<https://looperxx.github.io/CS224n-2019-20-The%20Future%20of%20NLP%20%2B%20Deep%20Learning/>

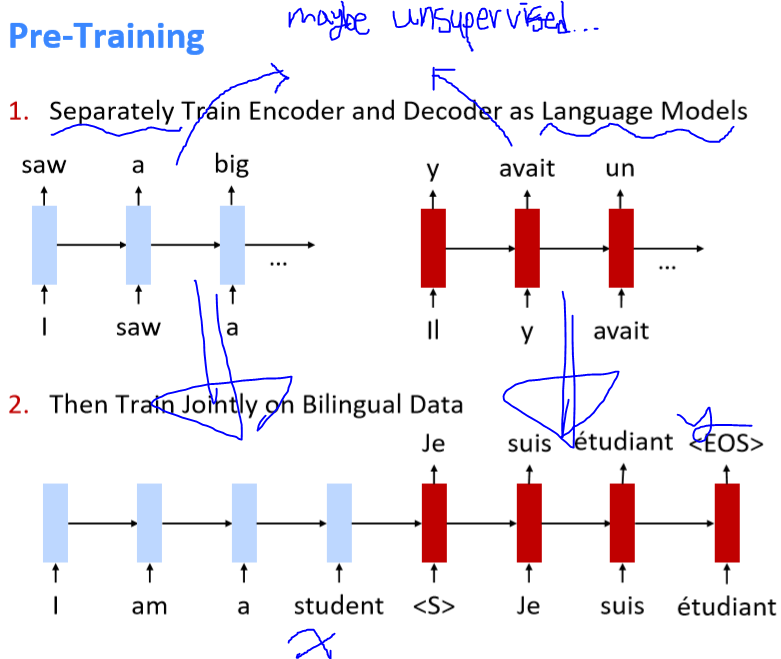
**Future of Deep Learning + LNP**

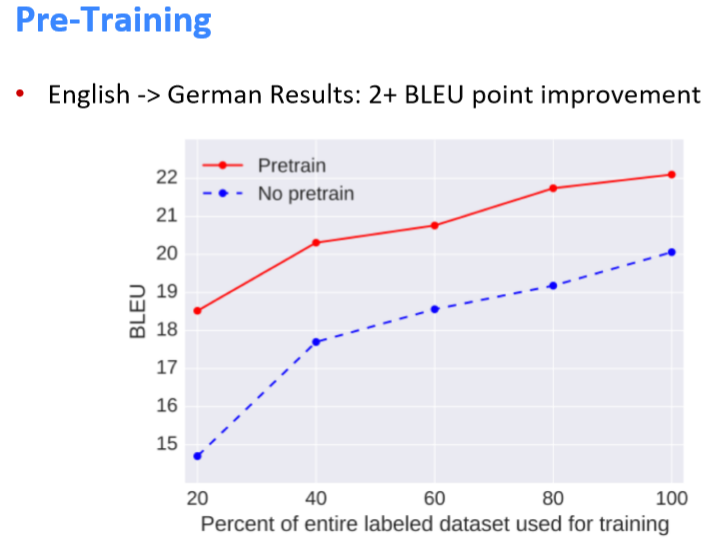
English: most tasks 10만개 이하의 데이터들 / 심지어 다른 언어는 더 적음. -> 답은 unlabeled data를 어떻게든 이용하면 된다.

**Using Unlabeled Data for Translation**

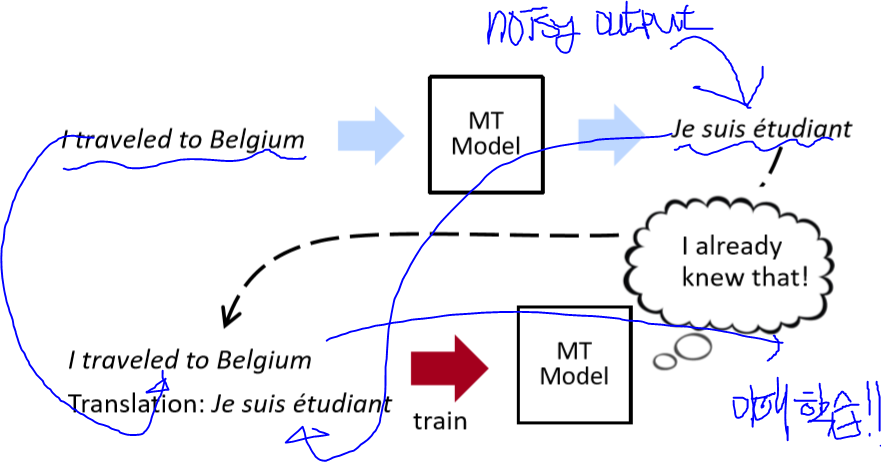
Monolingual text를 주로 이용하며 얻기가 쉽다.(unlabeled 상태)

Pre-Training에 이런식으로 사용할 수 있다. (fine-tune!)

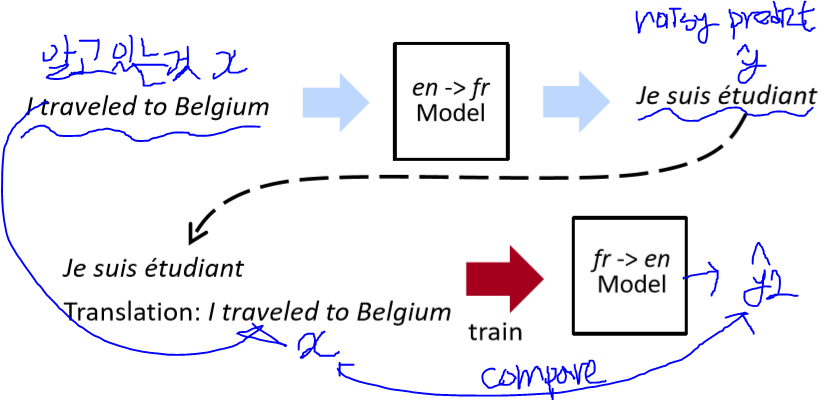




Pre-Training 시에 인코더와 디코더를 따로 학습하게 된다. 따라서 pre 학습 과정 중에 interaction이 없을 수가 있는 것이 문제이다. -> (pre)Self-Training



하지만 Circular 된다는 점이 뭔가 걸린다. -> 모순적. 따라서 이대로 사용하지 않고 Back Translation을 사용한다.(no longer circular, 중간의 output & input이 예측한 것이고 나머지는 이미 알고 있는 것이 된다.)

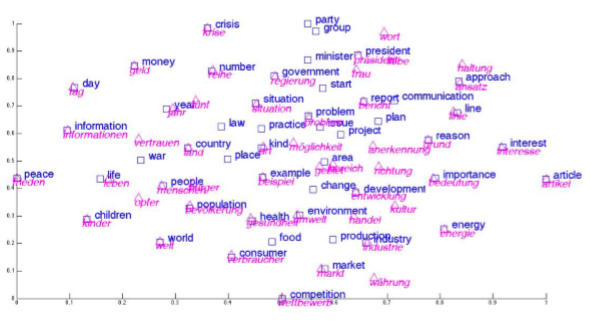


-> NMT에서 매우 좋은 성능을 보이고 있다. (NMT 1등) (labeled pairs = 4.5M / unlabeled sentence = 226M)

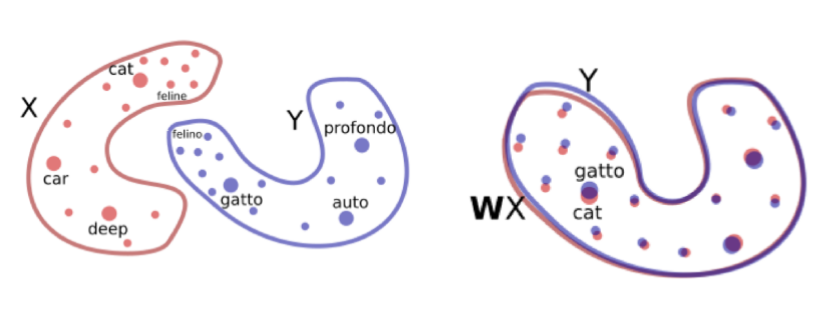
그렇다면 만약 labeled pairs가 아예 없고 only unlabeled data만 있을 때는??

일단 word translation 관점에서 살펴보자. 이것을 이용한다면 어떻게든 될 것 같다..?

각자의 언어에 맞는 word embedding이 있다 했을 때 space의 특성이 다를 것이다. 그것을 비슷하게 맞춰서 같은 의미에 해당하는 단어는 비슷한 위치에 있게 한다. (아래와 같이)



이것이 가능한 이유는 각 언어들에 대한 embedding 구조가 비슷하다고 가정했기 때문이다.



Word2vec으로 각 언어에 대한 embedding을 배운다.(X, Y) 그 다음에 (orthogonal mat -> overfit 방지)W를 통해서 비슷하게 포개 지도록 학습한다.(이때 학습은 adversarial training을 이용한다.)

Discriminator는 임베딩 단어가 Y에서 왔는지 또는 Wx에서 왔는지 구별하도록 한다.

그렇다면 이제 sentence to sentence 로는 어떻게 translation 할 것인가??

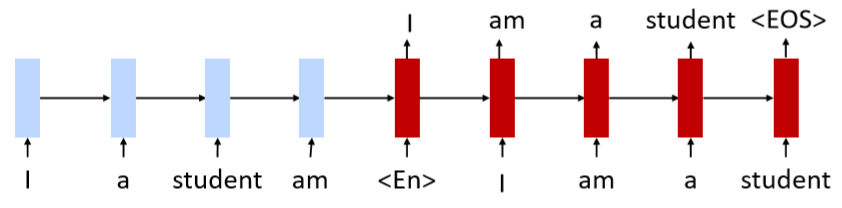
일단 seq2seq without attention을 사용. 하지만 same encoder & decoder를 사용하는 점이 다르다.(input & output 언어에 상관없이) / 일단 모든 언어는 cross word2vec으로 임베딩

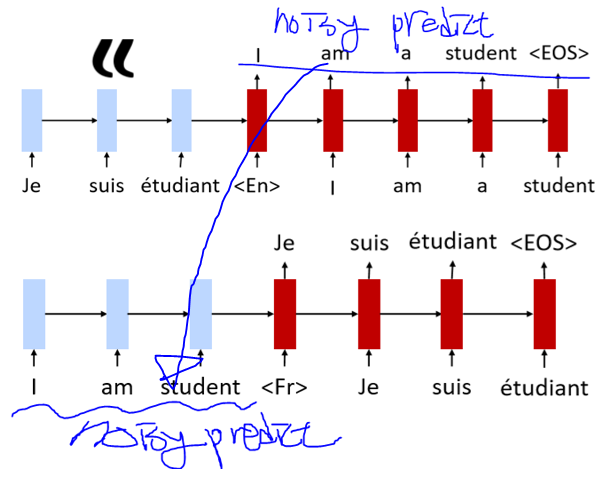
현재(english -> french라 생각한다.)

처음에는 de-noising 하기 위한 auto-encoder를 사용(input과 output이 같은!)

이러면 인코더가 인풋 문장의 정보를 압축해서 벡터로 디코더에 전해준다. 하지만 아웃풋이 역시 인풋과 같은 것이기 때문에 auto-encoder는 이런 상황에서 원래 문장을 recover하기에 좋다.

그 다음에 back translation을 하게 된다.

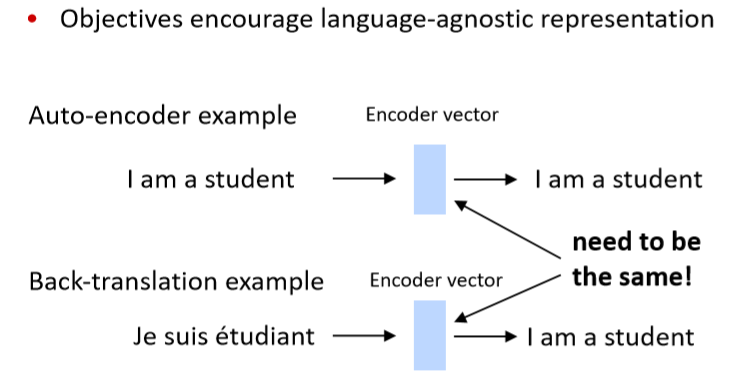


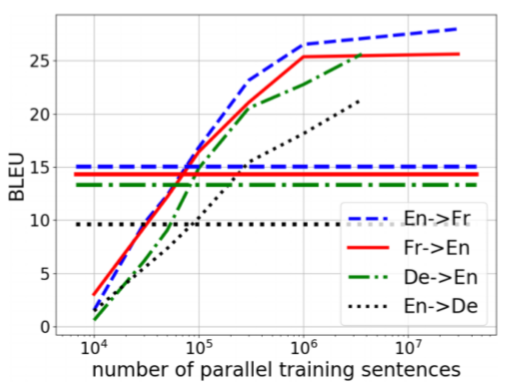


자꾸 공유하는 것을 강조하는 것 같은데 당연한데 왜 강조하지??

아마 pre-train 과정이 2단계로 나뉘는데(앞서는 unlabeled = pre-train / labeled = fine-tune) 이 2단계는 공유돼서 진행되야 하는 것을 강조하는 것 같음.

아래 그림 참조



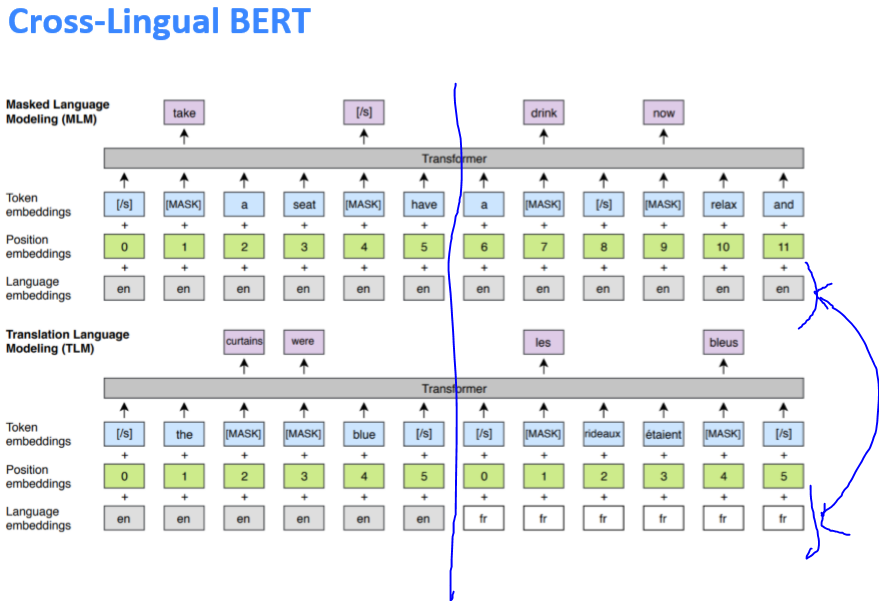


하지만 pure unsupervised learning은 좋지 않다고 한다.

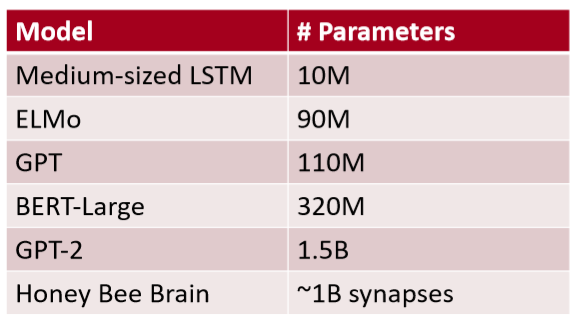
일단 vocab을 어디서 따로 구해야한다.(seed vocab 필요) 또한 학습할 데이터가 무지 많아져서 not so fast..?

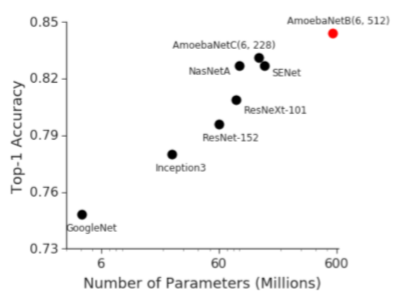
또한 NMT에서 Facebook이 BERT를 학습할 때 Cross-Lingual BERT를 이용하여 LM을 학습.

NMT Task에서 이런 방식을 기반으로 많이 학습한다.

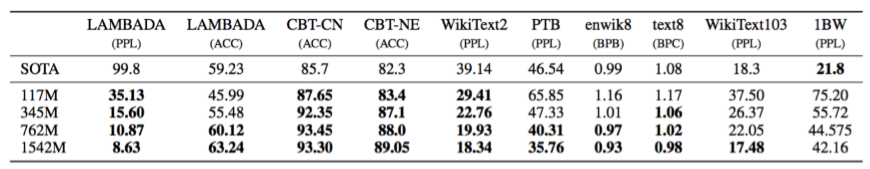


또 다른 LM인 GPT-2 (아마 unsupervised learning?)을 설명

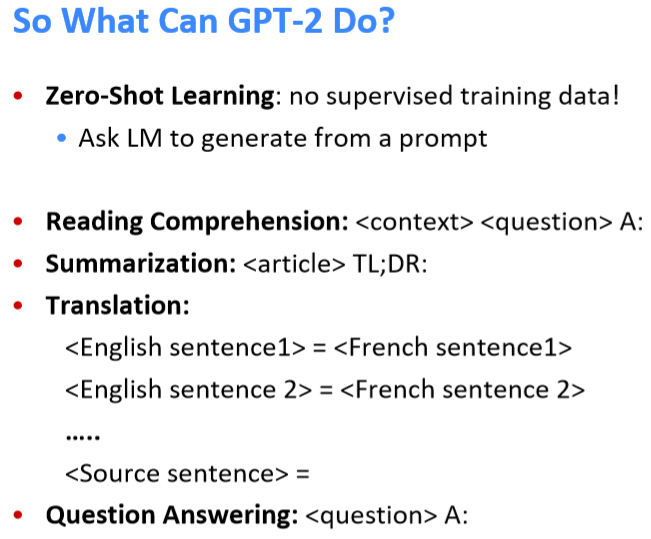
처럼 huge -> good results

better hardware & data, model parallel

GPT-2: 역시 big Transformer LM, 40GB Text

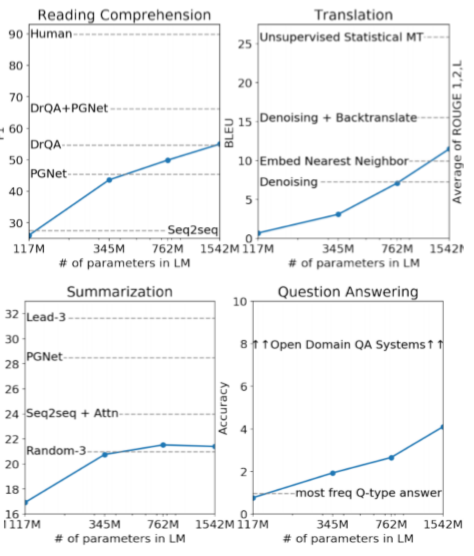


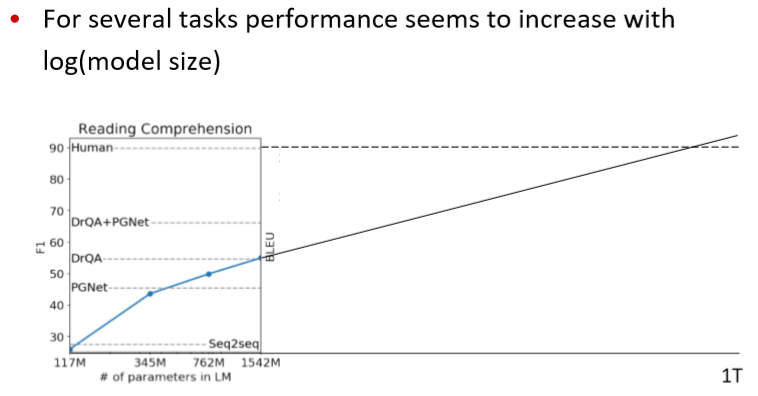
Pre-train만 했는데 이 정도.. fine-tune 안한 결과임.

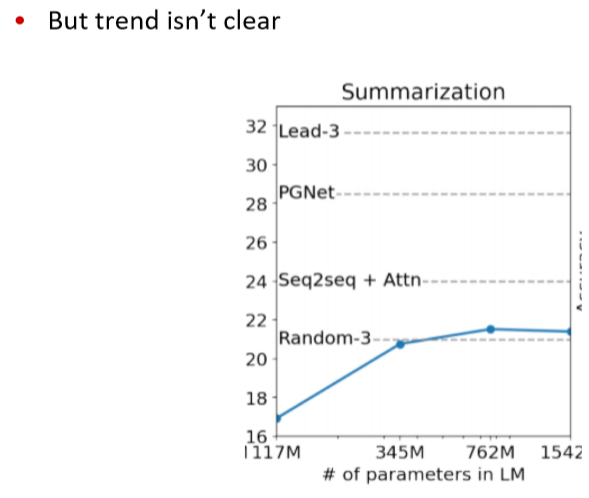




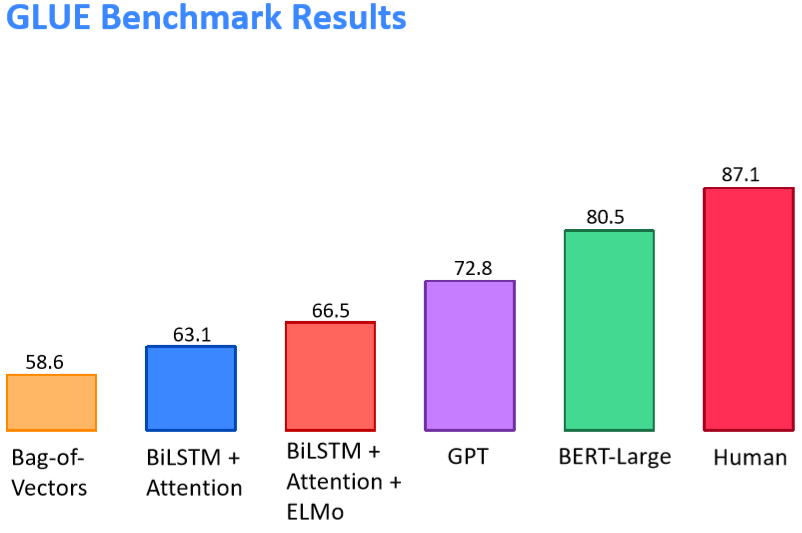
제로샷 러닝이란..(위에)



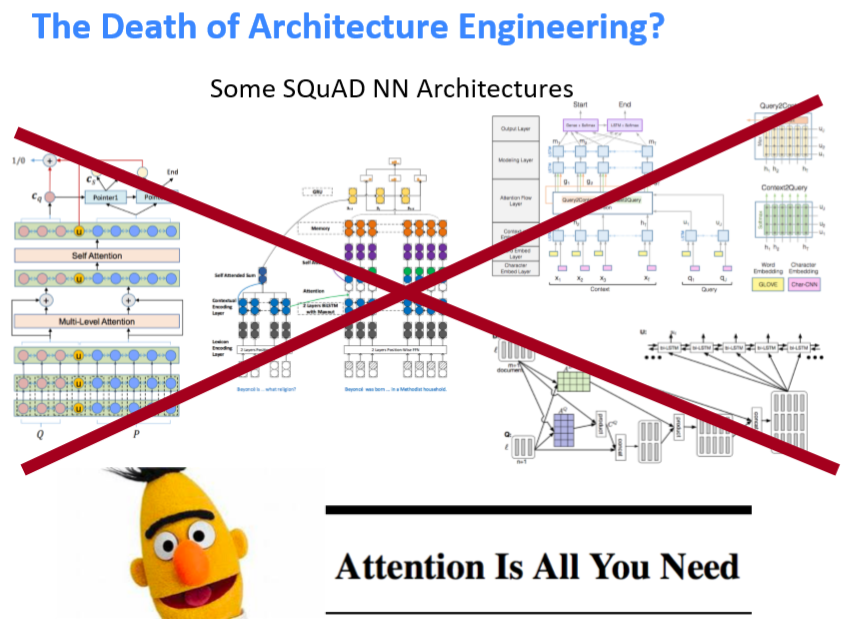




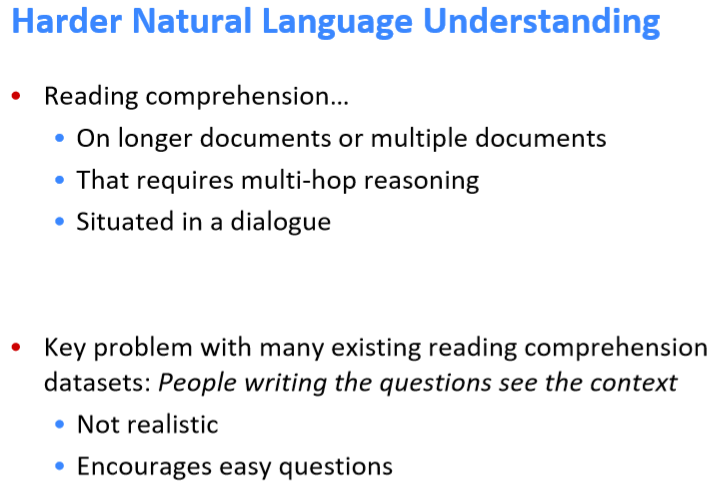
평가적인 부분 ~ SNS의 주관적 평가는 생략



Compendium of 10 natural language understanding tasks

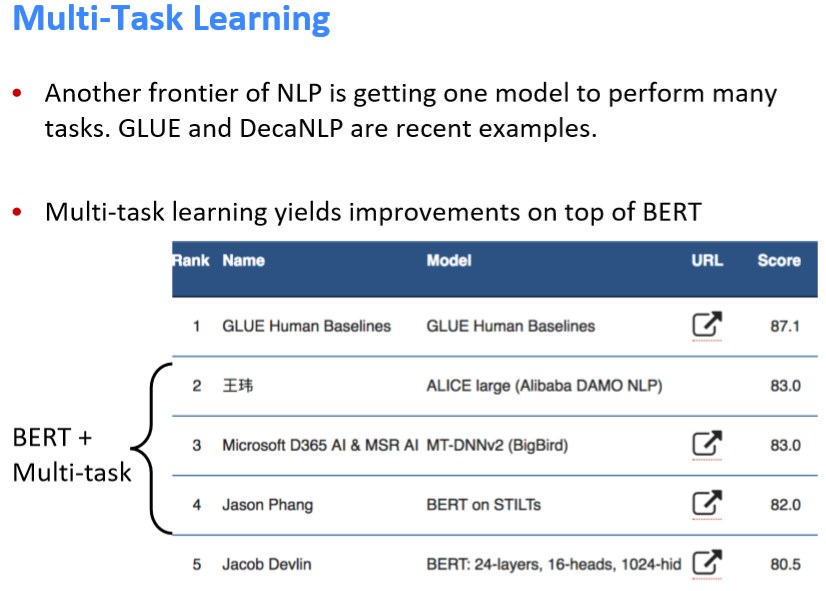


트랜스포머가 짱이다.

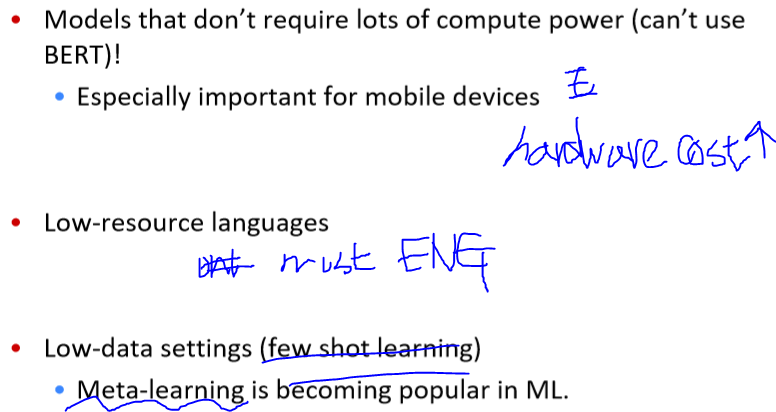


의 어려움이 있다.

하지만 여러 benchmark에서 BERT는 뛰어나며



Decanlp 처럼 multi-task learning도 같이 BERT와 이행되고 있다.



또한 model이 왜 이런 결과를 냈는지 어떻게 냈는지 왜 안 좋은 결과를 냈는지 이런 explainable한 것을 알아내는 것도 중요하다. – 끝 -