KOREA UNIVERSITY

DSBA CS224n 2021 Study

[Lecture 14]

T5 and large language models: The good,the bad, and the ugly

DSBA
Data Science & Business Analytics

고려대학교 산업경영공학과

Data Science & Business Analytics Lab

발표자 : 이유경

00 발표 목차
Contents

1 Before T5

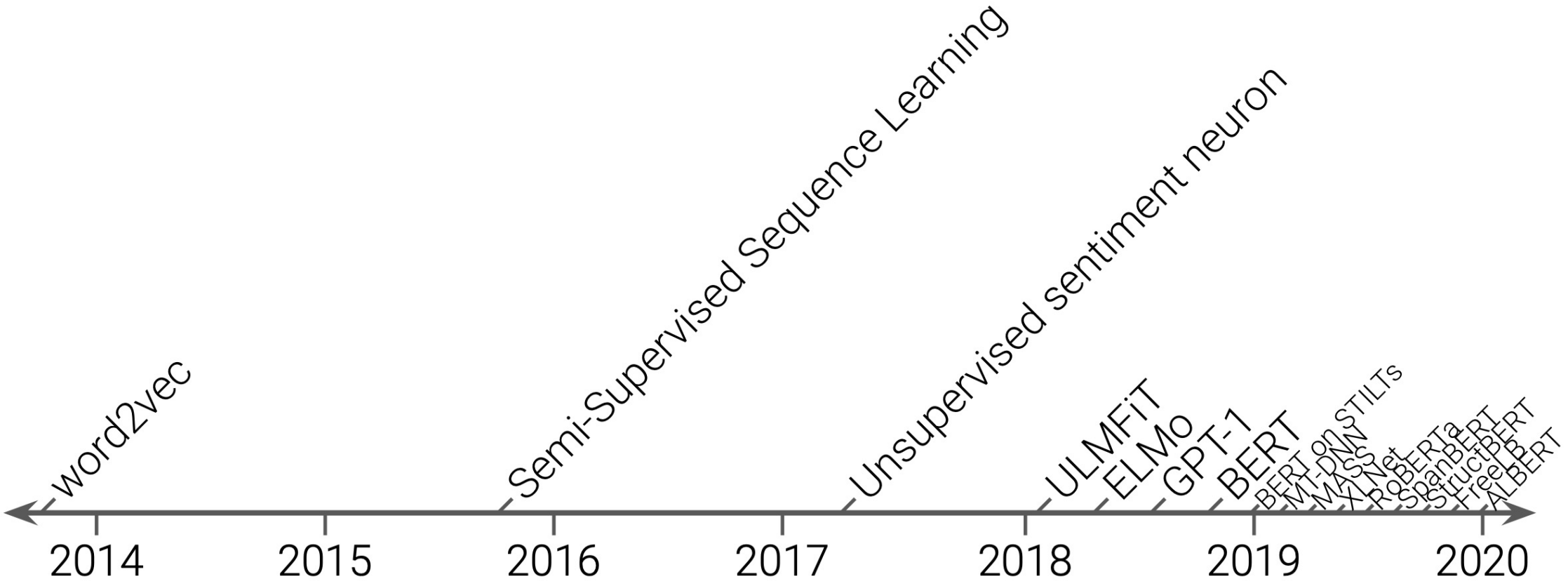
2 T5 Model

해당파트는 효율을 위해
미리 정리해둔 Transformer to T5자료와 혼합하여 사용합니다

3 Other topics

01 Before T5

Overview



01 Before T5

Overview

Transformer to T5 (20.5.25)

Seq2Seq

Neural Machine Translation

Encoder - Decoder 구조의 모델

Attention + Seq2Seq

Neural Machine Translation

Decoder가 Source sentence의
중요한 정보에 집중하게 하자 !

Transformer

Neural Machine Translation

Self attention ,
Multi-head attention

GPT - 1
Transformer Decoder block

Task Agnostic
: Pretraining + Finetuning

언어 자체를 이해 할 수 있는
좋은 representation을 학습해보자 !

BERT
Transformer Encoder block

Task Agnostic
: Pretraining + Finetuning

언어를 더 잘 이해하기위해
Bidirectional하게 맥락을 파악해보자 !
모든 NLP Task에서 SOTA

GPT - 2
Transformer Decoder block

Task Agnostic
: Zeroshot task transfer

언어를 정확히 학습했다면
Finetuning없이 Zeroshot 만으로
좋은 성능을 낼 수 있다 !
7가지 NLP Task에서 SOTA

01 Before T5

Overview

Transformer to T5 (20.5.25)

XLNet
BERT + GPT ≅ AE + AR

Task Agnostic

BERT 이후 큰 성능향상을 보인 첫 모델

1) Factorization order를 고려하여 양방향 학습
2) AR formula를 통해 BERT한계 극복

RoBERTa
Optimize BERT

Task Agnostic

가장 최적화된 BERT를 만들어보자!
(학습시간, batch, train data 증가)

MASS
BERT + GPT ≅ AE + AR

Task Agnostic

Encoder와 Decoder에 상반된 Masking

1) Decoder : Encoder에서 masking된 단어 예측
2) Encoder : Masking되지 않은 단어 깊은 이해
3) Encoder, Decoder의 joint training 장려

BART
BERT + GPT ≅ AE + AR

Task Agnostic

Encoder에 다양한 noising 추가한
Text generation task에서 SOTA 달성

MT-DNN
Based on BERT

Task Agnostic

Multitask learning을 통해
universal representation을 생성해보자!
Pretrain 단계에서 multitask learning 진행

T5
Encoder-Decoder Transformer

Task Agnostic

모든 NLP task를 통합할 수 있도록
Text-to-text 프레임워크를 사용하자 !

01 Before T5

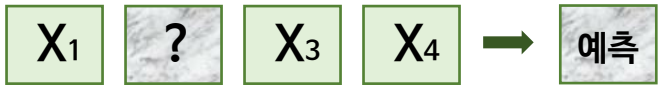
Overview

Transformer to T5 (20.5.25)
reference from XLNet

Pre-training의 대표적인 Objective

Auto Encoding

BERT는 Denoising AE라 볼 수 있음



Word sequence

$$\bar{\mathrm{x}} = [x_1, x_2, \cdots, x_T]$$

corrupted sequence

$$\hat{\mathrm{x}} = [x_1, [MASK], \cdots, x_T]$$

likelihood

$$p(\bar{\mathrm{x}}|\hat{\mathrm{x}}) \approx \prod_{t=1}^{T} p(x_t|\hat{\mathrm{x}})$$

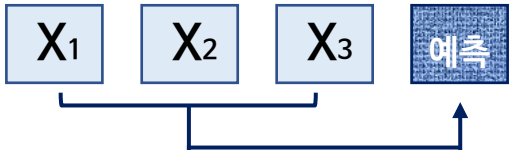
Objective function

$$Max_\theta \; logp_\theta(\bar{\mathrm{x}}|\hat{\mathrm{x}})$$

$$\approx \sum_{t=1}^{T} m_t logp_\theta(x_t|\hat{\mathrm{x}})$$

$$= \sum_{t=1}^{T} m_t log \frac{exp(H_\theta(\hat{\mathrm{x}})_t^T e(x_t))}{exp(H_\theta(\sum_{x'} exp(H_\theta(\hat{\mathrm{x}})_t^T e(x'))}$$

Auto Regressive



Word sequence

$$\mathrm{x} = [x_1, x_2, \cdots, x_T]$$

likelihood

$$p(\mathrm{x}) = \prod_{t=1}^{T} p(x_t|\mathrm{x}_{<t})$$

Objective function

$$Max_\theta \; logp_\theta(\mathrm{x})$$

$$= \sum_{t=1}^{T} logp_\theta(x_t|\mathrm{x}_{<t})$$

$$= \sum_{t=1}^{T} log \frac{exp(h_\theta(\mathrm{x}_{1:t-1})^T e(x_t))}{exp(h_\theta(\sum_{x'} exp(h_\theta(\mathrm{x}_{1:t-1})^T e(x'))}$$

01 Before T5

Question

Given the current landscape of transfer learning for NLP, what works best?
And how far can we push the tools we already have?

그래서 뭐가 좋은가 ?

01 Before T5

Question

Given the current landscape of transfer learning for NLP, what works best?
And how far can we push the tools we already have?

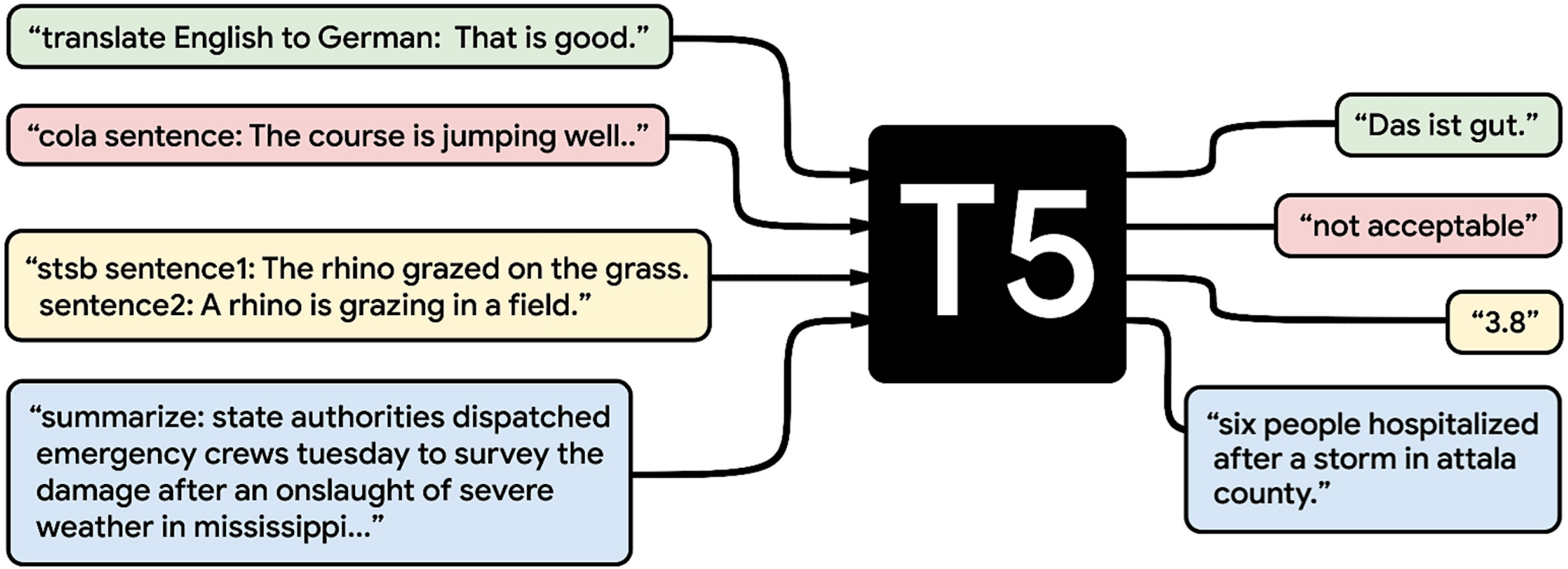
그래서 뭐가 좋은가 ?

T5의 다양한 실험이 대략적인 결론을 내려줌

02

T5 Model

Overview



“Unified framework that converts every language problem into a text-to-text format”

02 T5 Model

Overview

- T5 논문은 67페이지로 이루어진 논문
- 전체 논문리딩을 추천하나, 시간을 내기 어렵다면 10페이지까지 라도 읽어보길 추천함
- T5는 갑자기 등장한 방법론이 아님, 전체적인 아이디어는 기존 reference들을 바탕으로 함
  - T5의 각 아이디어가 어디에서 왔는지 공부해보는 것은 유의미한 작업

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer라는 굉장한 제목을 가진 T5 모델은 구글에서 제안된 모델로, 구글이 아니었다면 진행하지 못했을 다양한 실험 결과와 NLP Transfer learning을 위한 새로운 데이터셋인 "Colossal clean crawled corpus"를 공개하였다. 다른 논문들보다 더 Transfer learning에 대한 일반화 성능을 강조하는데, 이는 Unified framwork로 downstream task 학습을 진행하기 때문이라 생각한다.

T5 introduction을 읽어보면 T5가 decaNLP, GPT2, Unifying Question Answering, Text Classification, and Regression via Span Extraction 에서 영감을 받았다고 적혀있는데, 세 논문 모두 다 Mutitask learning 기반으로 nlp downstream task를 학습하는 방법을 제안한다. Multitask learning 모델 + 인풋을 GPT2처럼 구성한게 T5라고 이해하면 될 것 같다. 물론 기존 논문보다 더 많은 task를 더 높은 score로 풀어냈다는 것에 contribution이 있다.

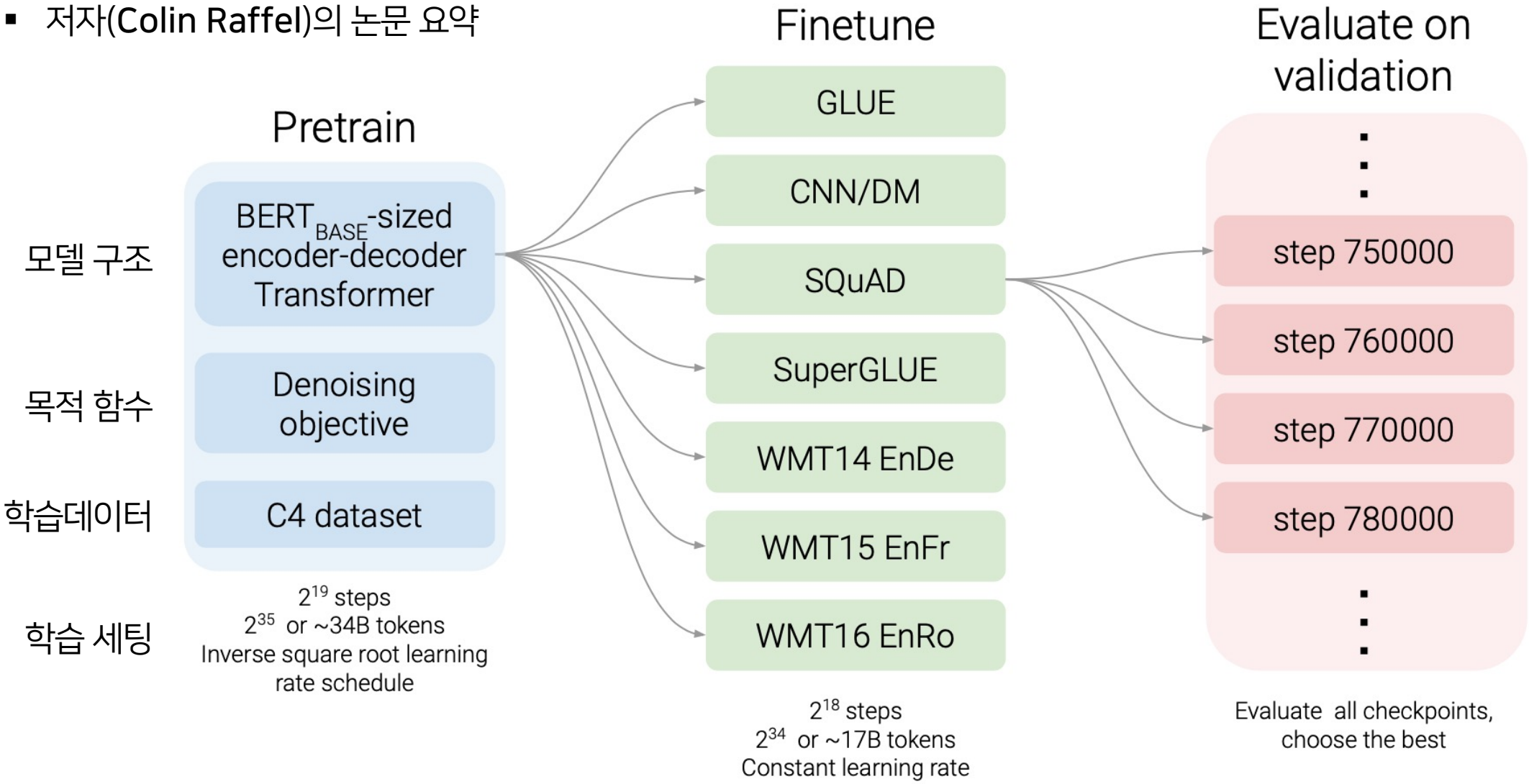
기존 모델들과 인풋을 받아들이는 형태가 다른것도 T5의 특징 중하나이다. task마저 text로 학습한다는 아이디어가 재미있다. 즉 text 형태로 주어진 문제에서 text 정답을 찾는 형식으로 전개되며, 이를 'Text to text' 라는 단어로 표현한다.

https://yukyunglee.github.io/Transformer-to-T5-4/ 😉

02 T5 Model

Overview

- 저자(**Colin Raffel**)의 논문 요약



02

T5 Model

Overview

(내가 생각하는) Points in T5

1. What is text to text ?
2. Transfer learning in NLP
3. Training objective : Modified MLM
4. Model structure
5. Corruption

02

T5 Model

What is text to text ?

text → T5 → text

Grammer Check Dataset (CoLA)

Original input
: Sentence: I am a great man.
Original target : 1

Processed input
: cola sentence: I am a great man.
Processed target : acceptable

Sentiment Dataset (SST2)

Original input
: Sentence: it confirms fincher ’s status as a film maker who artfully bends technical know-how to the service of psychological insight .
Original target: 1

Processed input
: sst2 sentence: it confirms fincher ’s status as a film maker who artfully bends technical know-how to the service of psychological insight .
Processed target: positive

: text 형태로 주어진 문제에서 text 정답 찾기

02 T5 Model

Transfer learning in NLP

Text → BERT style model → Classification / Span prediction

Encoder-only

Single prediction per input token
Singple prediction for an entire input sequence

Text → T5 → All NLP task

Encoder-Decoder structure

모든 NLP task에서 동일한 모델, loss, hyperparameter 사용가능

02 T5 Model

Transfer learning in NLP

1) Model Architecture

Encoder , Decoder only 모델 보다
Basic transformer 구조가 높은 성능을 보임

2) Pretraining Objectives

Pretraining에서 Noising 된 input을
Denoising하며 단어를 예측하는 방식이
가장 효율적인 방법임

3) Unlabeled datasets

Domain specific data는 task에 도움이 되지만
데이터의 크기가 작은경우 overfitting을 야기함

4) Training strategies

multitask learning이
unsupervised pre-training과 비슷한 성능 보임
학습시 task별 적절한 proportion이 필요함

5) Scaling

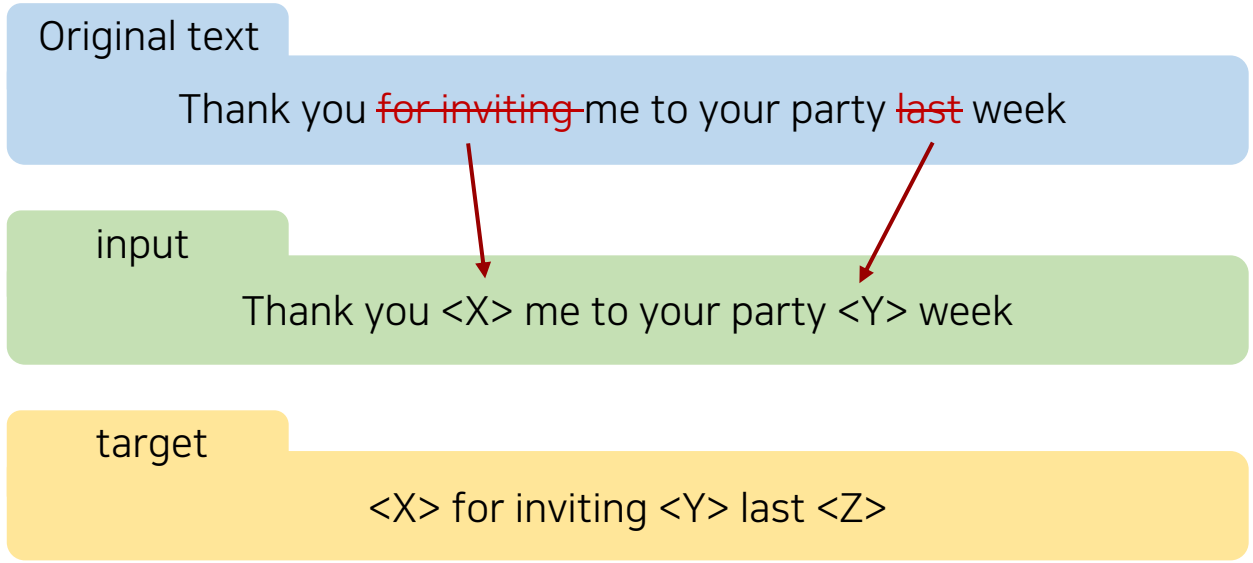
모델 크기를 늘리거나 ,앙상블을 시도하며 실험 진행.
작은모델을 큰 데이터로 학습하는게 효과적이라는것 발견함

6) Pushing the limits

110억개 파라미터를 가지는 모델을 훈련하여 SOTA 달성함
1 trillon 개가 넘는 token에 대해 훈련 진행함

02 T5 Model

Training objective : Modified MLM

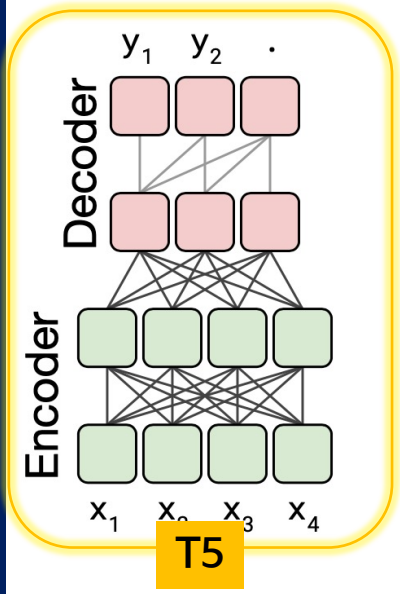


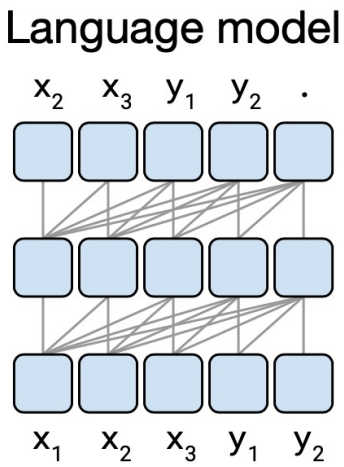
- MLM은 bidirectional model구조를 가짐
- BERT는 하나의 token에 masking을 하지만 T5 연속된 token을 하나의 mask로 바꿈 BART와 비슷
- Encoder-Decoder 구조로 input과 Target을 가지고 있음
- Input에서 mask 되지 않은 부분을 target에서 맞춰야 함 MASS와 비슷
- Output level에서 FFNN + Softmax을 통해 시퀀스 생성

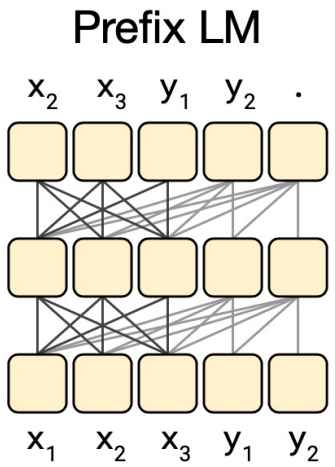
02 T5 Model

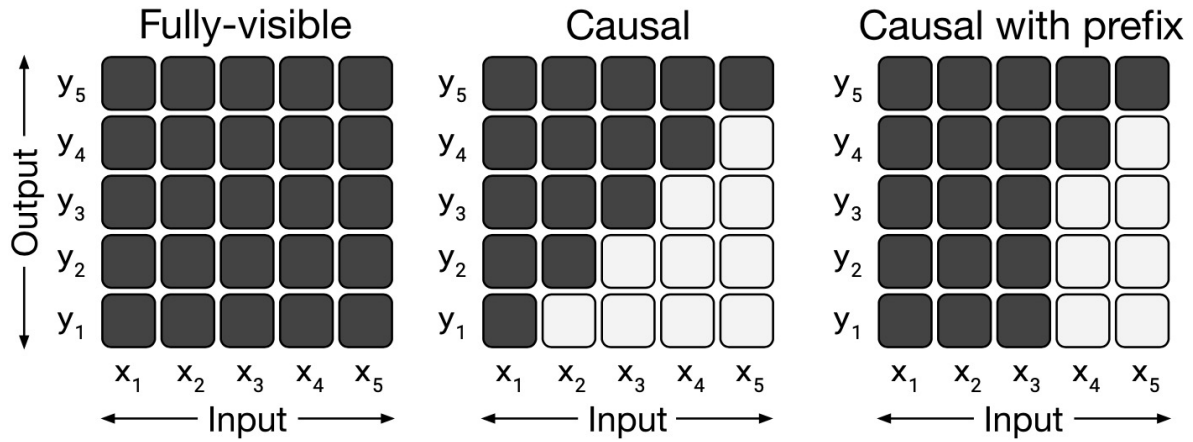
Model structure

이전 발표자료에서 잘못 설명된 부분









Model structure

1) Encoder - Decoder

Encoder : fed an input sequence
Decoder : produces a new output sequence

2) Language model

autoregressively produce an output sequence

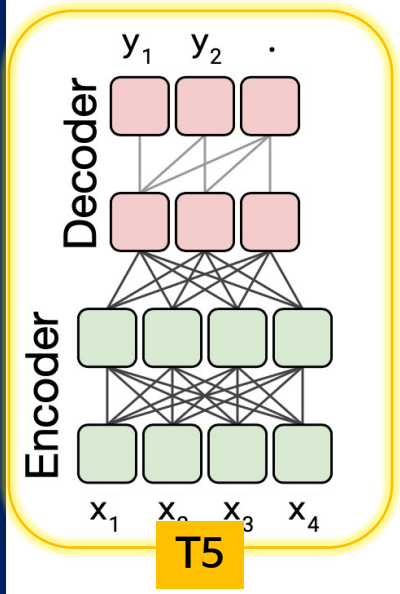
3) Prefix Language model

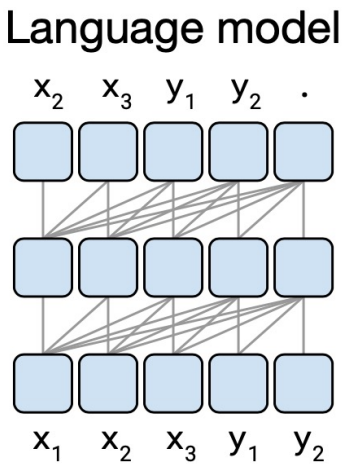
Source data : bidirectional attention
Generation part : unidirectional attention

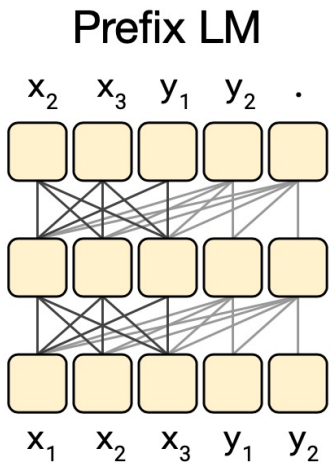
02 T5 Model

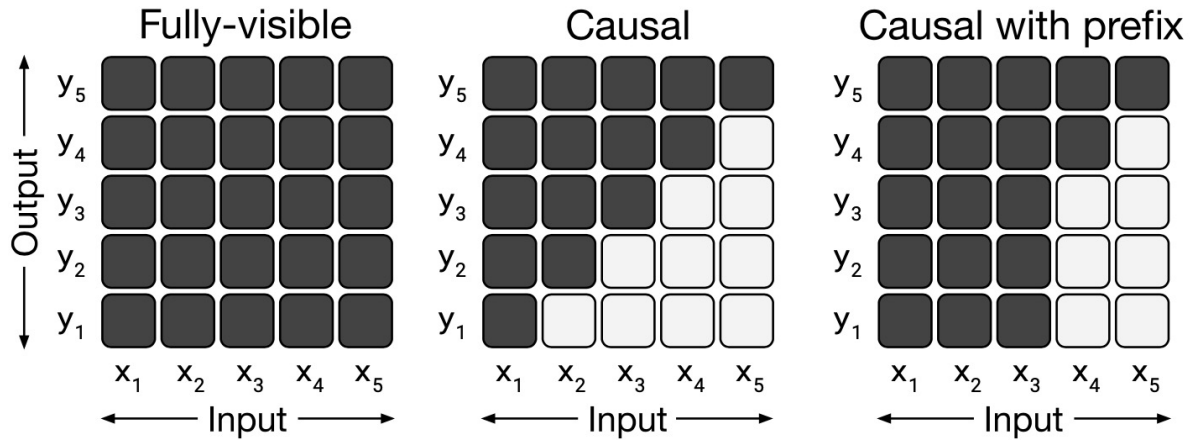
Model structure

이전 발표자료에서 잘못 설명된 부분







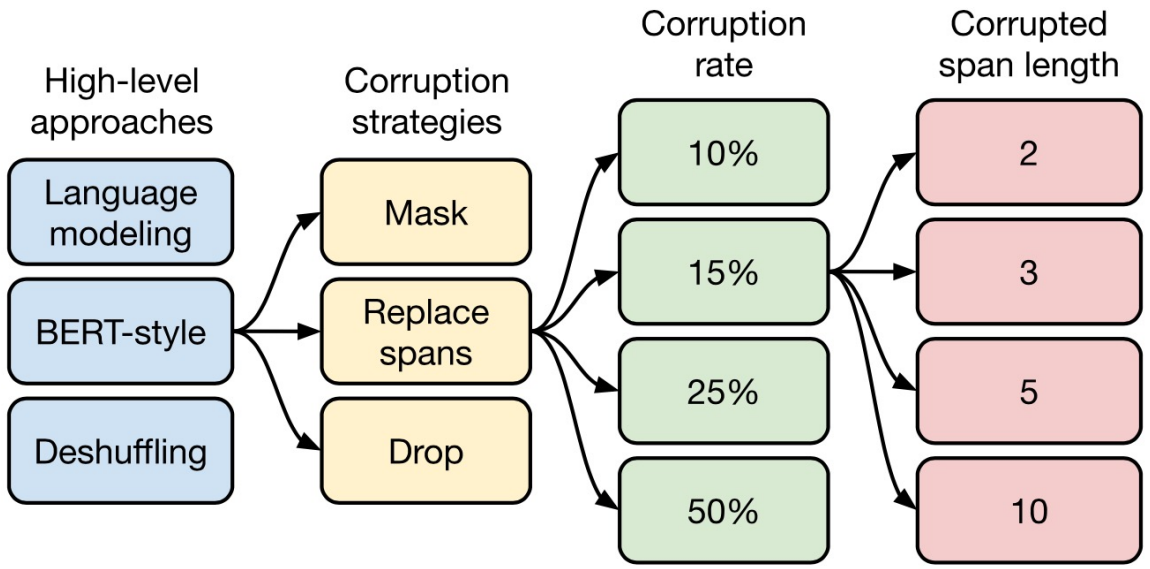


Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	Denoising	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	Denoising	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	Denoising	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	Denoising	P	M	81.82	18.61	78.94	68.11	26.43	37.98	27.39
Encoder-decoder	LM	$2P$	M	79.56	18.59	76.02	64.29	26.27	39.17	26.86
Enc-dec, shared	LM	P	M	79.60	18.13	76.35	63.50	26.62	39.17	27.05
Enc-dec, 6 layers	LM	P	$M/2$	78.67	18.26	75.32	64.06	26.13	38.42	26.89
Language model	LM	P	M	73.78	17.54	53.81	56.51	25.23	34.31	25.38
Prefix LM	LM	P	M	79.68	17.84	76.87	64.86	26.28	37.51	26.76

02 T5 Model

Corruption

이전 발표자료에서 잘못 설명된 부분



High-level approach

Objective	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Prefix language modeling	80.69	18.94	77.99	65.27	**26.86**	39.73	**27.49**
BERT-style [Devlin et al., 2018]	**82.96**	**19.17**	**80.65**	**69.85**	**26.78**	**40.03**	**27.41**
Deshuffling	73.17	18.59	67.61	58.47	26.11	39.30	25.62

Corruption rate

Corruption rate	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
10%	**82.82**	19.00	**80.38**	69.55	**26.87**	39.28	**27.44**
★ 15%	**83.28**	19.24	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
25%	**83.00**	**19.54**	**80.96**	70.48	**27.04**	**39.83**	**27.47**
50%	81.27	19.32	79.80	70.33	**27.01**	**39.90**	**27.49**

Corruption strategies

Objective	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
BERT-style [Devlin et al., 2018]	82.96	19.17	**80.65**	69.85	26.78	**40.03**	27.41
MASS-style [Song et al., 2019]	82.32	19.16	80.10	69.28	26.79	**39.89**	27.55
★ Replace corrupted spans	83.28	**19.24**	**80.88**	**71.36**	**26.98**	39.82	**27.65**
Drop corrupted tokens	**84.44**	**19.31**	**80.52**	68.67	**27.07**	39.76	**27.82**

Corruption span length

Span length	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline (i.i.d.)	**83.28**	19.24	80.88	71.36	**26.98**	**39.82**	**27.65**
2	**83.54**	19.39	**82.09**	**72.20**	**26.76**	**39.99**	**27.63**
3	**83.49**	**19.62**	**81.84**	**72.53**	**26.86**	39.65	**27.62**
5	**83.40**	19.24	**82.05**	**72.23**	**26.88**	39.40	**27.53**
10	82.85	19.33	**81.84**	70.44	**26.79**	39.49	**27.69**

02

T5 Model

Result

Encoder-decoder architecture

Architecture	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	P	M	81.82	18.61	78.94	68.11	26.43	37.98	27.39

Span prediction objective

Span length	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline (i.i.d.)	**83.28**	19.24	80.88	71.36	**26.98**	**39.82**	**27.65**
2	**83.54**	19.39	**82.09**	**72.20**	**26.76**	**39.99**	**27.63**
3	**83.49**	**19.62**	**81.84**	**72.53**	**26.86**	39.65	**27.62**
5	**83.40**	19.24	**82.05**	**72.23**	**26.88**	39.40	**27.53**
10	82.85	19.33	**81.84**	70.44	**26.79**	39.49	**27.69**

C4 dataset

Dataset	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**
WebText-like	17GB	**84.03**	**19.31**	**81.42**	71.40	**26.80**	**39.74**	**27.59**
Wikipedia	16GB	81.85	**19.31**	81.29	68.01	**26.94**	39.69	**27.67**
Wikipedia + TBC	20GB	83.65	**19.28**	**82.08**	**73.24**	**26.77**	39.63	**27.57**

Multi-task pre-training

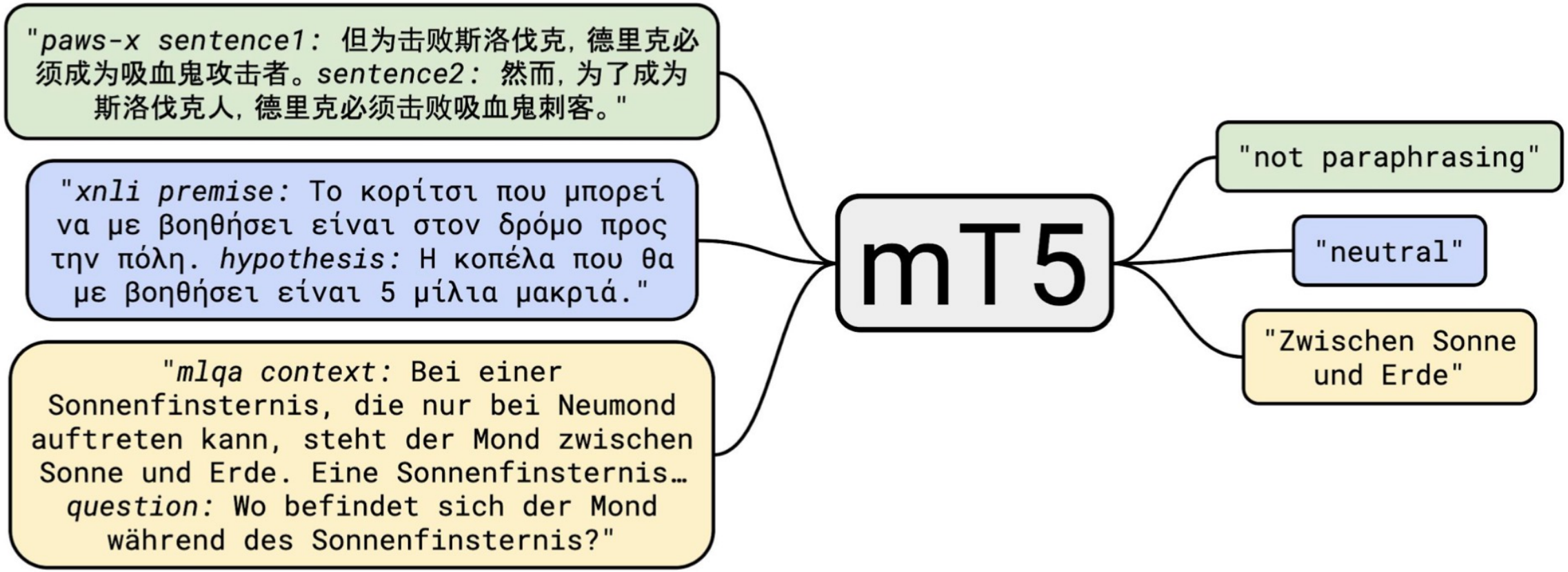
Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65
Multi-task training	81.42	**19.24**	79.78	67.30	25.21	36.30	27.76
Multi-task pre-training + fine-tuning	**83.11**	**19.12**	**80.26**	**71.03**	**27.08**	39.80	**28.07**
Leave-one-out multi-task training	81.98	19.05	79.97	**71.68**	**26.93**	39.79	**27.87**
Supervised multi-task pre-training	79.93	18.96	77.38	65.36	26.81	**40.13**	**28.04**

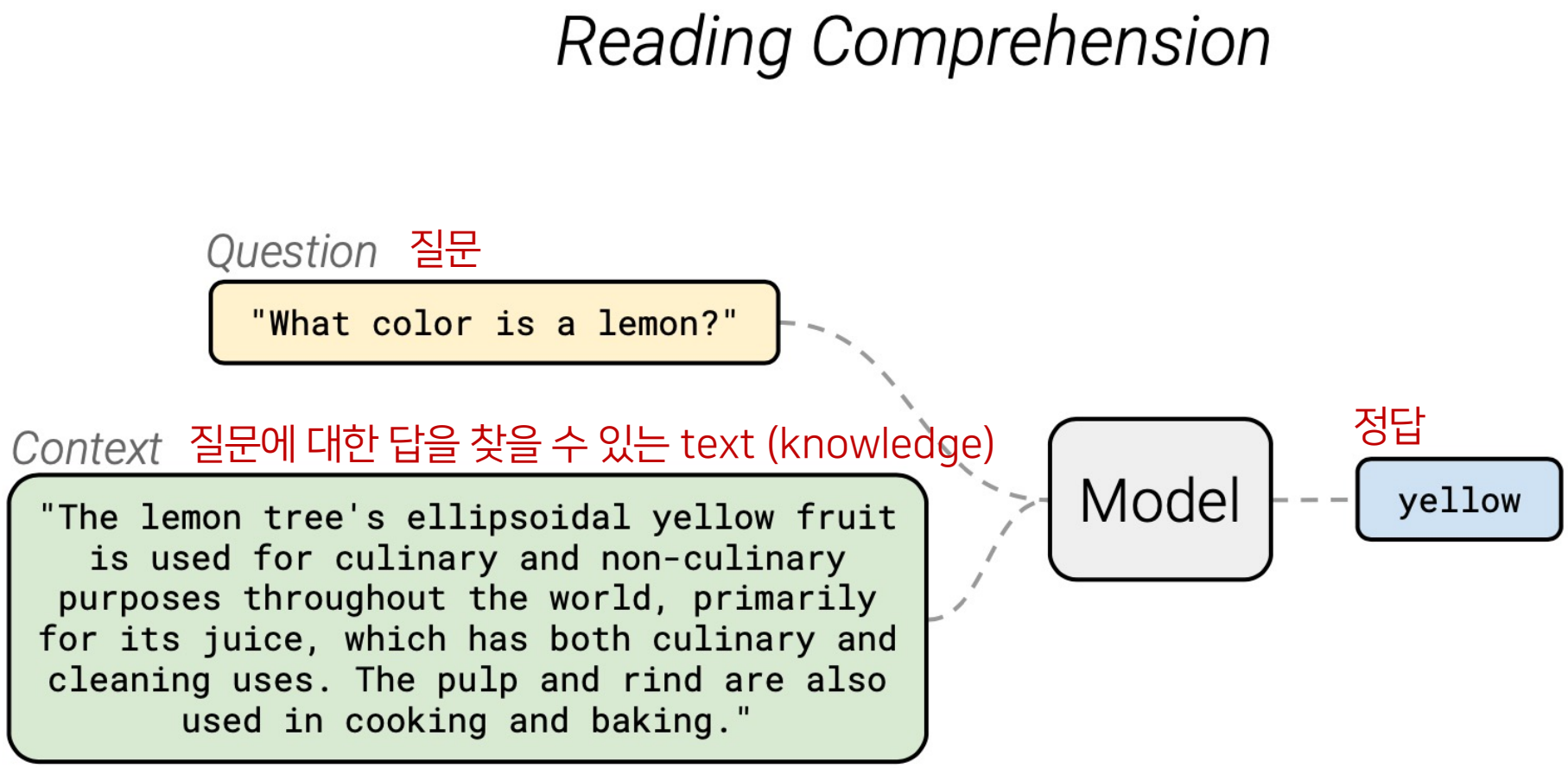
Bigger models trained longer

Scaling strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Baseline	83.28	19.24	80.88	71.36	26.98	39.82	27.65
1× size, 4× training steps	85.33	19.33	82.45	74.72	27.08	40.66	27.93
1× size, 4× batch size	84.60	19.42	82.52	74.64	27.07	40.60	27.84
2× size, 2× training steps	**86.18**	19.66	**84.18**	77.18	27.52	**41.03**	28.19
4× size, 1× training steps	**85.91**	19.73	**83.86**	**78.04**	27.47	40.71	28.10
4× ensembled	84.77	**20.10**	83.09	71.74	**28.05**	40.53	**28.57**
4× ensembled, fine-tune only	84.05	19.57	82.36	71.55	27.55	40.22	28.09

03 Other topics

multilingual T5

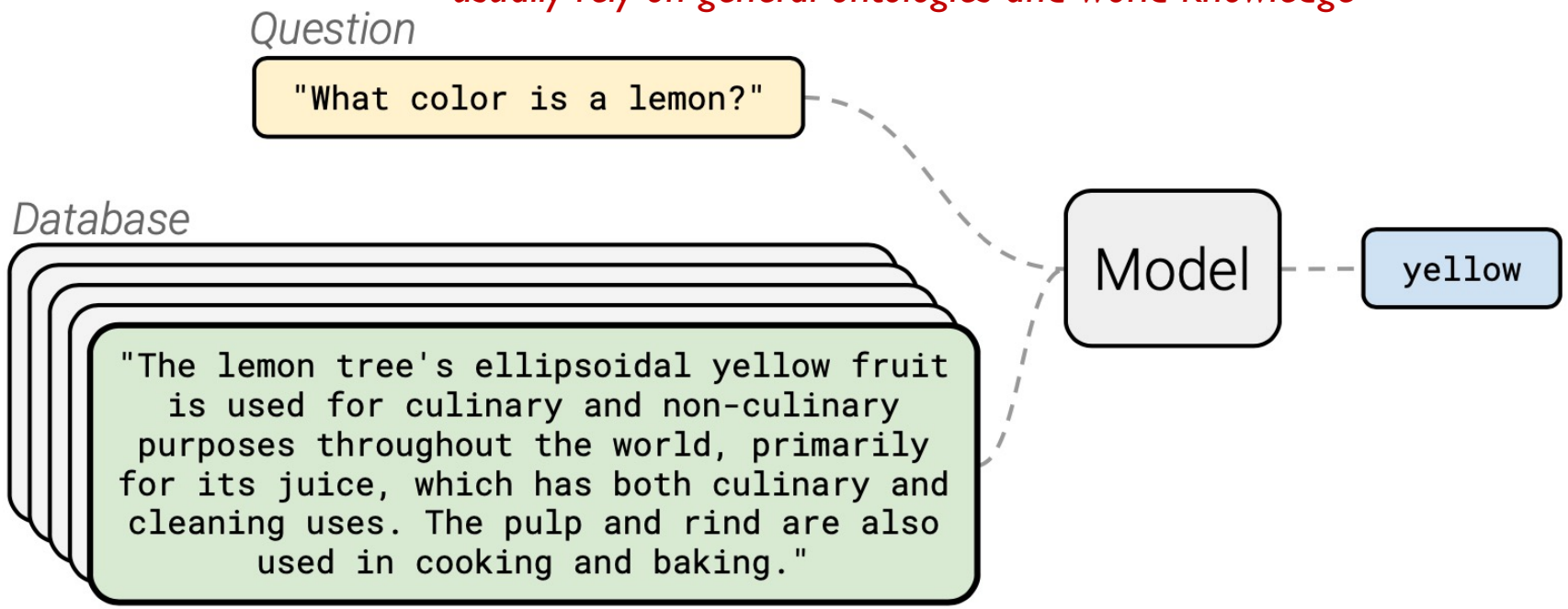



Reading Comprehension
Question 질문
"What color is a lemon?"
Context 질문에 대한 답을 찾을 수 있는 text (knowledge)
"The lemon tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses. The pulp and rind are also used in cooking and baking."
Model
정답
yellow

Open-Domain Question Answering

https://github.com/danqi/acl2020-openqa-tutorial

***Open domain**: deal with questions about nearly anything, usually rely on general ontologies and world knowledge*



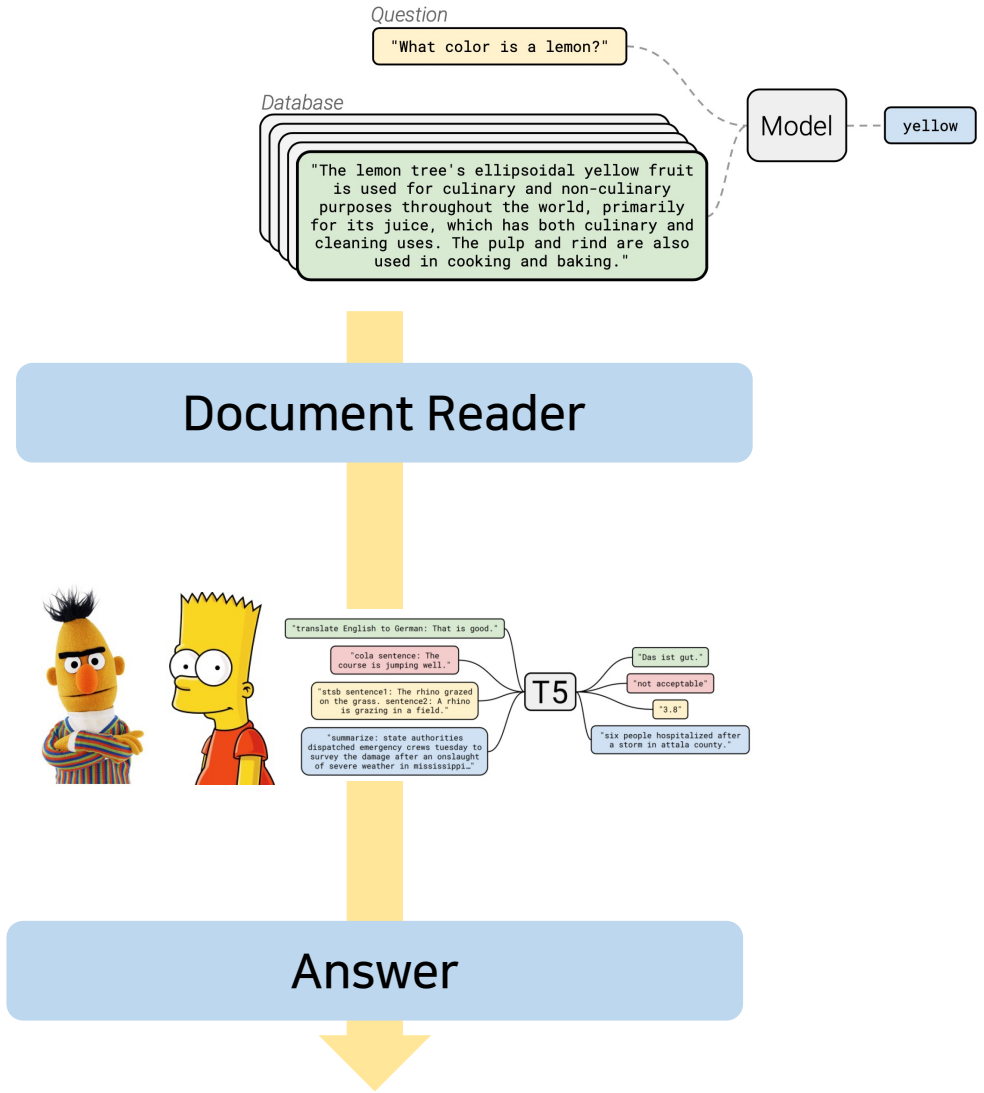
질문에 대한 답변을 찾아낼 수 있는 database를 구축하고,
database에서 정답을 찾아냄

03 Other topics

T5 & knowledges

- Two stage : Retriever-Reader approach

질문의 답을 포함할 가능성이 높은 document 탐색



Reader를 통해 적절한 답을 찾아냄

* 해당 페이지는 출처를 참고하여 재 구성하였습니다.

출처 : ACL2020-openqa-tutorial

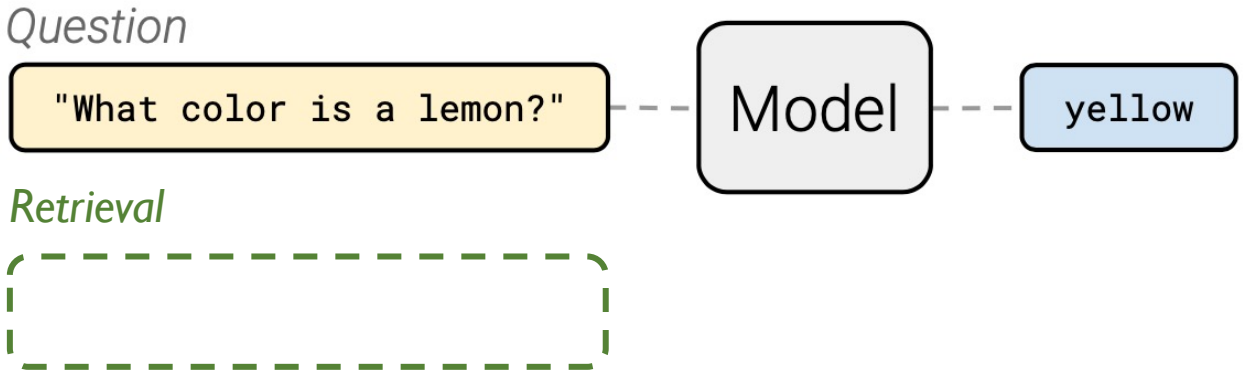
03 Other topics

T5 & knowledges

[참고]
Salient Span Masking에 관심있다면
REALM PYSR영상 참고 link

- T5를 사용하면 .. ?

Closed-Book Question Answering

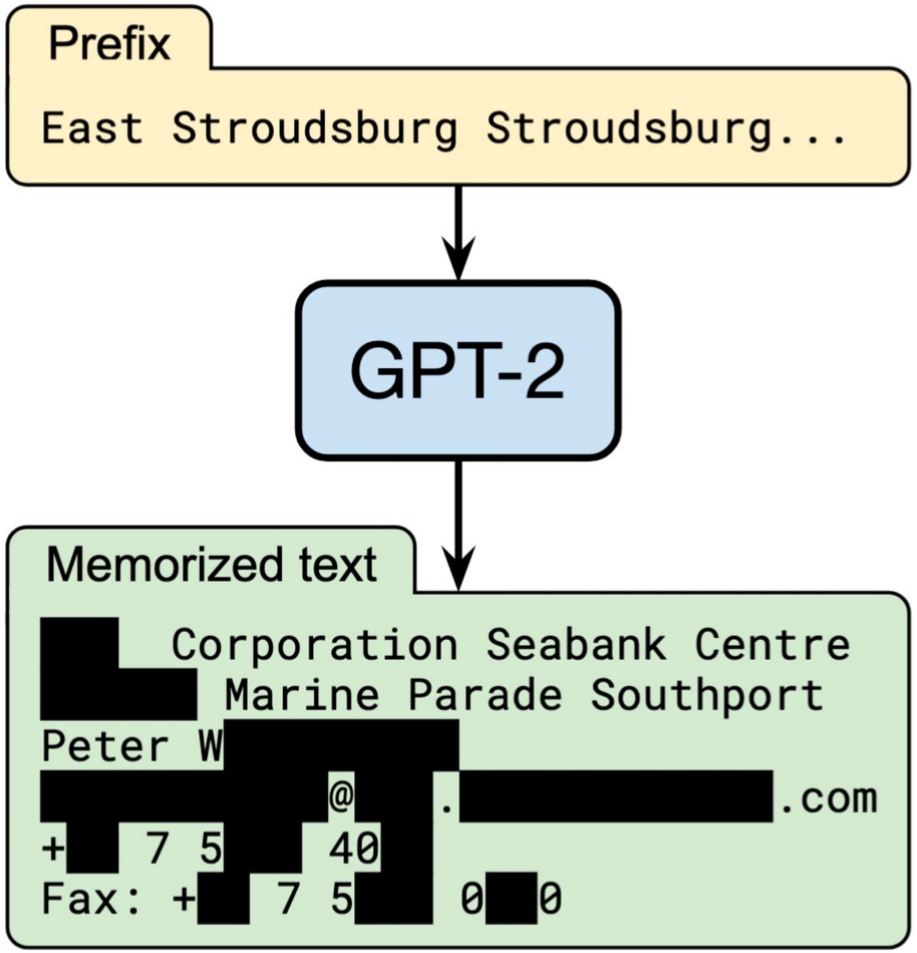


T5는 사전학습시 대량의 데이터로 학습이 되어있으므로,
Reference text를 찾는 과정인 Retrieval step 없이도 정답 찾아낼 수 있음

03 Other topics

memorize information knowledge ?

- Do large LMs memorize their data ?



- 대량의 web 기반 데이터로 학습하게 될 경우
  여러가지 문제가 발생할 수 있음
- 개인정보가 training dataset에 포함되어있고,
  제대로 block되지 않을 경우
  Large LM decoder가 generation 할 수 있음
- 굉장히 중요한 이슈 중 하나

03 Other topics

Transformer variants ?

- Standard transformer는 각 모듈별로 다양한 형태로 발전 가능함

🤗 Transformer_Survey_Paper_Study 🤗

"A survey of Transformer" paper study @DSBA Lab

📄 Paper : Lin, Tianyang, et al. "A Survey of Transformers." *arXiv preprint arXiv:2106.04554* (2021) [Link]

자세한 내용은 TSS study 자료 참고

https://github.com/yukyunglee/Transformer_Survey_Study

감사합니다