<https://yhdosu.github.io/2019/11/12/T5.html>

<https://jeonsworld.github.io/NLP/t5/>

**Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer – T5  
(Text-to-Text Transfer Transformer)  
(JMLR 2020)**

**Abstract**

NLP의 다양한 Task를 한번에 처리하는 프레임워크가 많음

여기서 입출력 방법을 바꿨음 -> 인코더와 디코더 모두 이용

단, 모든 문제를 Text to Text 형태로 변경해서 해결함  
ex) classification 문제라도 label 분류하는 것이 아니라 decoder에서 text를 생성하는 형태  
(마치 decanlp에서 다양한 데이터들을 QA 형태로 바꾸는 개념이랑 비슷한 듯)

Summarization, QA, Text Classification 등에서 SOTA 달성

**Introduction**

* Pre-Training 모델의 변천사 & 왜 이 모델이 필요한지
* Seq2Seq 변천사 & 왜 Text2Text 까지 왔는가

**Setup**

T5 모델 & Colossal Clean Crawled Corpus(C4) 데이터셋의 설명

**Model**

일반적인 Transformer 형태

**C4**

Common Crawl을 발전

* 한달에 약 20TB 수집
* 수집 시 HTML Tag 등의 Markup or Non-Text들은 제거됨
* 하지만 아직 Menu & Error Message 등은 남아 있음 = Heuristic Cleanup 필요
  + JS Warning 시작 or JS 라인 날리기
  + “lorem ipsum” 포함 페이지 날리기
  + 나쁜 단어 포함된 것 버리기
  + {} 등의 코드 포함 버리기
  + Langdetect를 사용해 영어가 아닌 것은 제거
  + Terminal punctuation mark (.)으로 끝나는 문장만 포함
* 정제하니 총 750GB, Tensorflow Dataset에 공유

**Downstream Tasks**

GLUE, Super GLUE, 번역, QA, 요약, 분류 등 전 방위에 걸쳐 테스트

**Input & Output format**

Unified Text2Text 모델이여서 입력 앞에 Task 정보를 Prefix

또한 각 Task에 맞게 입력을 조금 변형시킴

* MNLI(classification task)  
    
  Input: mnli premise: I hate pigeons. hypothesis: My feelings towards pigeons are filled with animosity  
  Output: entailment  
    
  참고: entailment, contradiction, neutral 외에 출력한다면 contradiction으로 간주
* STS-B(regression task)  
    
  21개의 클래스를 가진 분류 문제로 변환해서 진행 (유사도를 1~5 사이의 0.2 stride로 위처럼 classification 한다)
* DPR(Winograd Tasks)  
    
  찾아야 할 pronoun의 위치를 표시해야 함

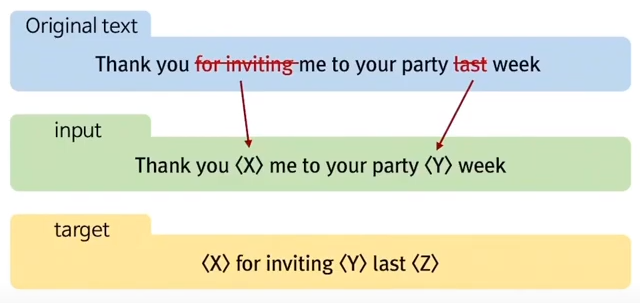
**Architectures**

기존 BERT 스타일은 Encoder-only여서 Classification & Span Prediction에만 특화됐다고 볼 수 있다 (single prediction per input token & single prediction for an entire input sequence)

하지만 T5는 인코더-디코더 구조로 모든 NLP에서 동일한 모델, loss, hyper parameter 사용 가능 = 강력

즉 Fine-Tuning 시에 T5는 모든 Task에서 Text2Text 형태로 입출력하기 때문에 어느 정도 더 범용적이다?

역시 여기서도 BERT의 기본 스타일을 변형시켜 학습하는 Modified MLM의 pretrain이 있다



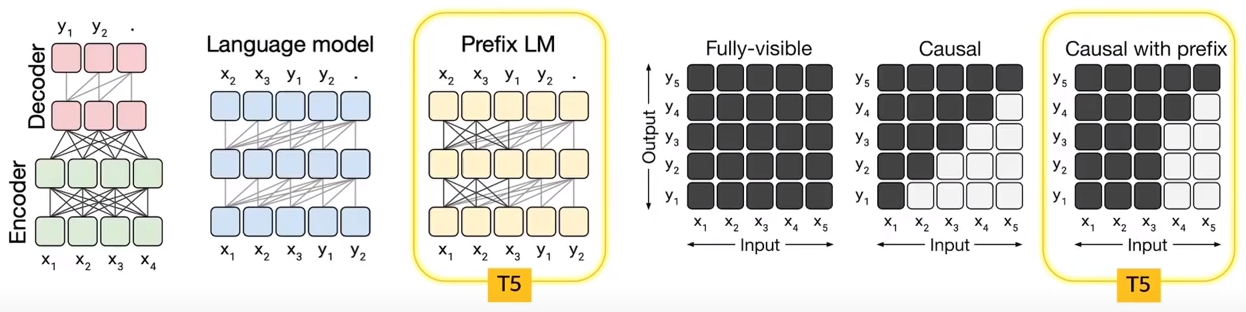
BERT는 하나의 token에 masking -> 연속된 tokens를 하나의 mask로 바꿈(BART와 비슷)

Input에서 masking 되지 않은 부분을 Target에서 맞춤(MASS와 비슷)

Output level에서는 FFNN + Softmax로 시퀀스 생성

MLM – bidirectional Model

인코더-디코더 구조로 입력 & 타겟 가짐



**Baseline**

Standard Transformer using a simple denoising objective 방법 사용 (noise를 원래 문장으로 복원하는 것이 목표, 기존 BERT 처럼)

Standard 인코더-디코더 트랜스포머, 파라미터 사이즈는 BERT BASE로 세팅(단 인코더 & 디코더라 딱 2배), Generative & Classification에서 좋은 성능

**Training**

Text2Text Task 형태로, Learning Rate schedule에 Inverse Square Root 사용, seq\_length = 512, batch size = 128, pretrain steps = 524288, finetune steps = 262,144 (2의 제곱수)

**Vocab**

Sentence Piece