NLP 자연어 처리

발표자 KUGGLE 12기 김세영

2025.11.18

목차

NLP? 컴퓨터가 인간의 언어를 이해하고 해석하도록 하는 기술 분야

NLP가 겪는 어려움

1. 모호성 : "배" - 신체, 과일, 선박

2. 문맥 의존성 : "날씨가 '죽인다'" - 긍정적 의미

3. 신조어 및 변화 : 끊임없이 생성되는 새로운 언어

중요성 및 활용 사례

중요성

디지털 사회의 소통과 자동화를 이끄는 핵심 기술

활용 사례

기계 번역, 챗봇 및 가상 비서, 감정 분석, 텍스트 요약, 정보 검색

텍스트 전처리

자연어를 모델이 학습할 수 있도록 정제

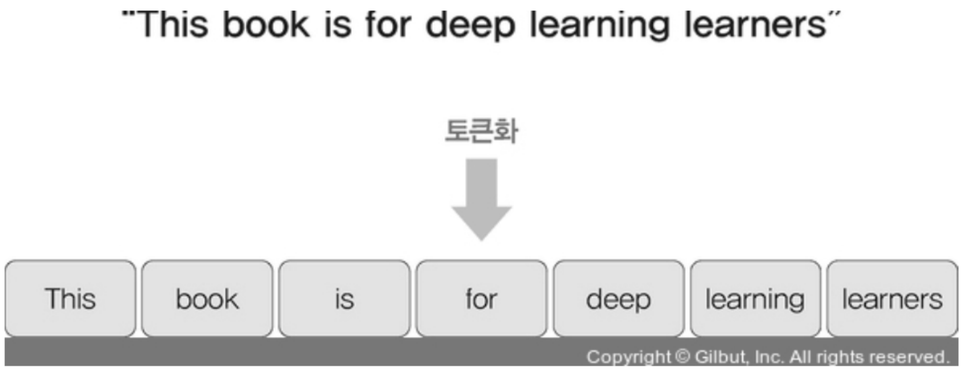
주요 단계

1. 텍스트 수집 : 뉴스 기사, 리뷰, 문서

2. 토큰화 : 텍스트를 '토큰'이라는 작은 단위로 분리 (문장 → 단어)

3. 정제 : 의미 분석에 도움 되지 않는 단어(불용어) 제거, 단어를 어근으로 축소

토큰화 (Tokenization)



텍스트 조각을 토큰이라고 하는 더 작은 단위로 분리하는 방법으로 띄어쓰기, 글자, 형태소 등 다양한 방식으로 처리

불용어 제거

불용어 : 문장에서 자주 등장하지만 실제 의미분석에 도움 되지 않는 조사나 접미사

```
stopwords = ['가입', '기여', '역대','오리지널','회사','최신','소개','관련', '이', '시간', '나오', '있', '가지
'되', '생각하', '수', '그러', '이', '속', '생각', '보', '하나', '않', '집', '없', '살',
'나', '모르', '사람', '적', '주', '웰', '아니', '데', '등', '자신', '같', '안', '우리', '어떤',
'때', '내', '년', '가', '경우', '한', '명', '지', '생각', '대하', '시간', '오', '그녀',
'말', '다시', '일', '이런', '그렇', '앞', '위하', '보이', '때문', '번', '그것', '나',
'두', '다른', '특징', '말하', '어떻', '알', '여자', '남자', '그러나', '개', '받', '전',
'못하', '들', '일', '사실', '그런', '이렇', '또', '점', '문제', '싶', '더', '말', '사회', '정도',
'많', '좀', '그리고', '원', '좋', '잘', '크', '통하', '따르', '소리', '중', '놓']
```

불용어를 포함한 채로 모델 학습 시키면 불필요한 패턴을학습할 위험이 있어 이를 제거 해야함

텍스트의 수치화

필요성

기계는 텍스트를 직접 이해하지 못하므로 수치적 표현(벡터)으로 변환해야 함

빈도 기반 벡터 Bag of Words (BoW), TF-IDF

예측 기반 임베딩 Word2Vec / GloVe

BoW (Bag of Words)

단어의 순서는 고려하지 않고 각 단어가 문서에 몇 번 등장했는지 빈도에만 집중

전체 문서에서 고유 단어의 '사전'을 만들고 각 문서를 '사전'에 있는 단어의 출현 횟수로 표현

'a', 'the' 같이 자주 나오지만 의미 없는 불용어에도 높은 숫자가 부여되어 문서의 실제 의미 왜곡

TF-IDF

단순 빈도(TF)에 얼마나 희귀한 단어인가(IDF)라는 가중치를 곱해 문서 내 단어의 실제 중요도를 반영

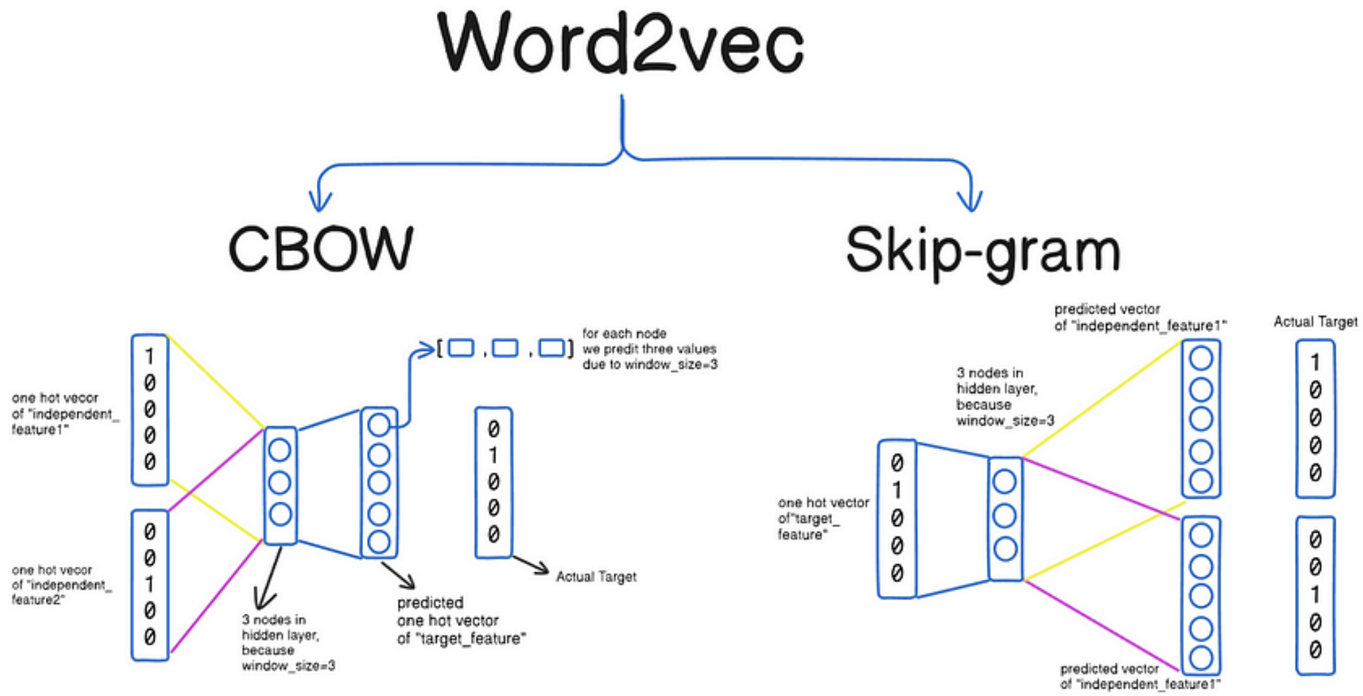
TF-IDF = TF x IDF

모든 문서에서 자주 나오는 단어난 IDF를 낮춰 가중치를 감소하고 특정 문서에만 자주 나오는 단어는
TF와 IDF가 모두 높아 가중치 증가

Word2Vec

단어를 '의미'를 함축한 다차원 공간의 벡터로 변환하는 '예측 기반' 임베딩 모델

'비슷한 문맥에서 등장하는 단어는 비슷한 의미를 가질 것이다'를 가정한 모델



NLP 모델링

코사인 유사도 두 벡터 간 방향 유사도를 측정

전통적 머신러닝 텍스트 분류

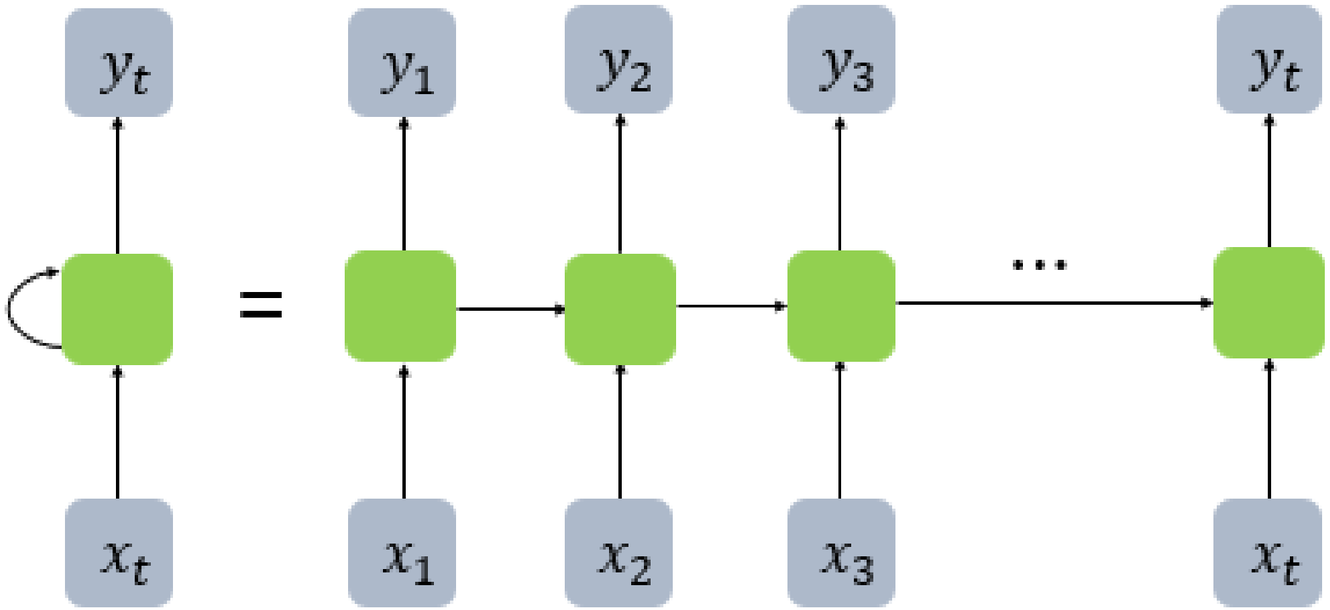
시퀀스 모델링

1. RNN : 단어의 순서를 학습하는 딥러닝 모델

2. LSTM / GRU : RNN의 장기 기억 문제를 개선한 모델

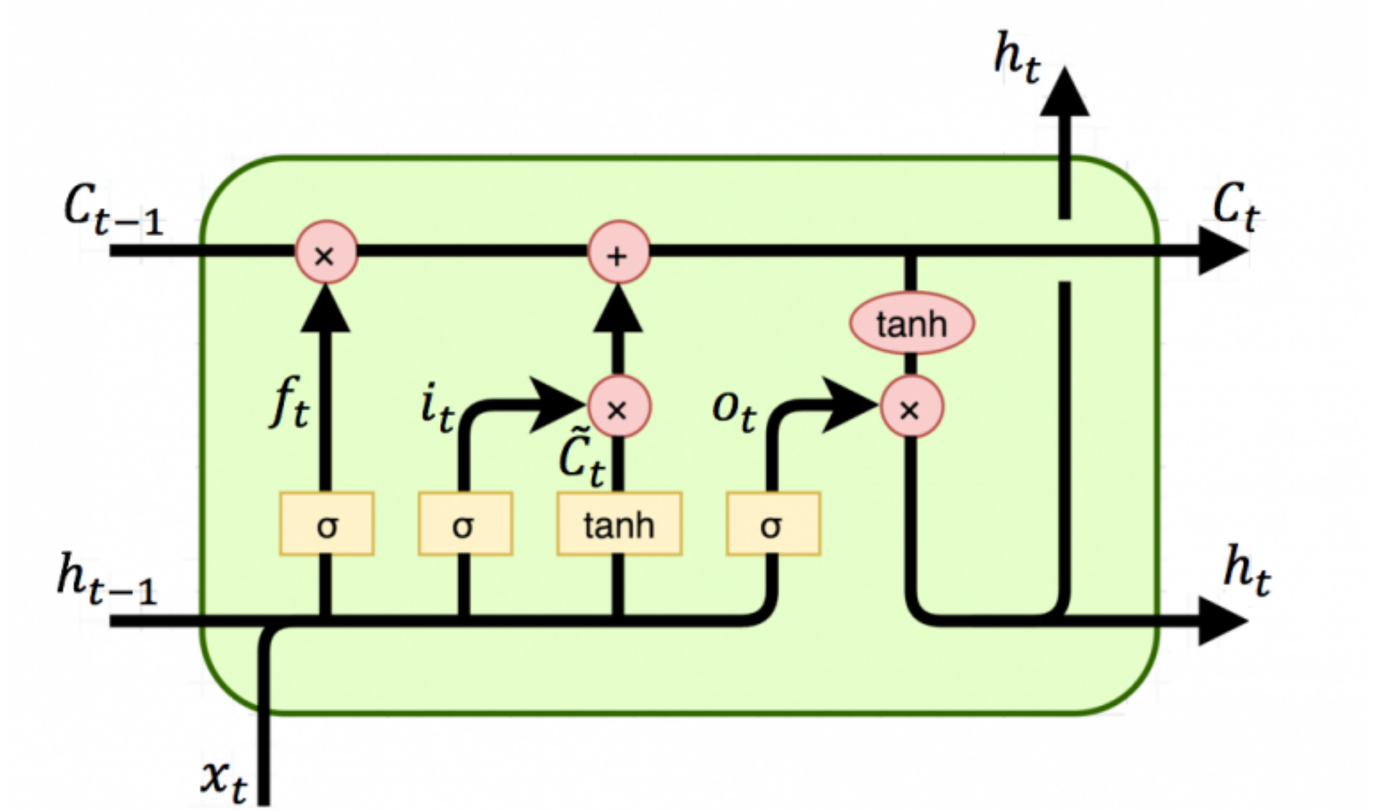
RNN

이전 단계의 정보를 다음 단계로 넘겨주며 Hidden State를 순환시킴



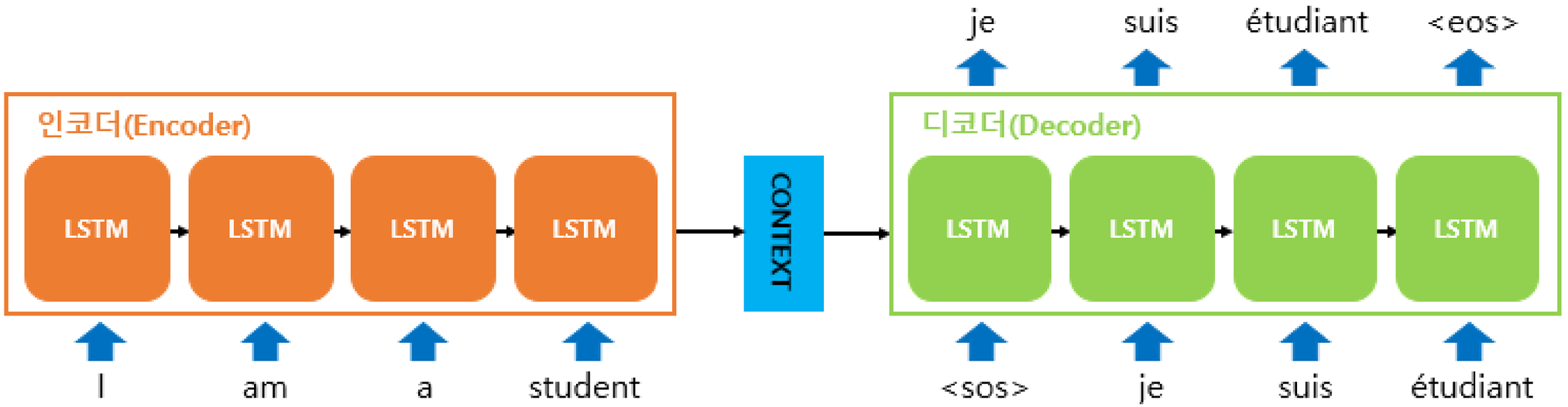
LSTM

RNN에 Gate라는 특별한 장치를 추가해 기억할 것과 잊을 것을 학습



Seq2Seq

입력된 시퀀스로부터 다른 도메인의 시퀀스를 출력하는 모델



최신 동향

트랜스포머 Attention 메커니즘만으로 구현한 자연어 처리 모델

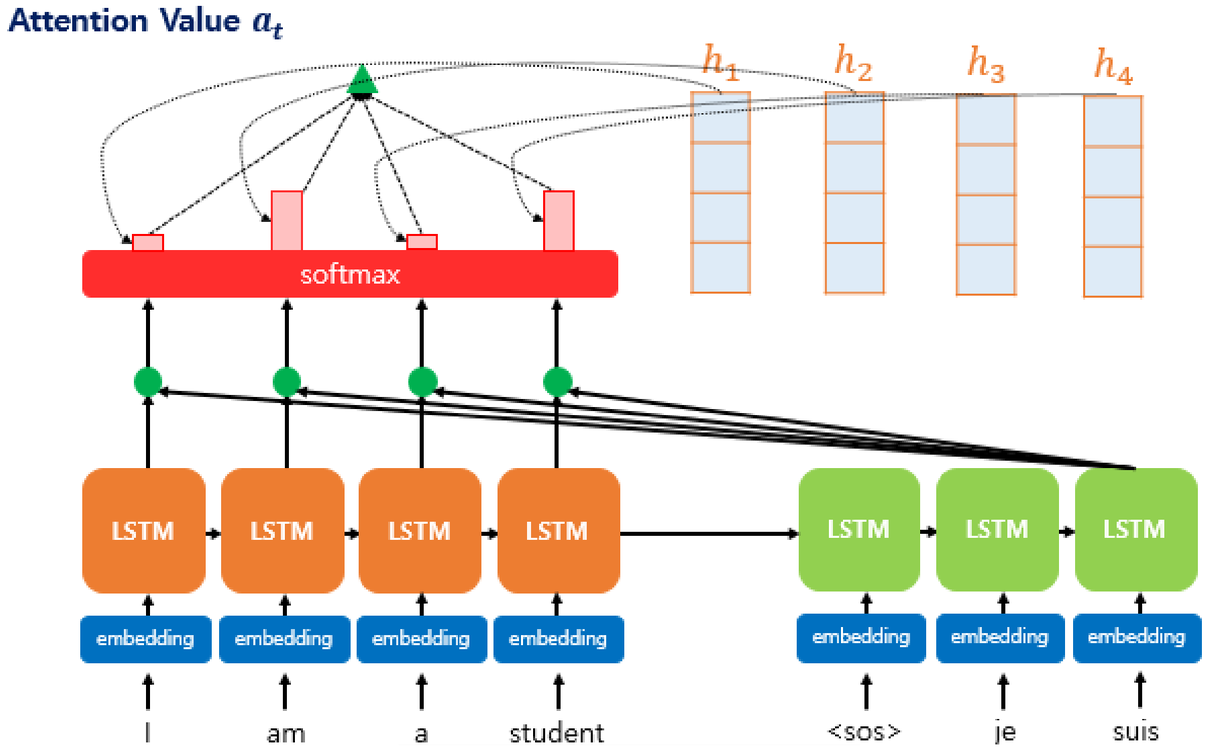
LLM 대량의 텍스트 데이터와 수많은 파라미터를 기반으로 학습한 거대 언어 모델

1. BERT : 트랜스포머의 Encoder 구조를 활용. 문장의 양방향 문맥을 모두 학습해 문맥 이해에 탁월함

2. GPT : 트랜스포머의 Decoder 구조 활용. 다음 단어를 예측하며 문장을 생성하는데 특화됨

Attention

디코더에서 출력 단어를 예측하는 매 시점마다, 인코더에서의 전체 입력 문장을 다시 한 번 참고
이때 해당 시점에서 예측해야할 단어와 연관이 있는 입력 단어 부분을 좀 더 **집중(attention)**해서 보게 됨



트랜스포머 (Transformer)

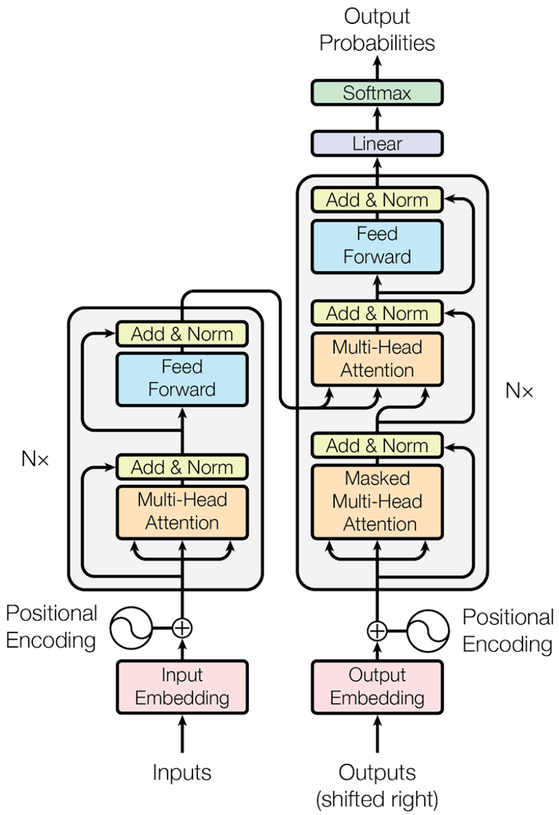


Figure 1: The Transformer - model architecture.

Attention Is All You Need

LSTM의 순환 구조를 없애고 오직 어텐션 만으로 문맥 파악

1. Self-Attention : 문장 내 단어들이 스스로 서로의 관계 파악
ex) 그 동물은 길을 건너지 않았다 왜냐하면 그것은 너무 피곤했기 때문이다
→ '그것'이 가리키는게 길이 아니라 동물임을 단어 간의 Attention만으로 파악

2. 압도적인 학습 속도 : RNN/LSTM은 1번 단어 계산 → 2번 단어 계산 ...처럼 순서대로만 처리 가능했지만 트랜스포머는 모든 단어를 한 번에 병렬로 계산

GPT / BERT

GPT 트랜스포머의 Decoder만으로 아키텍처 구성

- GPT는 오직 왼쪽에서 오른쪽으로만 문장을 읽음
- 지금까지 나온 단어들을 바탕으로 그 다음에 올 단어가 무엇인지 예측하는 방식으로 학습

BERT 트랜스포머의 Encoder만으로 아키텍처 구성

- BERT는 문장 전체를 한 번에 봄
- 문장의 단어 몇 개를 Masking하고 주변 문맥을 이용해 그 빈칸을 맞추는 퀴즈를 푸는 방식으로 학습

수고하셨습니다