NLP Model

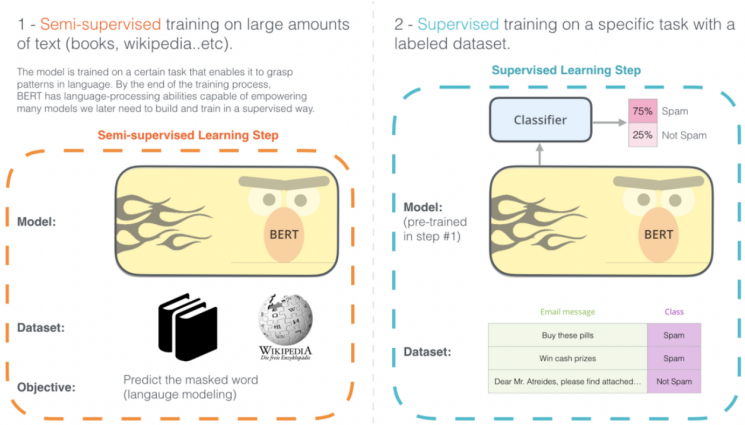
DA 20 기 김소정

NLP(자연어 처리)는 인공지능의 한 분야로써 사람의 언어를 컴퓨터가 알아듣도록 처리하는 인터페이스이다. NLP 의 최종적인 목표는 컴퓨터가 사람의 언어를 이해하고 여러가지 문제를 수행할 수 있도록 하는 것이다.

BERT

Bidirectional Encoder Representations from Transformers 의 약자인 BERT 는, 2018 년 10 월에 논문이 공개된 구글의 새로운 Language Representation Model 이다. 입력 문장을 양방향으로 분석하는 모델로, 주어진 시퀀스 다음 단어를 맞히는 것에서 벗어나, 일단 문장 전체를 모델에 알려주고 빈칸(MASK)에 해당하는 단어가 어떤 단어일지 예측하는 과정에서 학습한다. 이를 마스크 언어 모델이라 하며, 이 덕분에 BERT 의 Embedding 품질이 기존 다른 모델보다 좋다.

BERT 모델은 이름에서도 알 수 있듯이, Transformer 모델에 기반해 있다. Transformer 모델은 Input Text 를 입력 받아 Attention 메커니즘을 통해 인코딩, 디코딩하는 방식의 모델이다. LSTM + RNN 과 같이 각 단위 워드 벡터가 시간의 연속성을 기억하고 있을 필요가 없기에 좋은 성능을 보인다.



http://jalammar.github.io/illustrated-bert/

BERT 는 이러한 Transformer 모델 중 인코더만 쓰는 형태이다. 또한 BERT 는 방대한 양의 Corpus(위키피디아, 웹문서, 책정보 등)를 이미 트레이닝 시킨 언어 처리 모델이라는 점에서 큰 장점을 갖는다. 앞서 언급한 MASK LM 기법과 NSP(Next Sentence Prediction) 기법을 이용해서 pre-train 이 된 모델이다.

해당 논문과 학습된 모델이 세상에 등장했을 때 자연어 처리 분야에서 큰 패러다임의 전환이 있었을 만큼 획기적이고 중요한 모델이다.