텍스트 마이닝과 데이터 마이닝

Part 06. 자연어 처리와 실습

정 정 민

programmers

Chapter 16. 문장 분류 문제

1. 문장 분류 문제
2. 딥러닝 모델을 활용한 문장 분류 접근
3. 분류 문제를 넘어

문장 분류 문제

문장 분류 (Sentence Classification)

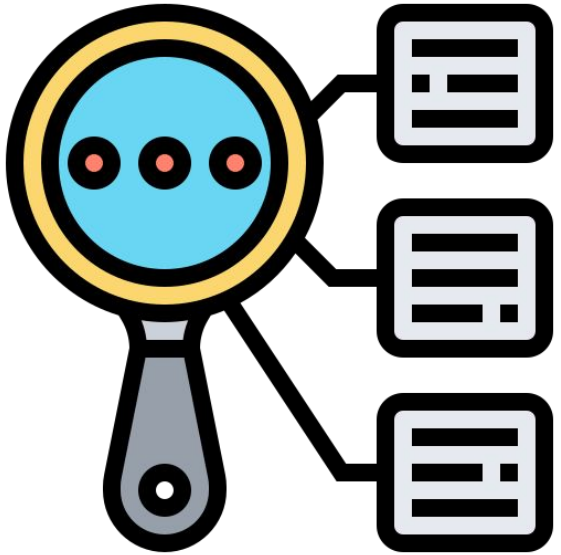
- 텍스트 데이터를 활용해 분류 문제를 푸는 것
- 정해진 클래스 중 어떤 클래스에 속하는지를 판단

- 텍스트의 의미를 이해하고 구조화된 방식으로 분류를 하는 것이 목표

- Part4에서 다룬 감정 분석도 문장 분류의 한 종류
  - 클래스 : 긍정 / 부정
    - '중립' 감정도 다루는 경우가 있음

- 감정 분석 말고도 다양한 하위 테스크가 존재

- 분류 이외의 다른 복잡한 문제에서 문장 분류에 특화된 기술 모델을 사용
  - 분류의 특화된 모델은 문장을 분석하는 능력이 좋고
  - 이 능력을 이용해 다른 문제에 활용



세부 문장 분류와 문장 분류를 활용한 상위 문제들

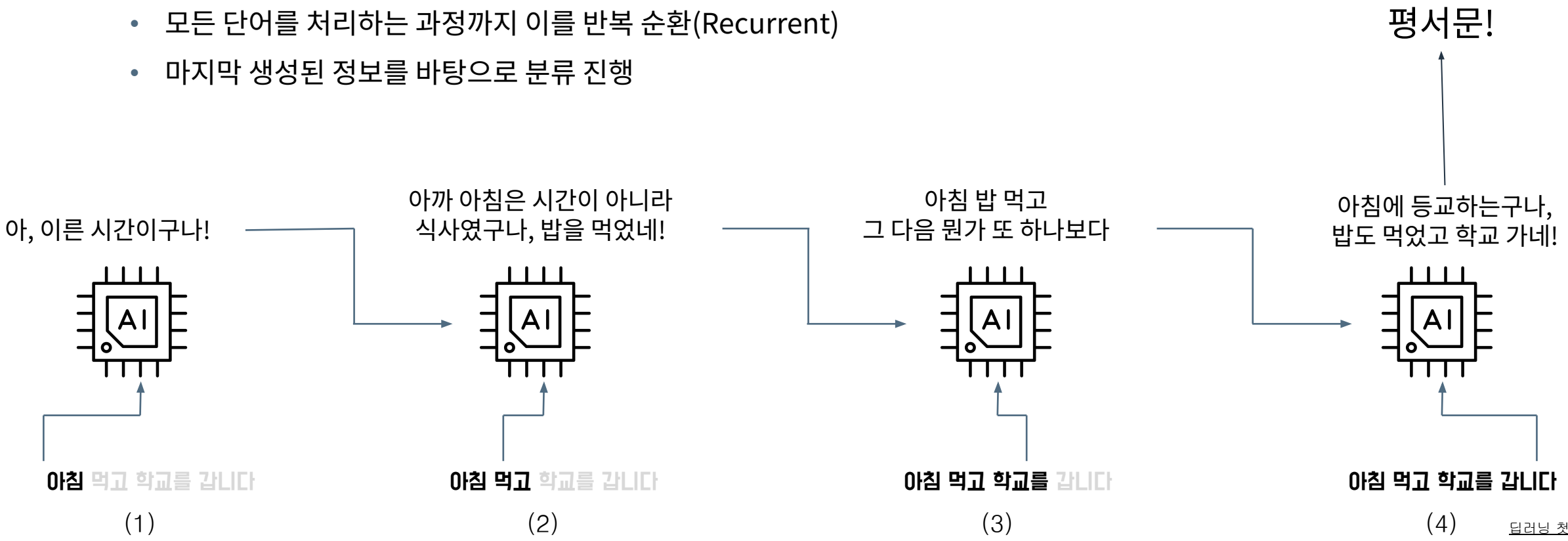
- 문장 분류에는 다양한 하위 문제가 존재
  - 감정 분석 (Sentiment Analysis)
  - 주제 분류 (Topic Classification) : 글이 속한 주제 탐지 (스포츠, 정치, 엔터 등)
  - 의도 분석 (Intent Detection) : 발화의 의도 파악 (정보 요청, 구매, 예약, 질의 등)
    - 빠른 업무 배분 가능
  - 언어 감지 (Language Detection) : 번역에서 언어 감지 등에 사용
  - 등등

- 복잡한 문제를 풀기 위한 베이스 모델로 문장 분류 모델을 사용
  - 텍스트 요약
  - 텍스트 생성
  - QA 챗봇
  - 등등

딥러닝 모델을 활용한 문장 분류 접근

순환 신경망 (Recurrent Neural Network, RNN)

- 딥러닝을 활용한 초기 텍스트 처리 모델
- 사람이 글을 읽고 이해하는 과정을 모방해 모델을 설계
  - 단어를 하나씩 입력 받고
  - 이전에 이해한 내용을 바탕으로 새로운 정보를 생성
  - 모든 단어를 처리하는 과정까지 이를 반복 순환(Recurrent)
  - 마지막 생성된 정보를 바탕으로 분류 진행

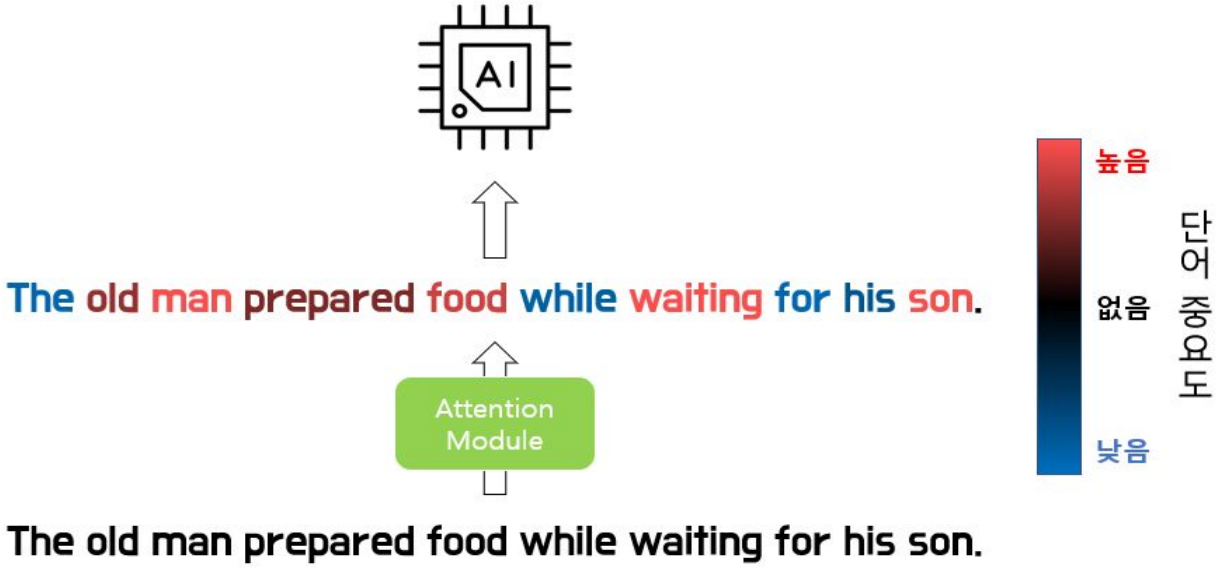


주의 메커니즘 (Attention Mechanism)

- 순환 신경망 (RNN) 이후에 나타난 새로운 분석 방법
- 입력으로 받은 텍스트 정보에서 딥러닝 모델이 주의(Attention)를 집중할 단어를 자동을 판단

- 집중된 단어를 바탕으로 NLP 문제를 처리

- 단어의 정보를 모델 스스로 판단
- 높은 성능, 다양한 정보 추출 가능



https://blog.naver.com/mmismin/223021119710

문장 분류 with 주의 메커니즘

- 주의 메커니즘을 활용해 분류 문제를 풀기 위해서는
- 이것으로 생성된 문장 임베딩 값을 활용

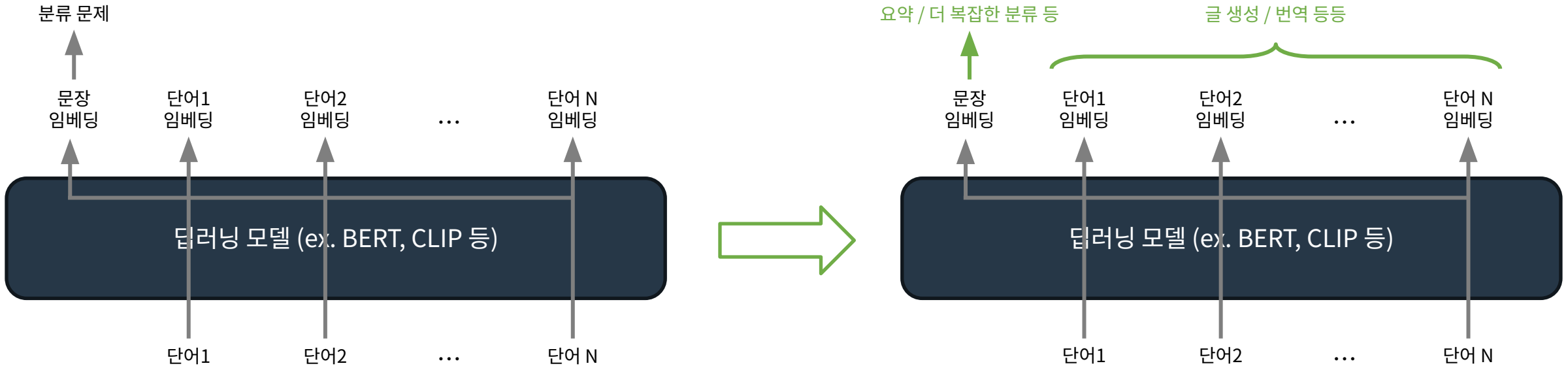
- 정리된 문장의 정보를 바탕으로 원하는 타겟 클래스를 예측
- 문장 임베딩은 Attention이 입혀진 단어를 바탕으로 생성
  - 중요도가 적용된 전반적인 문장의 의미를 생성

분류 문제를 넘어

분류 문제를 넘어

- '분류'는 (대부분의 데이터에서) 가장 기본이 되는 문제
- 또한, 학습을 위한 데이터 준비 입장에서도 분류 데이터가 만들기도 쉬움
  - 특히 데이터 레이블을 만드는 상황에서!

- 이러한 이유로, 분류 문제로 데이터의 특성을 익힌 모델을
- 분류보다 더 복잡한 문제에 적용하는 상황이 많음
- 이를, 전이 학습(Transfer Learning) 혹은 미세 조정(Fine-Tuning)이라고 함



E.O.D