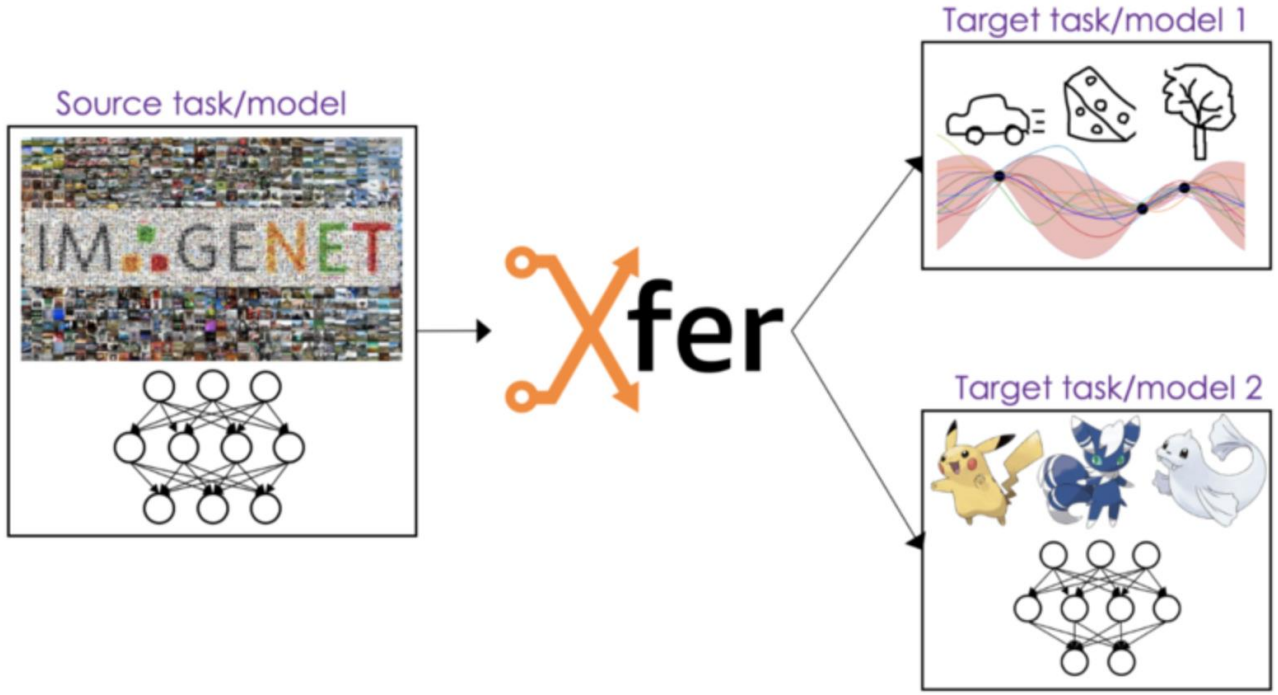
Figure 1: The Transformer - model architecture.



Attention만을 사용하여 Encoder-Decoder 구조를 구현한 모델

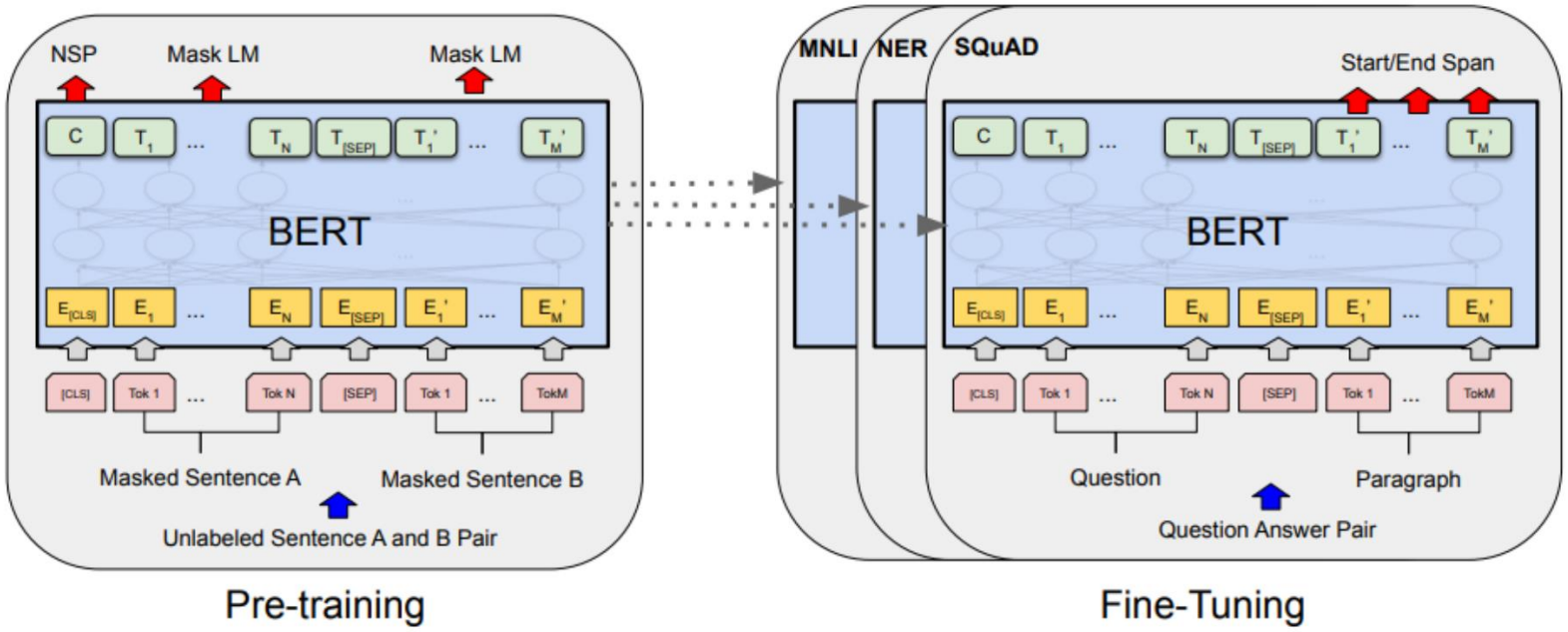
Attention is All You Need (NIPS, 2017)

Transfer Learning in NLP



매우 큰 데이터셋에 훈련된 모델의 가중치를 가지고 와서
우리가 해결하고자 하는 과제에 맞게 재보정해서 사용

BERT



Transfer Learning의 개념을 착용한 Pre-trained 모델
Transformer의 Encoder를 적층시켜 만든 구조로, NSP과 MLM기법 적용

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (NAACL-HLT, 2019)

Technology

Electra

GPT-1

XLNet

KoBERT

T5

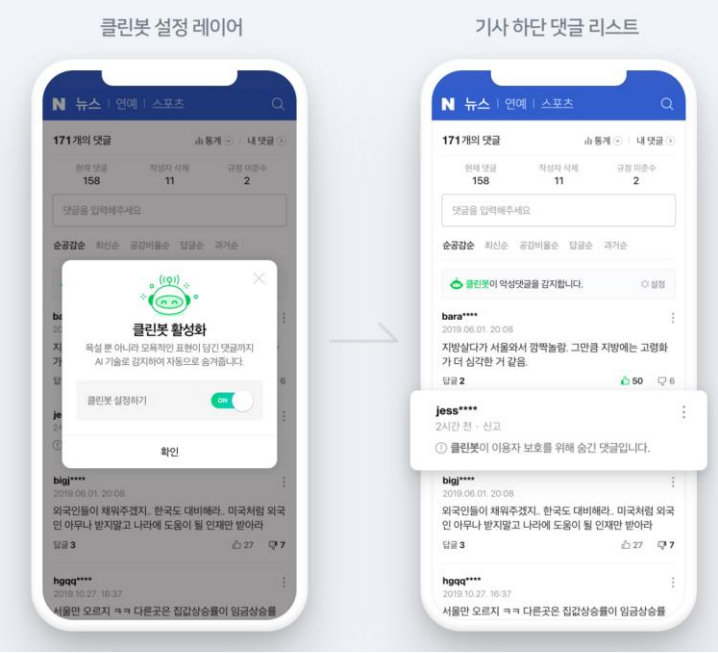
RoBERTa

BART

GPT-3

자연어처리로 어떤 것들을 할 수 있는가?

Text classification



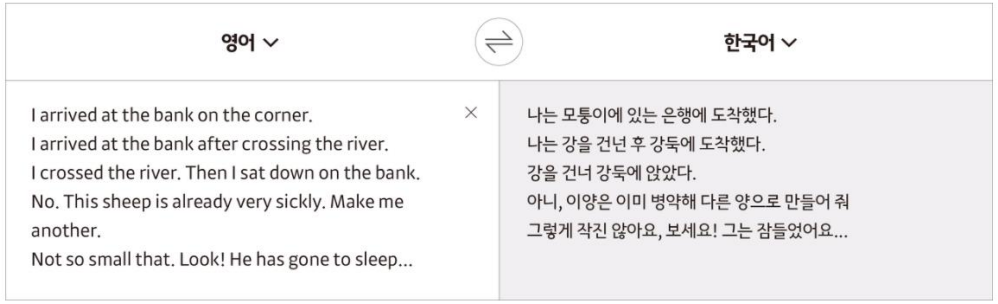
네이버 클린봇 - 욕설분류

텍스트(메일의 내용)	레이블(스팸 여부)
당신에게 드리는 마지막 혜택! ...	스팸 메일
내일 뵐 수 있을지 확인 부탁...	정상 메일
쉿! 혼자 보세요...	스팸 메일
언제까지 답장 가능할...	정상 메일
...	...
(광고) 멋있어질 수 있는...	스팸 메일

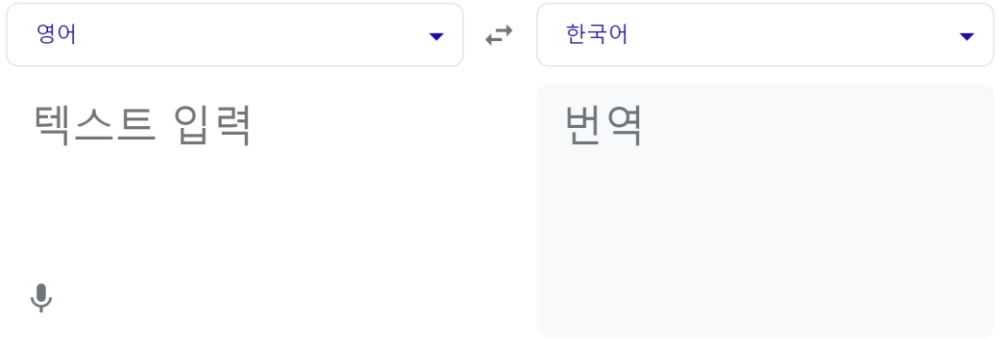
메일 스팸분류

텍스트를 입력으로 받아 범주(label)별로 분류

Machine Translation



카카오 i 번역기



구글 번역기

한 언어에서 다른 언어로 글자나 음성을 변환

Back-translation기법을 통해 데이터 증강에도 사용

Text summarization

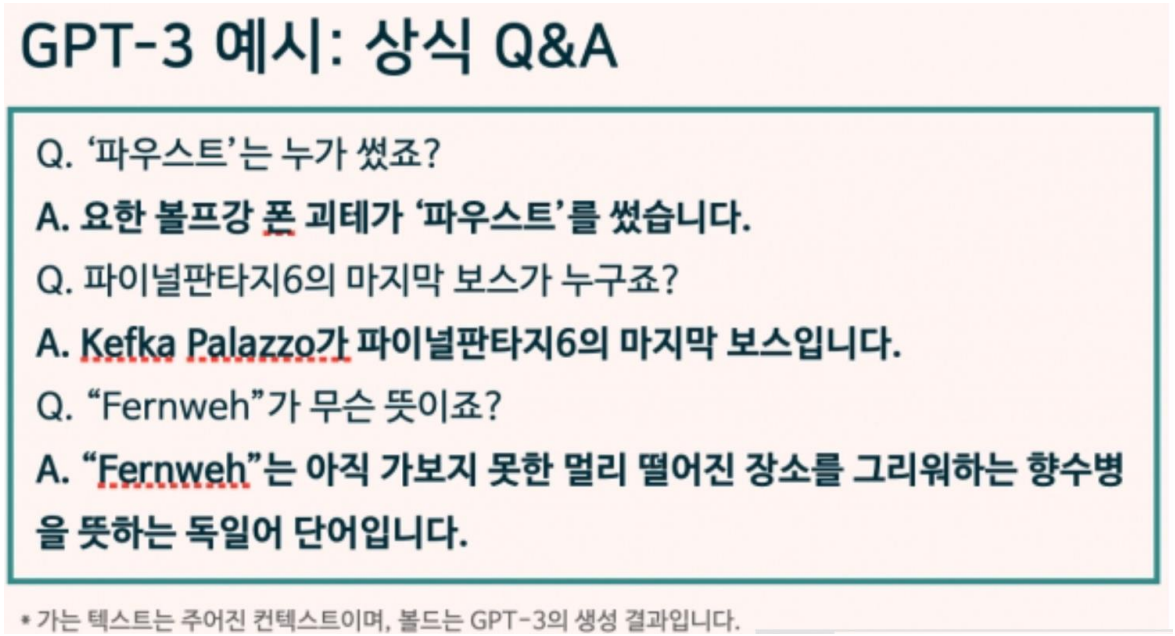
네이버 뉴스 요약봇

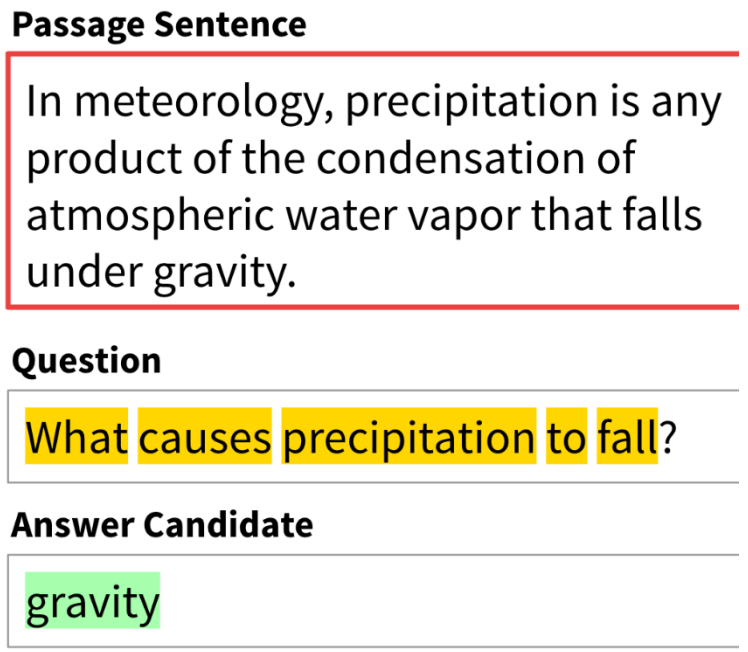
네이버 쇼핑 리뷰 요약

매우 긴 원문을 핵심 내용만 간추려서 작은 문장으로 요약하는 것

크게 추출 요약과 생성 요약으로 나누어짐

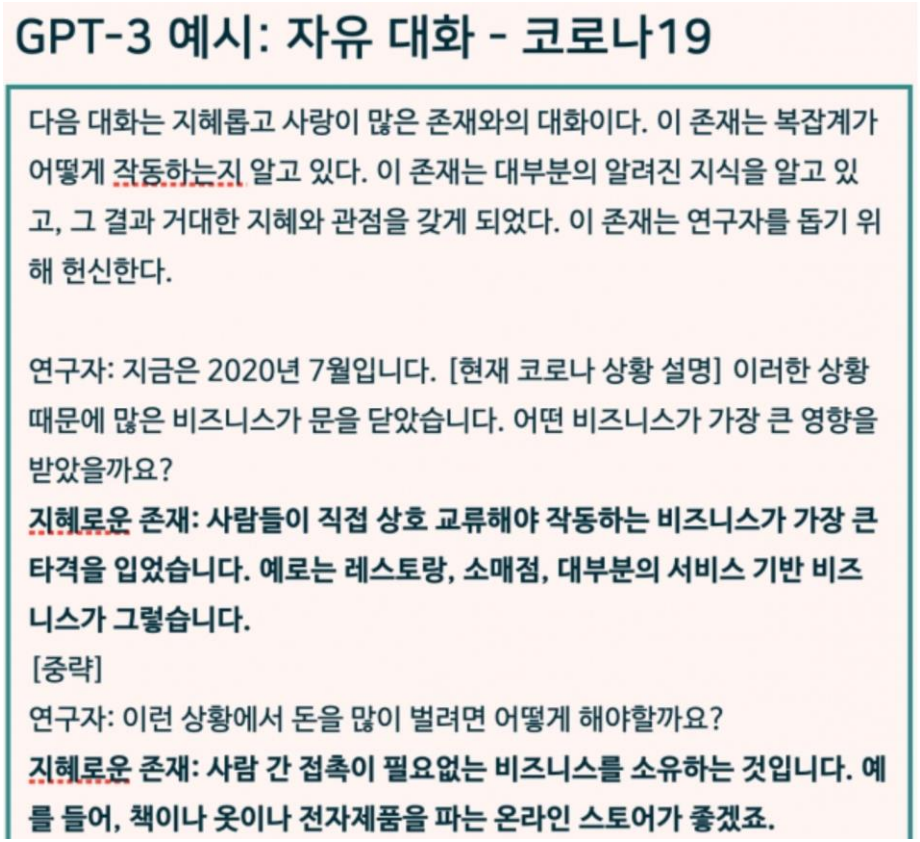
Question Answering





인간이 제기한 질문에 자동으로 응답하는 시스템

Text Generation



GPT-3 예시: 자유 대화 - 코로나19

다음 대화는 지혜롭고 사랑이 많은 존재와의 대화이다. 이 존재는 복잡계가 어떻게 작동하는지 알고 있다. 이 존재는 대부분의 알려진 지식을 알고 있고, 그 결과 거대한 지혜와 관점을 갖게 되었다. 이 존재는 연구자를 돕기 위해 헌신한다.

연구자: 지금은 2020년 7월입니다. [현재 코로나 상황 설명] 이러한 상황 때문에 많은 비즈니스가 문을 닫았습니다. 어떤 비즈니스가 가장 큰 영향을 받았을까요?

지혜로운 존재: 사람들이 직접 상호 교류해야 작동하는 비즈니스가 가장 큰 타격을 입었습니다. 예로는 레스토랑, 소매점, 대부분의 서비스 기반 비즈니스가 그렇습니다.

[중략]

연구자: 이런 상황에서 돈을 많이 벌려면 어떻게 해야할까요?

지혜로운 존재: 사람 간 접촉이 필요없는 비즈니스를 소유하는 것입니다. 예를 들어, 책이나 옷이나 전자제품을 파는 온라인 스토어가 좋겠죠.

텍스트를 생성하여 자동으로 글을 작성하거나
대화할 수 있도록 하는 것

Nlp + ?

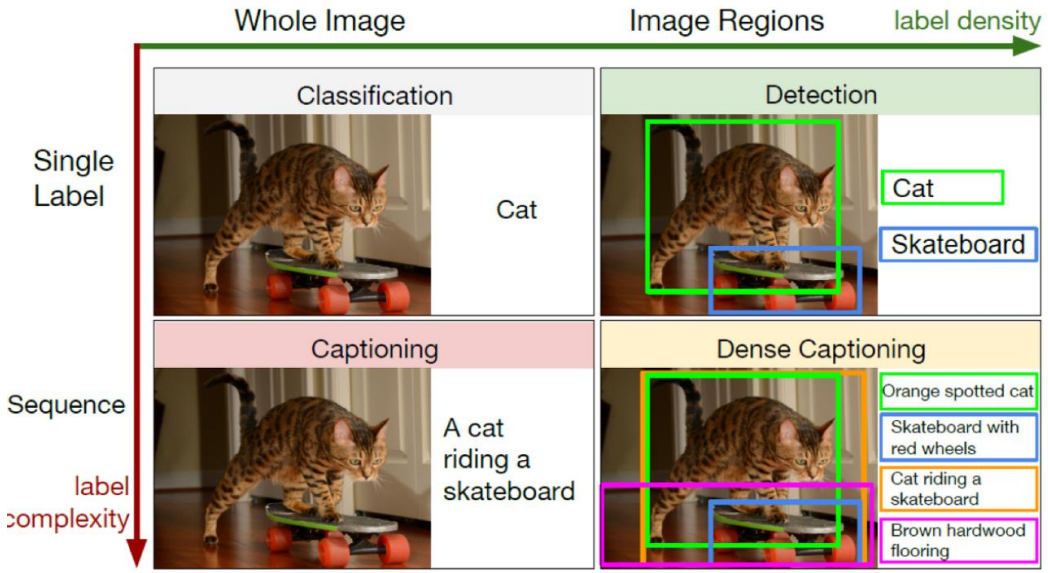
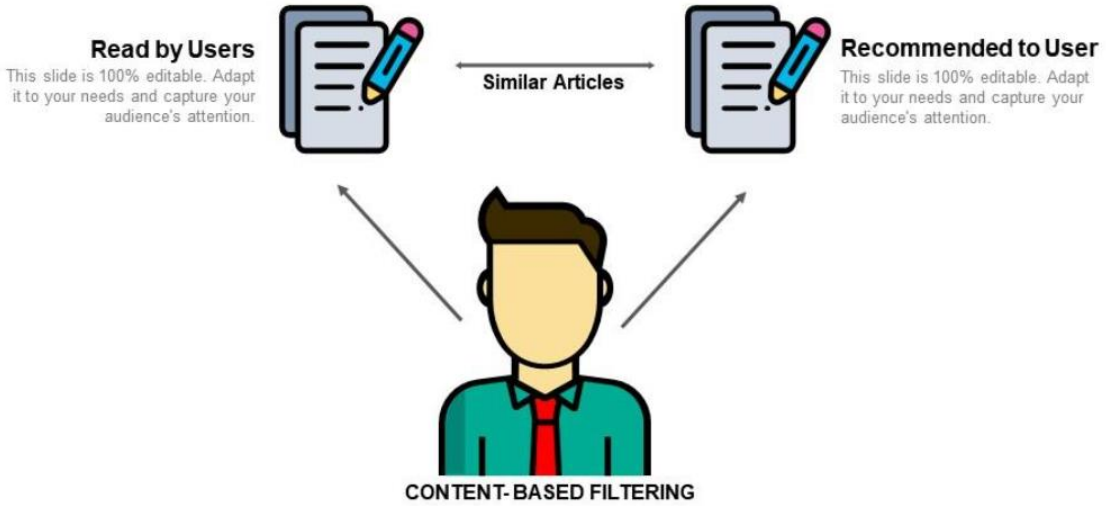


Image captioning

이미지를 설명하는
캡션(설명)을 만들어 내는 것



Content(text)-based filtering

문서 간의 유사도를 구해
추천해주는 시스템

Task

토픽 모델링

문서쌍 유사도 검사

자연어 추론

기계독해

언어적 용인 가능성

감정분석

의도 분류

자연어처리의 문제점은?

자연어처리의 문제점

예시) 개꿀잼이다

substring count

- count(개) = 20000
- count(개꿀) = 1500
- count(개꿀잼) = 1200
- count(개꿀잼이) = 30
- count(개꿀잼이다) = 15

복잡하고 번거로운 전처리

- **알잘딱깔센** : 알아서 잘 딱 깔끔하고 센스있게
- **억텐** : 억지 텐션
- **당모치** : 당연히 모든 치킨은 옳다
- **자강두천** : 자존심 강한 두 천재의 대결

꾸준히 생겨나는 신조어

crawl	CC-MAIN-2021-21	CC-MAIN-2021-25	CC-MAIN-2021-31
language	%	%	%
eng	44.8424	45.1857	45.5605
rus	7.2658	6.7665	7.2898
deu	5.6906	5.5856	5.9029
zho	4.6508	5.0691	4.1251
jpn	4.5850	4.3859	4.9273
fra	4.4747	4.3725	4.5090
spa	4.3315	4.3107	4.4105
<unknown>	2.8840	2.3539	1.7425
ita	2.4277	2.4112	2.4354
por	2.1373	2.1200	2.2068
nld	1.8082	1.7758	1.9176
pol	1.6050	1.5885	1.6471
ces	1.1124	1.0004	1.1233
tur	1.0099	1.0407	1.0210
ind	0.8311	0.8450	0.8500
vie	0.8036	1.5225	0.9478
swe	0.7398	0.7236	0.7292
fas	0.6425	0.6692	0.6570
kor	0.6378	0.6697	0.5698

[표 1] Common Crawl 웹 데이터

압도적으로 부족한 한국어 데이터

자연어처리가 중요한 이유
-> 한글의 특수성

자연어처리 분야 채용

주요업무
- 딥러닝/머신러닝 분야의 최신 알고리즘 연구
- text preprocessing, sentiment classification, emotion recognition
- text generation, text style transfer, conversation management, multimodal Q&A

자격요건
아래의 요건중 하나에 해당하면 됩니다.

- 최신 딥러닝 논문을 코드로 구현하고 실험, 분석 할 수 있는 능력있음
- 최신 딥러닝/머신러닝 동향에 대해서 이해하고 있음

또는

- 전통적인 언어처리 알고리즘에 대한 실무적인 경험 있음
- 언어처리 분야의 새로운 알고리즘을 제안하고 구현해본 경험이 있음

우대사항
- 오픈소스 활동
- 관련 분야 논문 개제 경험
- 챗봇등 언어처리관련 개발 경험

업무내용
- 사용자의 의도를 이해하는 NLU(Natural Language Understanding) 개발
- 기계 독해(Machine Reading Comprehension)
- smalltalk(챗봇) 개발
- 검색에 사용되는 여러 가지 언어 처리 기술(Natural Language Processing)

지원자격
- Tensorflow, keras, pytorch 등 프레임 워크가 능숙하신 분
- 자연어 처리 기술 주제에 경험과 관심을 갖고 계신 분

우대사항
- 자연어 처리 전공자 / Machine Learning 전공자
- 주요 학회(ACL, EMNLP, COLING 등)에 논문을 게재한 분
- Deep Learning 프로그램 숙련자
- C/C++ python 프로그램 숙련자
- 개발 경력 3년 이상

자연어처리 분야 채용

[주요 업무]

- 전세계 다양한 언어를 활용한 language model (Transformer, GPT, LSTM 등) 연구 및 서비스 개발
- Education 도메인에 특화된 인공지능 시스템 연구 및 개발
- 다양한 ML/AI 알고리즘 및 모델링을 통한 관다 서비스 향상

[자격 요건]

- 인공지능 분야 석사 이상의 학위 혹은 그에 준하는 경력이 있으신 분
- 머신 러닝 알고리즘에 대한 기본기를 갖추신 분
- 오픈소스를 활용한 모델의 학습 및 개선 경험이 있으신 분

[우대 사항]

- 국내외 유수 학회/저널에 논문을 게재한 경험이 있는 분
- 인공지능 기술을 활용한 프로젝트를 주도적으로 진행한 경험이 있는 분
- AWS/GCP 등을 이용한 머신러닝 서비스 개발 및 배포 경험이 있는 분
- 최소 하나 이상의 ML 라이브러리(PyTorch, TensorFlow 등)를 익숙하게 활용하실 수 있으신 분

주요업무

1. [자연어처리 부문 인공지능 서비스 연구개발]
 - 질의응답, 기계독해, 대화모델, NLU, NLG 등의Conversational AI 기술
 - 형태소분석, 개체명 인식, 구문분석, 언어모델 등의 NLP기반기술
 - 정보추출, 평판분석, 문서요약, 문서분류, 토픽 모델링 등의 텍스트마이닝 기술
 - 자연어처리 분야 선행 연구

2. [정보검색 부문 인공지능 서비스 연구개발]
 - 정보검색 모델링 및 검색엔진 설계
 - 텍스트 유사도 및 순위화 알고리즘
 - 키워드 추출, 정제, 유사도, 순위화 알고리즘
 - 데이터 수집, 정제 및 구조화
 - Neural IR 등 정보검색 선행연구

3. [지능형 지식 지원시스템 연구개발]
 - 금융데이터 기반 지능형 지식 지원 시스템 연구개발
 - 온톨로지(Ontology) 기반 지식 모델링 및 지식 베이스 구축
 - 상담 지원 시스템, Q&A 시스템, 챗봇 관련 알고리즘 연구개발
 - 텍스트 기반 기계 학습 및 딥러닝 모델 연구개발

자격요건

- 석사 학위 이상 (필수)
- 원활한 커뮤니케이션 스킬 보유자
- 컴퓨터공학/통계학/수학/산업공학 관련 전공
- 자연어처리/텍스트마이닝/머신러닝/딥러닝 모델 연구개발 경력자
- 프로그래밍(Python, C/C++, Java 등) 가능자
- PM/PL 등 과제 및 조직관리 경험자 우대

우대사항

- NLP/IR/KE 관련 딥러닝 지식 및 활용 경험
- NLP/IR/KE 관련 솔루션 설계 및 개발 경험
- AI관련 최신 논문의 작성, 이해 및 활용 경험

자연어처리 분야 채용

Job Description

- 고객 경험관리, 제품전략 지원을 위한 분석과제 수행 및 인사이트 발굴
- 예측 / 최적화 / 추천 모델 설계 및 개발
- AI 기반 지능형 자동화 과제 기획 및 수행
- 빅데이터 분석 활용 / 확산 지원 (교육 및 프로젝트 지원)

Job Requirements

- 프로그램 언어 활용 능력 (R, Python, SQL 등)
- 데이터 기반 업무 분석 및 이해 역량
- 분석 자료 프리젠테이션 문서 작성 역량
- 관련 경력 3~12년

우대사항
- 통계학, 산업공학, 컴퓨터 공학 관련 전공자
- 데이터마이닝, 최적화 등 수리 알고리즘 이해 및 개발 역량
- 텍스트마이닝 개발 및 활용 경험
- 클라우드 이해 및 사용 역량
- 인공지능(AI/ML) 이해 및 활용 역량
- 인프라 관리 및 기획 경험
- CPG 산업, 유통, 온라인 비즈니스 경험자
- 빅데이터 시스템 설계 및 분석 데이터 마트 구축
- 분석 솔루션 구축 경험

업무 소개

- 구매데이터와 동물병원 영수증 진료내역,구강검사키트 검진 데이터와 사용자가 등록한반려동물의 정보, 이 밖에 도 함께 모아갈 여러 종류의 반려동물 데이터를 지금보다 더의미있게 활용하려고 합니다.
- 가장 적합한 사용자에게 핏펫몰 상품 노출과알림을 위한 추천 알고리즘을 개발
- 이미 세상에 나와있는 모델을 적용해보고,최적화되고 개선된 모델을 개발
- 사용자의 반려동물 헬스케어 정보를 적시에 제공해 도울 수 있도록 데이터를 활용하는 방안 계획, 연구

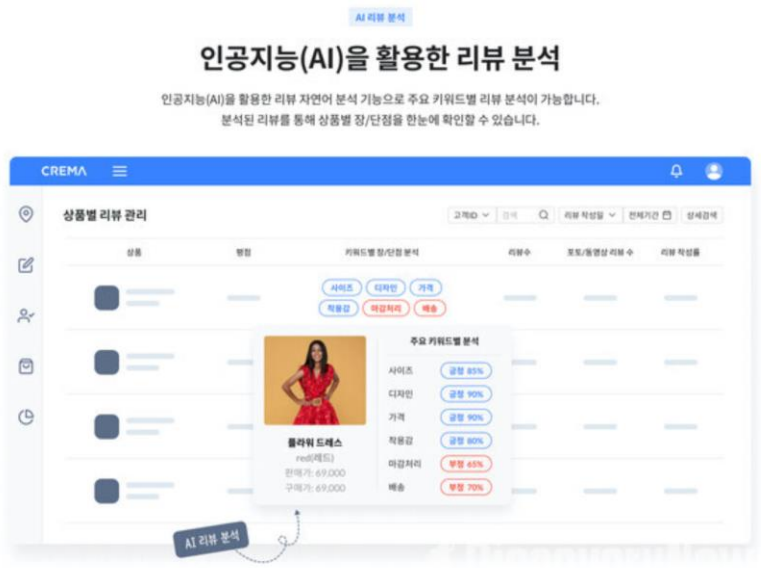
자격 조건

- 머신러닝, 딥러닝 모델의 주요 개념들을 이해하고 계신 분
- 머신러닝 논문들을 이해하고 구현하여 비즈니스 모델에적용할 수 있는 분
- 개인화 추천 알고리즘 개발 경험이 있으신 분
- 주요 머신러닝 프레임워크(TensorFlow, PyTorch, Keras, Scilkit)사용 경험이 있으신 분
- 한국어 자연어처리/텍스트마이닝 경험이 있으신 분
- 이미 만들어진 모델을 시스템 및 알고리즘 효율 관점에서 개선한 경험이 있으신 분
- 서비스와 비즈니스를 도메인 관점에서 분석하고 데이터와 연결하여 모델을 설계해본 경험이 있으신 분

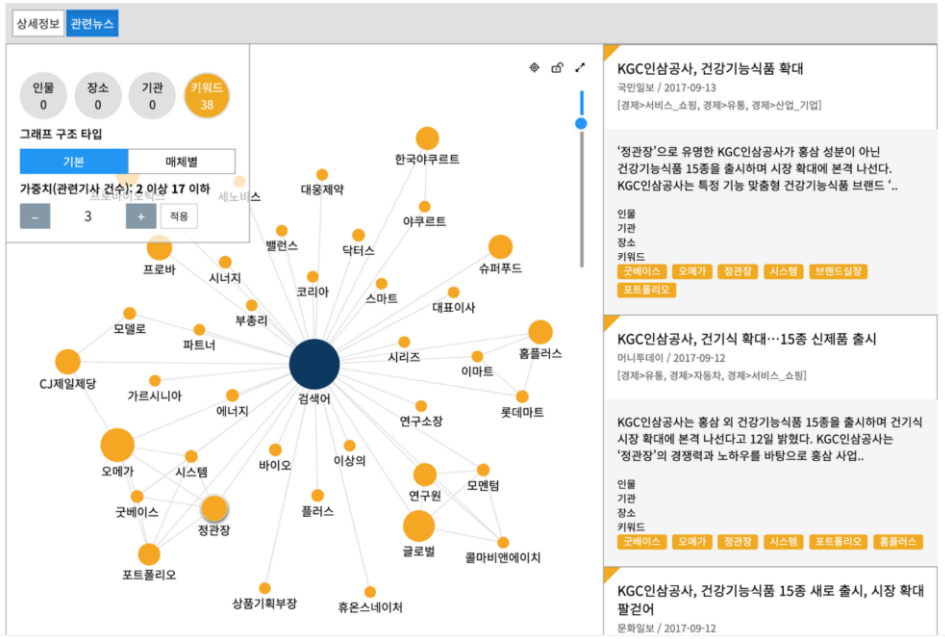
우대사항

- 머신러닝/추천시스템 관련 프로젝트를 경험하셨거나 추천 알고리즘에 대한 경험을 보유하신 분
- 데이터 인프라, 데이터 분석, 머신러닝을 포괄하는 팀 빌딩 및 리딩을 해보신 분
- 통계학/컴퓨터공학 석사 이상의 학위를 가진 분
- 관련 분야 학회에서 논문을 발표하신 분

AI엔지니어/리서처만??



이커머스의 핵심 상품 리뷰 분석



개체명 인식을 통한 뉴스 연결망 분석

-> 데이터 사이언티스트, 분석 분야에서도 텍스트 데이터를
활용해 분석해 본 경험을 매우 큰 이점

자연어처리의 알고리즘, 최신 기술을 깊게 파고 들고 싶다.
-> NLP(AI) Researcher

자연어처리를 활용하여 직접 서비스를 개발하고 개선하고 싶다.
-> AI 엔지니어

자연어처리뿐만 아니라 머신러닝, 딥러닝을 활용해 효율적인 모델링을 설계하고 싶다.
-> 데이터 사이언티스트

텍스트 데이터 전처리, 토픽 모델링 등 기본적인 텍스트 마이닝 기법을 도메인에 사용하고 싶다.
-> 도메인 전문 데이터 분석가

딥러닝 중 한 기술인
자연어처리를 다룬다고 생각하지 말고,
텍스트 데이터를 분석한다고 생각

최신 기술(ex-BERT)을 아무 생각 없이
코드 복붙해서 가져다 쓰는 것이 아니라,

텍스트 데이터가 근본적으로
어떻게 표현되고 학습되는지 알아야 함

밑바닥부터 시작하는 딥러닝2 발제