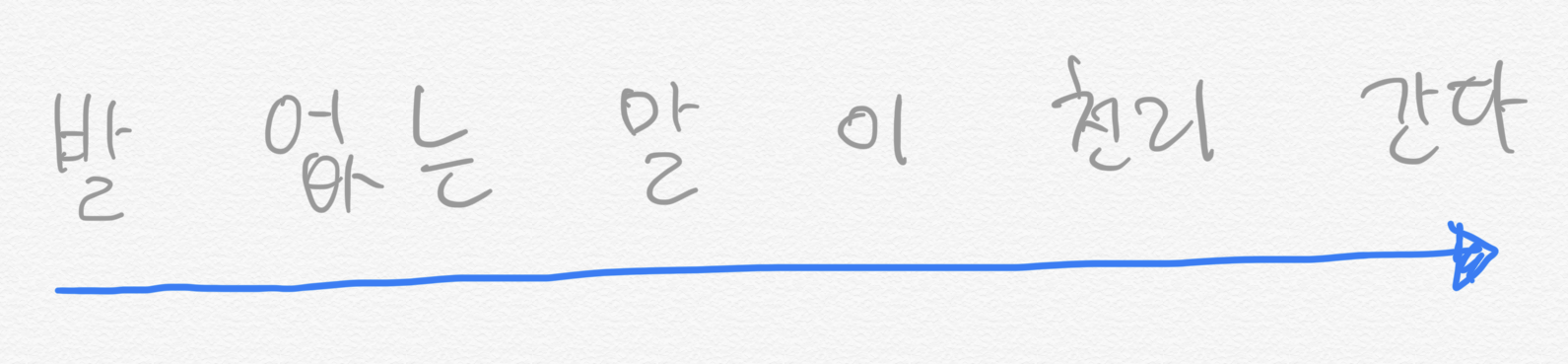
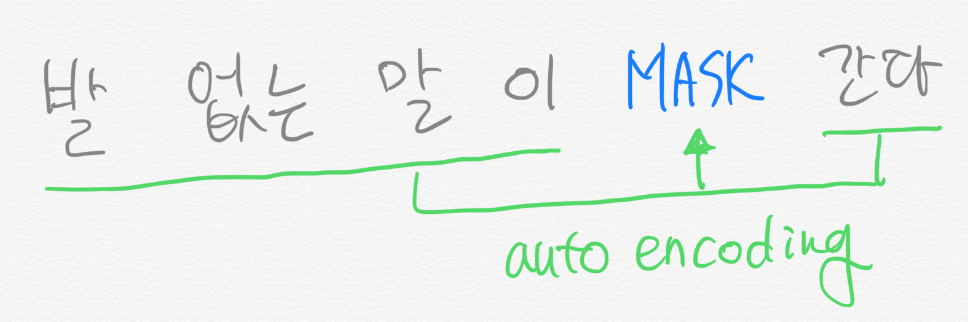
<https://ratsgo.github.io/natural%20language%20processing/2019/09/11/xlnet/>

XLNet(2019, Yang, google)

공개 당시 20개 데이터셋에서 최고 성능, 기존 강자인 BERT보다 크게 앞섬.(일부 데이터셋에서는)

트랜스포머를 개선한 트랜스포머-XL의 확장판(eXtra-Long), 기존 트랜스포머보다 좀 더 넓은 범위의 문맥을 볼 수 있다.(트랜스포머-XL은 이 점을 강조)

임베딩 모델의 최근 흐름 크게 두 개!

1. AutoRegressive  
     
   ELMo나 GPT를 여기에 분류할 수 있다.  
   두 모델 모두 이전 문맥을 바탕으로 다음 단어 예측하는 과정!  
   
2. AutoEncoding(입력과 출력이 같다?!)  
     
   BERT가 대표적, 일부에 마스킹(노이즈)을 주어 문장을 원래대로 복원!  
   = 실제로 어떤 단어일지 맞추는데 중나점을 둔다.  
   이런 맥락에서 BERT를 Denoising Autoencoder라고 표현한다.(노이즈가 포함된 입력을 받아 해당 노이즈를 제거한 원본 입력을 출력으로 한다.)  
   

AR -> 문맥을 양방향으로 볼 수 없다.

AE -> BERT는 양방향이다. 그림 2처럼 앞뒤 문맥을 모두 살피기 때문이다. = 덕분에 downstream task에서 상당 기간 절대 강자  
하지만 큰 문제는 마스킹 처리한 토큰들을 서로 독립이라고 가정한다는 점이다.  
이 경우에 마스킹 토큰들 사이에 의존 관계(dependency)를 따질 수 없다..

더군다나 fine-tune 시에는 masking을 하지 않는데, 이 달라진 방식의 학습 방법은 일반화를 낮출 수 있다.

마지막으로 BERT는 긴 문맥을 학습하기 어렵다고도 한다.

따라서 이 두 방식에 대한 단점을 극복하기 위해 permutation LM 제시.

간단히 말하면 토큰을 랜덤으로 셔플한 뒤 그 뒤바뀐 순서가 마치 원래 그랬던 것인 양 언어 모델을 학습하는 기법이다.