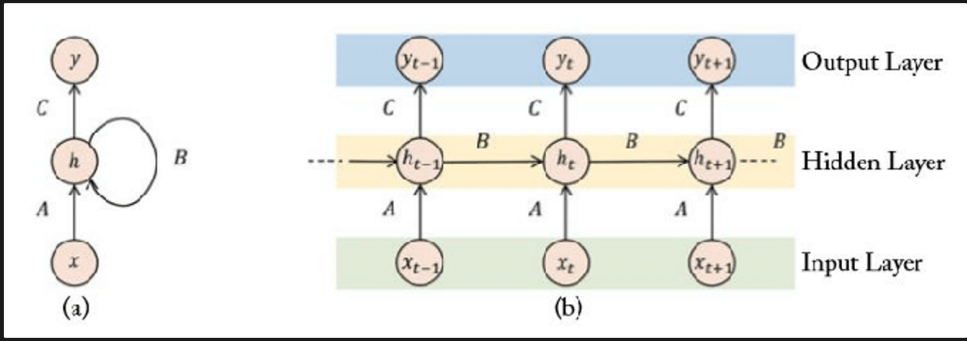
NLP 1주차 스터디

작성일시	2024년 1월 18일 오후 1:58
강의 번호	비어 있음
유형	스터디 그룹
자료	https://colab.re...
복습	☐
텍스트	트랜스포머를 활용한 자연어 처리 1장

1장 트랜스포머 소개 📖

▪ 인코더-디코더 프레임워크

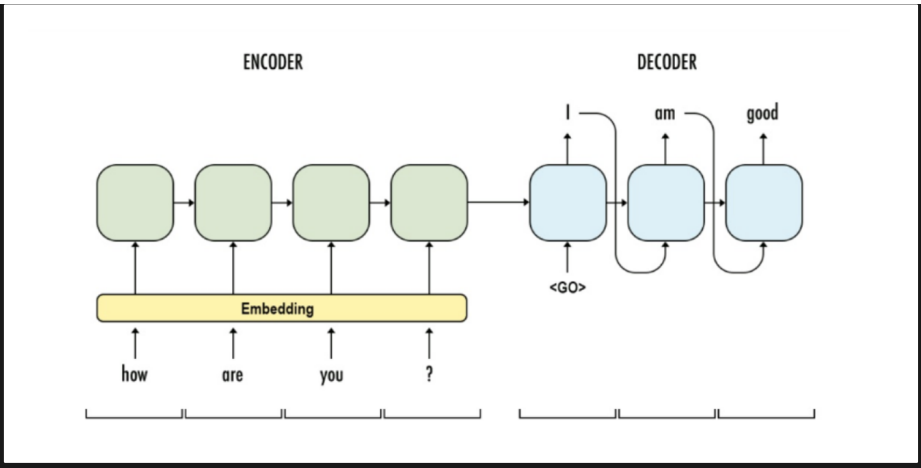
RNN



- 내부에 순환 구조 (feedback loop)가 포함
- 시간에 의존적이거나 순차 데이터 학습에 활용
- 단어 시퀀스를 한 언어에서 다른 언어로 매핑하는 기계 번역 시스템의 핵심

Seq2Seq / encoder-decoder

- 입력 단어는 순차적으로 인코더에 주입
- 인코더는 입력 시퀀스의 정보를 'last hidden state' 수치 표현으로 인코딩
- 이 상태가 디코더로 전달되어 출력 시퀀스 생성
- 출력 단어는 위에서 아래 방향으로 한 번에 하나씩 생성

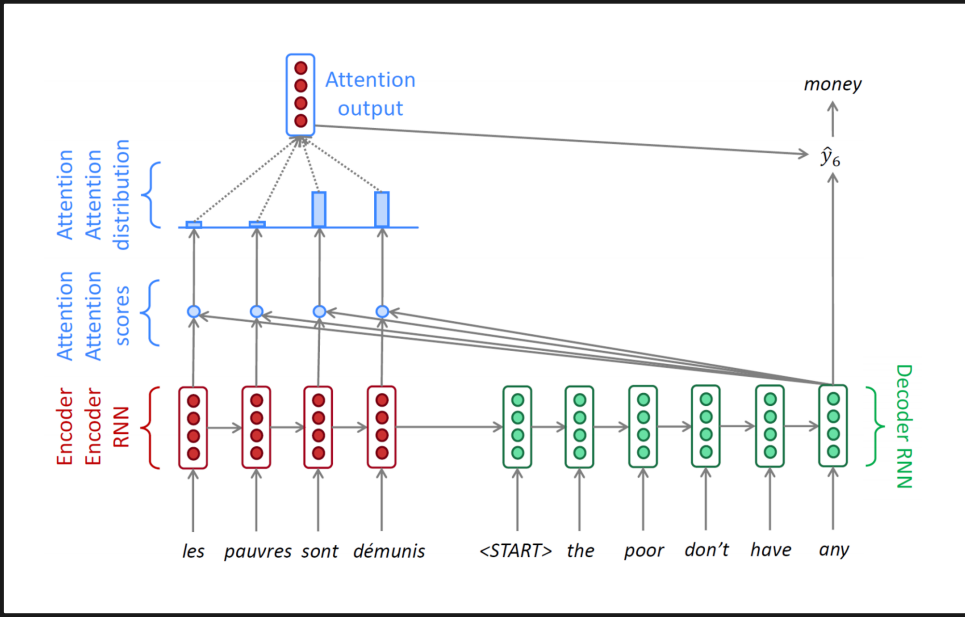


- Embedding : 컴퓨터가 알아들을 수 있도록 언어를 수치화하는 과정

■ 어텐션 메커니즘

디코더가 인코더의 모든 은닉 상태에 접근해 Seq2Seq의 병목을 제거하는 메커니즘

어텐션 메커니즘이 적용된 구조



- 입력 시퀀스에서 스텝마다 인코더에서 디코더가 참고할 은닉 상태 출력
- 모든 상태 동시 사용 > 디코더에 많은 입력 발생하므로 어떤 상태를 먼저 사용할지 우선 순위를 정하는 메커니즘
- 디코더가 모든 디코딩 타임스텝마다 인코더의 각 상태에 다른 가중치 할당

■ 셀프 어텐션

신경망의 같은 층에 있는 모든 상태에 대해 어텐션을 작동시키는 방식

■ NLP의 전이 학습

사전에 훈련된 모델을 이용하는 것

ULMFit 프레임워크의 프로세스

1. **사전 훈련**

: 위키디피아 등 대규모 말뭉치에서 언어모델링 수행

+) 언어모델링 : 이전 단어를 바탕으로 다음 단어를 예측하는 것

2. **도메인 적응**

: 도메인 내 말뭉치에서 언어 모델링 수행

3. **미세 튜닝**

: 언어 모델을 타깃 작업을 위한 분류 층과 함께 미세 튜닝

■ HuggingFace transformers 🤗

요약하자면 여러 회사에서 개발한 트랜스포머 모델을 누구나 쉽게 사용하게 도와주는 라이브러리

텍스트 분류

▶ 🤗 transformers로 예시 텍스트의 감성 분류하기

개체명 인식 (NER)

- 개체명 : 제품, 장소, 사람 등의 실제 객체
- 개체명 인식: 텍스트에서 개체명을 추출하는 작업

▶ NER 로드하기

질문 답변

텍스트 구절과 함께 답을 얻고 싶은 질문을 모델에 전달하고, 모델은 답변 텍스트를 반환

▶ 고객의 피드백에 대해 질문했을 때의 답 확인하기

요약

긴 텍스트를 입력으로 받고 관련 사실이 모두 포함된 간단한 버전을 생성하는 것을 목표로 함

▶ 텍스트 요약 예제

번역

▶ 영어 텍스트 독일어로 번역하기

텍스트 생성

일종의 자동 완성 기능

▶ 고객의 피드백에 빠르게 응답하기

2장 텍스트 분류

텍스트 분류의 예 : 감성 분석, 고객 피드백을 여러 카테고리로 분류 등

그림 2-2 🤗 데이터셋, 🤗 토크나이저, 🤗 트랜스포머스 라이브러리를 사용해 트랜스포머 모델을 훈련하는 전형적인 파이프라인

■ 예제

DistilBERT를 사용해 사람들이 자사 제품에 드러낸 감정 상태를 자동으로 인식하는 시스템 만들기

DistilBERT ?

- BERT의 한 종류로 BERT와 비슷한 성능을 내지만 훨씬 작고 효율적
- 대규모 BERT 모델을 훈련해야 한다면 사전 훈련된 모델의 체크포인트를 바꾸면 됨

+) 체크포인트 : 모델 학습 과정 중 특정 시점의 모델의 가중치와 파라미터를 저장한 상태

■ 데이터셋

분노, 혐오, 두려움, 기쁨, 슬픔, 놀람의 여섯 개의 감정으로 나뉘는 데이터셋 활용

📢 나머지는 코드를 작성하며 마크다운으로 코랩에 표기! 첨부된 링크 확인하기