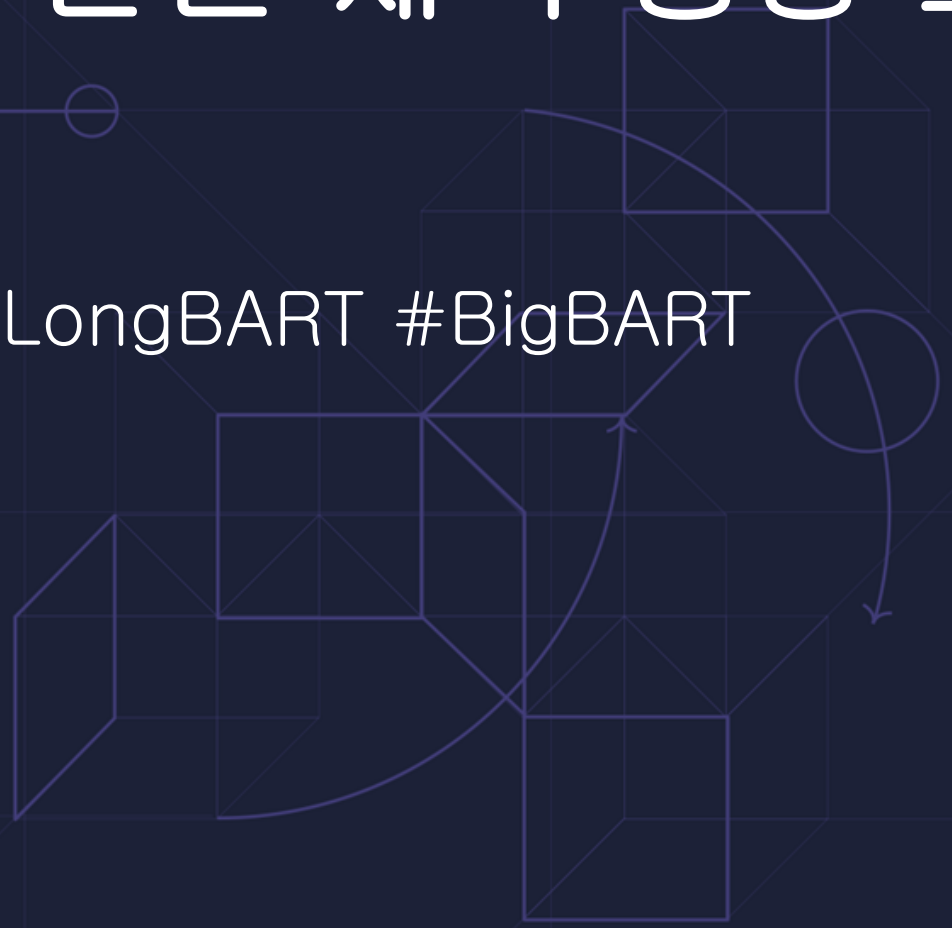
제목 짓기 어려워서 만든 제목 생성 모델

#비전공자의반란 #생성요약 #LongBART #BigBART

NLP-02
(메타몽팀)



boostcamp ai tech

목차

1. Introduction

#team 소개
#프로젝트 협업
#주제_선정_배경
#시스템 아키텍처
#시연영상

Introduction
Team 소개

데이터 수집
모델 개발 및 튜닝
Text infilling
Teacher forcing
scheduler

고창용_T2008
#PM(Main)+코드검사 담당

R-drop
시각화
Serving

이예빈_T2165
#PM(Sub) 담당

모델 개발 및 튜닝
Serving 프로토타입 제작

이기성_T2149
#코드검사 담당

Baseline 구축
데이터 수집
모델 개발 및 튜닝

정유석_T2204
#리서치 담당

모델 개발
모델 경량화
ElasticSearch

안명철_T2126
#리서치 담당

Baseline 구축
한국어 평가지표 개발
시각화
데이터 후처리

박범진_T2081
#기록 담당

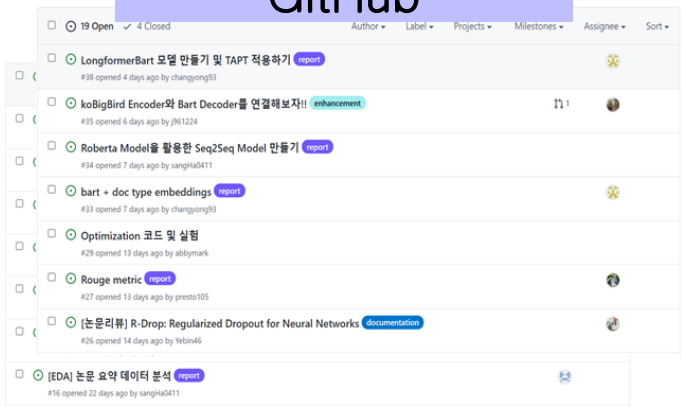
데이터 전처리
모델 개발
Text infilling
Teacher forcing
scheduler

박상하_T2084
#기록 담당

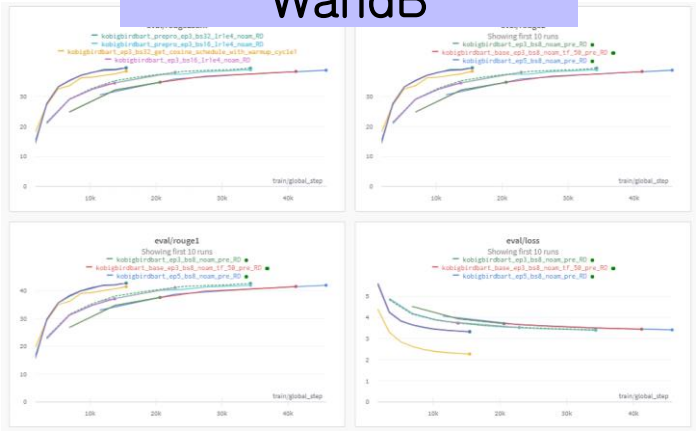
Introduction

프로젝트 협업

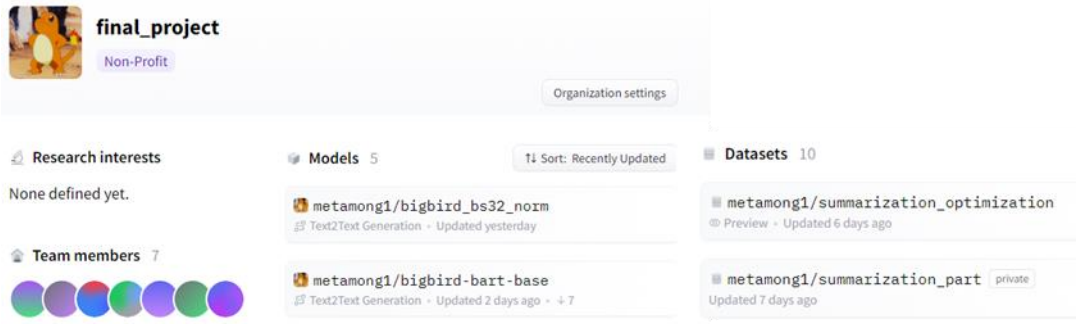
GitHub



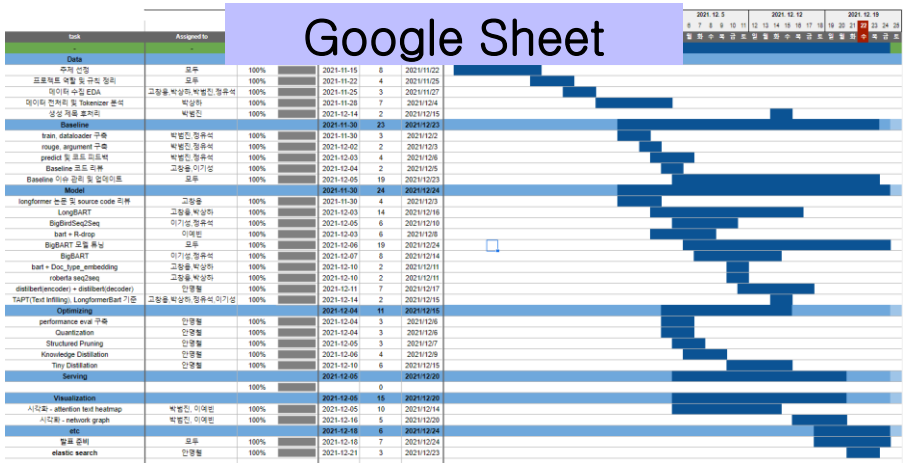
WandB



HuggingFace Hub



Google Sheet



다양한 툴을 활용하여 협업

논문, 보고서… 제목 짓기 너무 어렵지 않나요?

□

……?

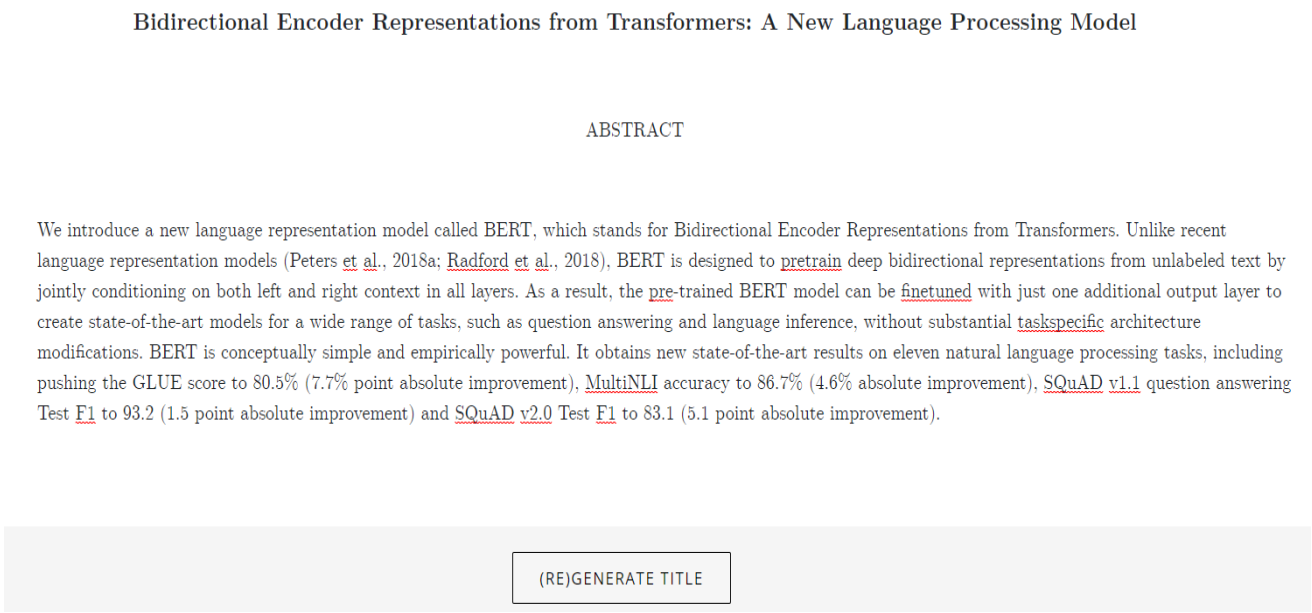
Introduction
주제 선정 배경

“생성 요약을 활용해 제목을 만들어주는 모델 개발”

- 논문, 보고서… 제목 짓기 너무 어렵지 않나요?

영어 제목 생성기는 있는데 한국어는 없네?

- 우리가 만들자!



Introduction
주제 선정 배경

1 다양한 시도를 통해
우리의 성장을
가장 잘 보여줄 수 있는가?



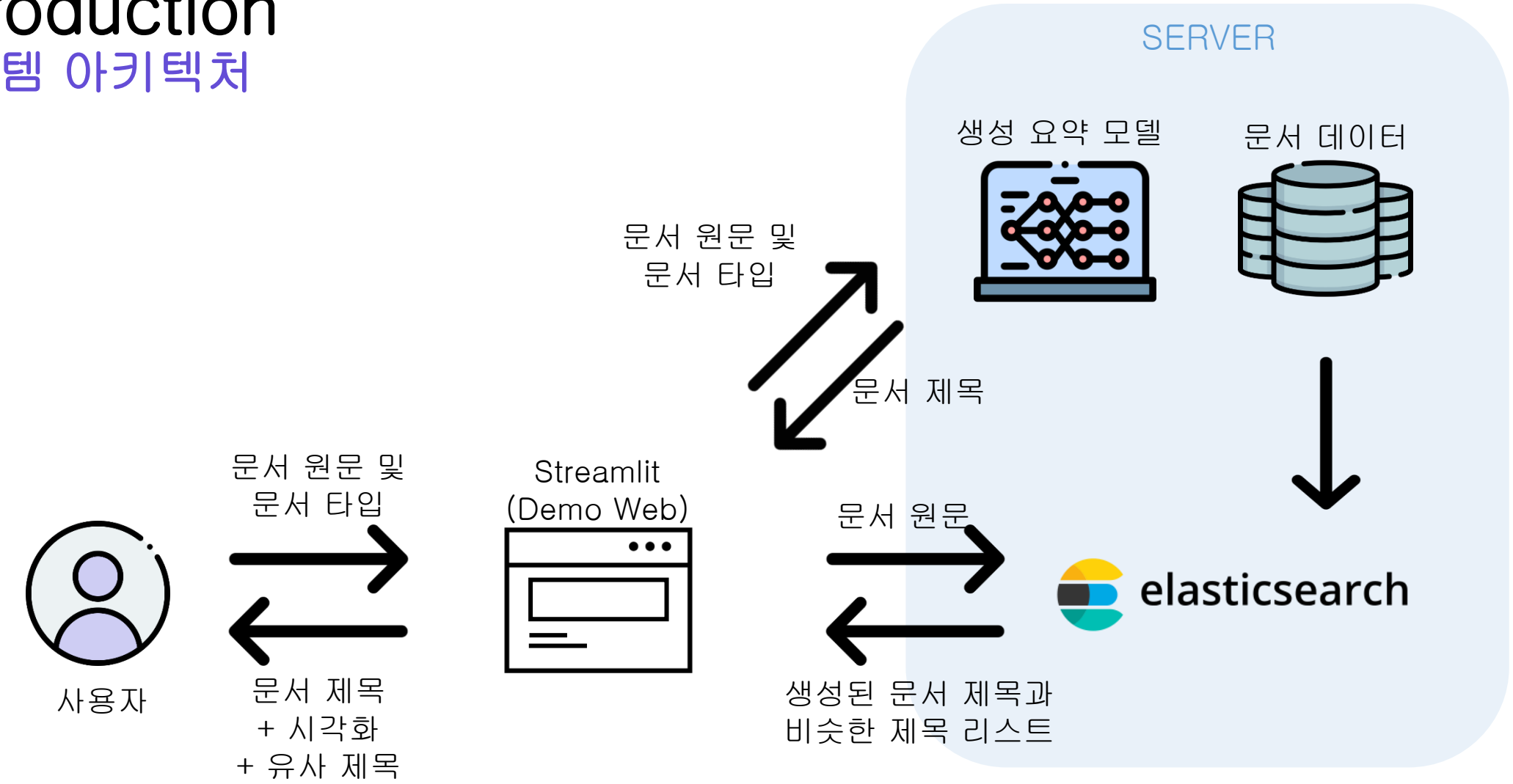
부스트캠프에서 배운 내용들을 모두 녹아낼 수 있는 프로젝트

배운 내용보다 더 나아갈 수 있는 프로젝트

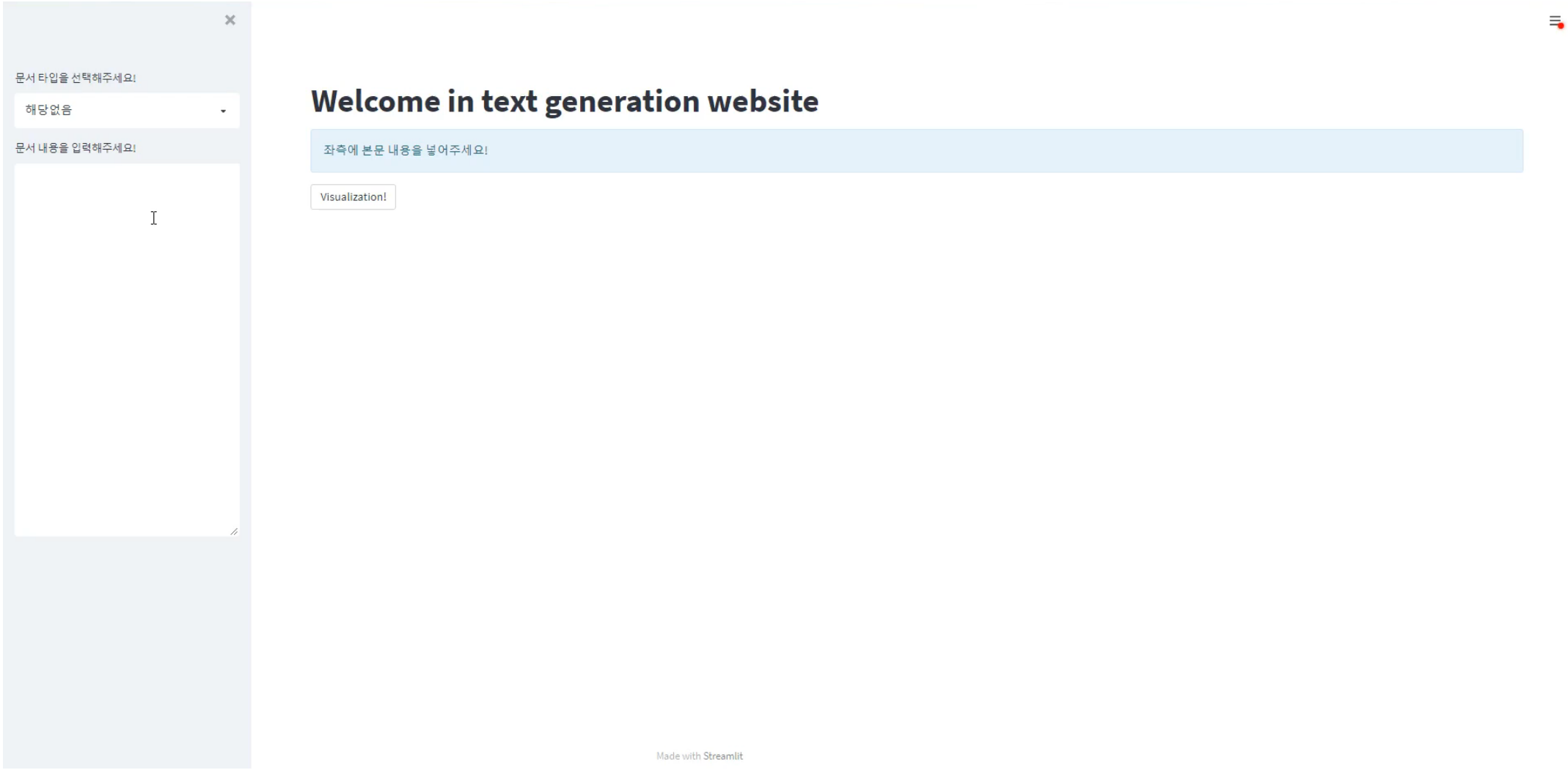
2 주어진 환경에서
완성할 수 있는가?


개발자로서 주어진 리소스 내에서
최선의 결과물을 만들 수 있는 프로젝트

Introduction
시스템 아키텍처



Introduction
시연 영상

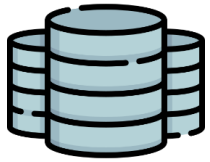


2. 학습 파이프라인

#Model Pipeline
#EDA
#Pre/Post-Processing
#Model
#평가 방법

학습 파이프라인

Model Pipeline

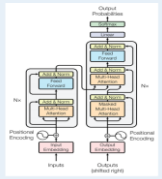


Train 50%

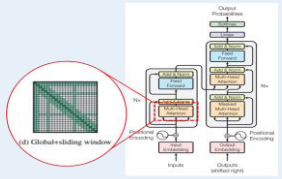
Preprocessing

Model

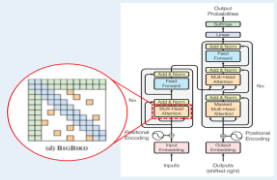
+document type embedding



BART



LongBART



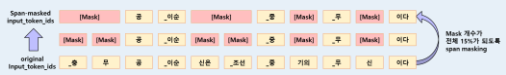
BigBART

Yes

doc type embedding

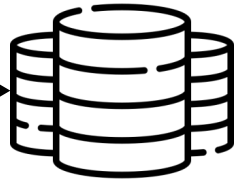
No

Pretraining / TAPT



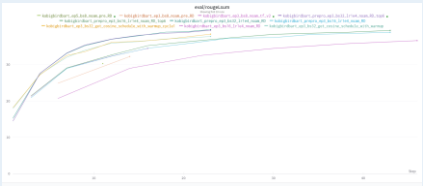
Finetuning

Noam Scheduler with warmup
Linear Scheduler with Warmup
Regularized Dropout
Teacher Forcing Scheduler
Hyperparameter
Rouge Score specialized in Korean.

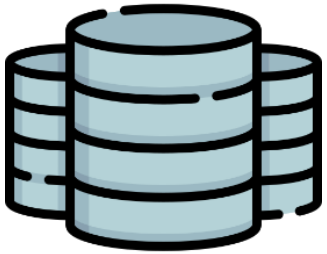


Eval 50%

Selecting Models



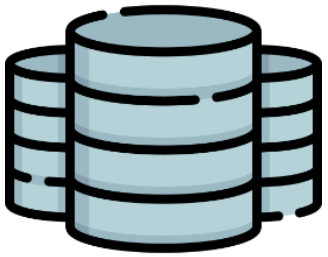
Train 50%로 학습한 모델 중 성능이 좋은 후보 모델을 전체 데이터로 재학습



Train 100%

학습 파이프라인

Model Pipeline

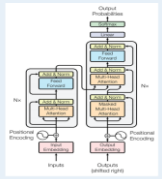


Train 100%

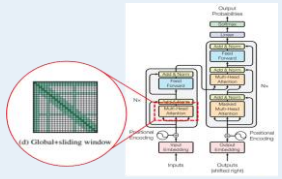
Preprocessing

Model

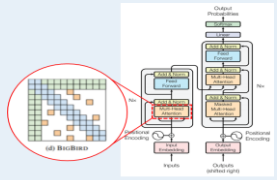
+document type embedding



BART



LongBART



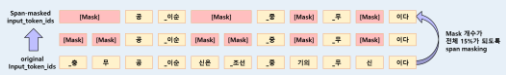
BigBART

Yes

No

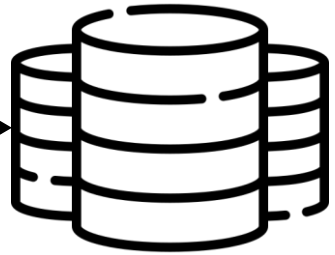
doc type embedding

Pretraining / TAPT

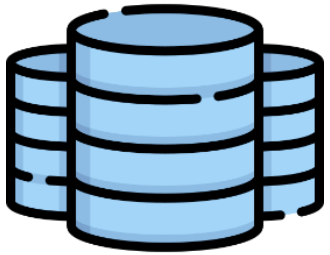


Finetuning

Noam Scheduler with warmup
Linear Scheduler with Warmup
Regularized Dropout
Teacher Forcing Scheduler
Hyperparameter
Rouge Score specialized in Korean.



Eval 100%



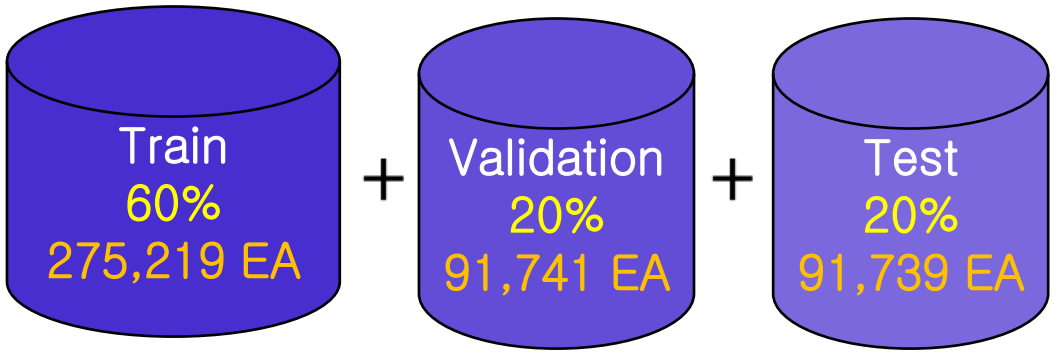
Test 100%

최종 모델 선정

Optimization

(optional)
Quantization
Knowledge Distillation

학습 파이프라인
데이터 분석(EDA)



데이터 본문 문체 및 특성 분석 (max length : 1887)

1. 논문 ➔ 목적, 증거
   본 연구의 목적은 청소년의 스마트폰 중독에 대한 학교부적응의 영향과 이에 대한 우울의 매개효과를 검증하는데 있다. 첫째, ~~~~. 둘째, ~~~~.

2. 뉴스 ➔ 발언 인용
   "재판관 관심사항과 논리를 보강한 보충서면을 준비하고 대법원 현장검증 시 대응에 철저를 기해 당진땅을 꼭 수호하자"고 강조했다.

3. 사설 잡지 ➔ 주장, 의견 제시
   핵담판이 난기류에 빠진 만큼 우리 정부는 북·미의 움직임을 예의 주시하면서 한·미 간 대북 공조를 강화해야 할 것이다.

데이터 제목 문체 및 특성 분석 (max length : 128)

1. 논문 ➔ 연구, 분석
   불변시장점유율모형과 BCG 매트릭스를 이용한 항만의 철강 수출변동 분석

2. 뉴스 ➔ 지역, 단체
   GS홈쇼핑, `벤처투자`에 사활 걸었지만…`지분법손실` 확대

3. 사설 잡지 ➔ 주장, 의견 제시
   개각으로 들썩인 공직사회, 이젠 개혁에 전념하라
   한국판 라인-야후재팬 빅뱅' 막는 지주사 규제, 완화해야

학습 파이프라인
Post/PreProcessing

Pre-processing

Model이 이해하지 못하는 문자를 최대한 제거

- 괄호 안에 있는 한자, 영어로 된 부연 설명 제거
- 한자, 한글, 영어 및 구두점을 제외한 특수 문자 제거

도애 홍석모의 금강산 유기, 『간관록』 일고 (陶厓 洪錫謨의 금강산 유기, 『艮觀錄』 一考) → 도애 홍석모의 금강산 유기, 간관록 일고

Post-processing

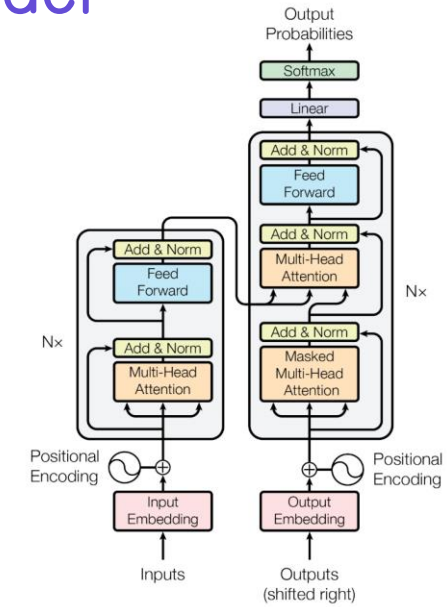
마무리가 제대로 지어지지 않은 부분 제거

- 쌍을 이루지 못한 문장기호와 마무리 되지 못한 내용 제거: (), [], {}, ‘ ’, “ “ 등
- 개행문자, 문장 끝 특수문자 제거

물도 껍질이 있는 걸까? - [물에 관한 → 물도 껍질이 있는 걸까?

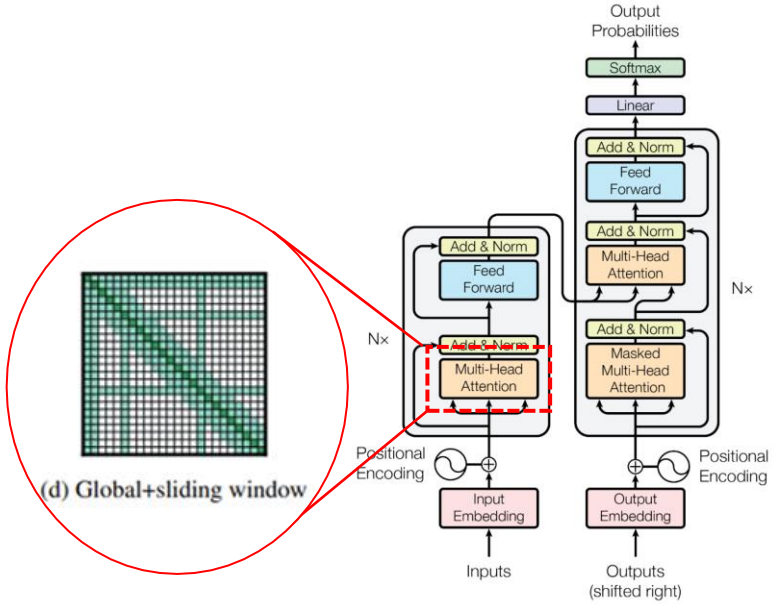
학습 파이프라인

Model



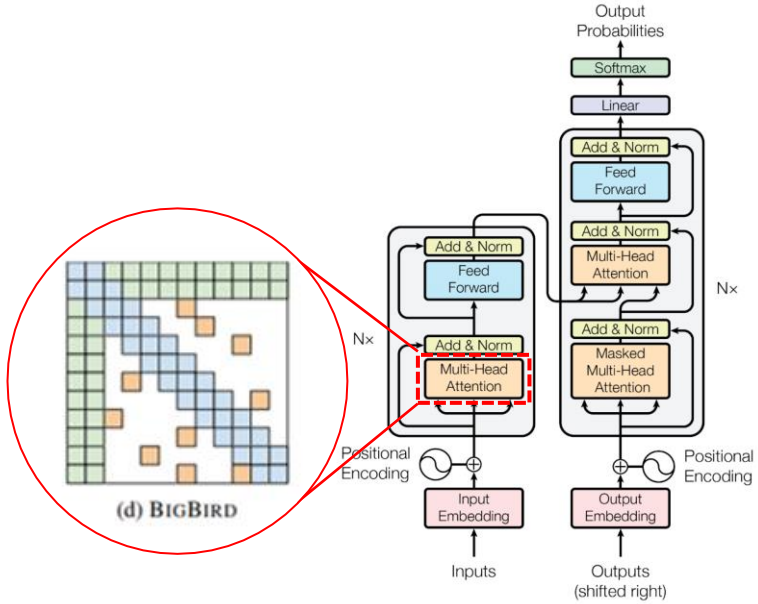
BART

CNN/Daily Mail Dataset
Vanilla 모델로 8위



LongBART

Longformer + BART



BigBART

BigBird + BART

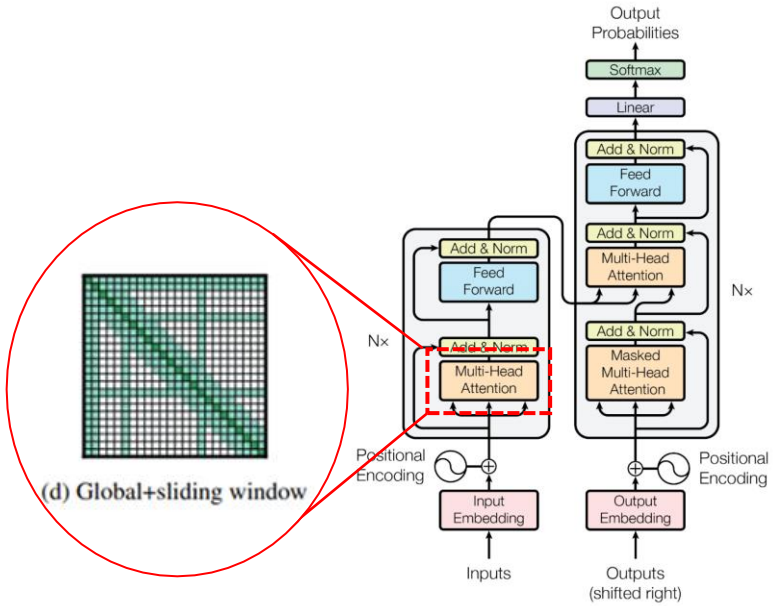
8	**BART**	44.16	21.28	40.90
9	**ERNIE-GENLARGE**	44.02	21.17	41.26
10	**T5**	43.52	21.55	40.69

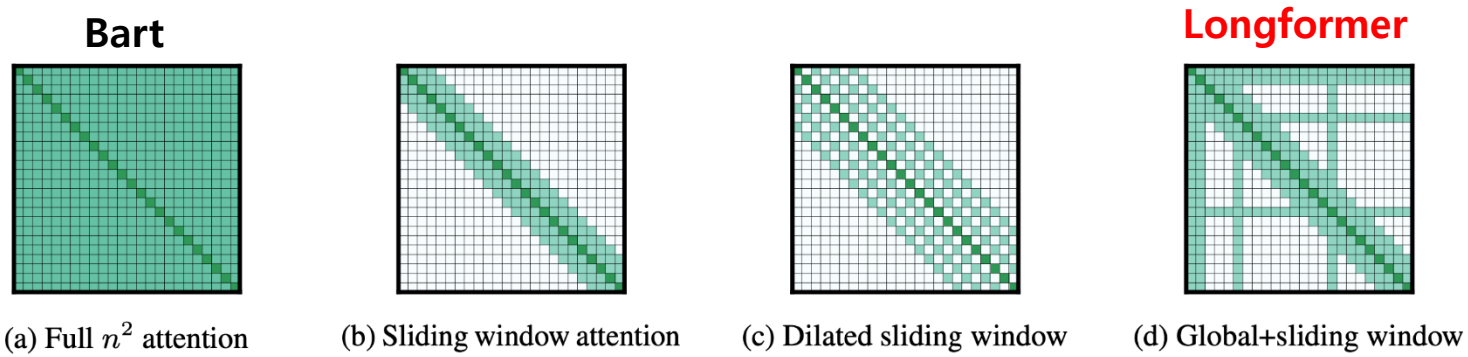
BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

학습 파이프라인

Model – LongBART (Longformer + BART)

- 모델 선정 이유
  - Time Complexity(Self Attention)
    - Bart ➔ O(N^2) ,Longformer ➔ O (N)
  - Long Sequence를 효율적으로 학습할 수 있는 모델
- 모델 구현 과정 중 제한 사항
  - Longformer 모델은 한국어 데이터로 pretrained 된 모델이 없음
  - Pretraining 과정이 필요

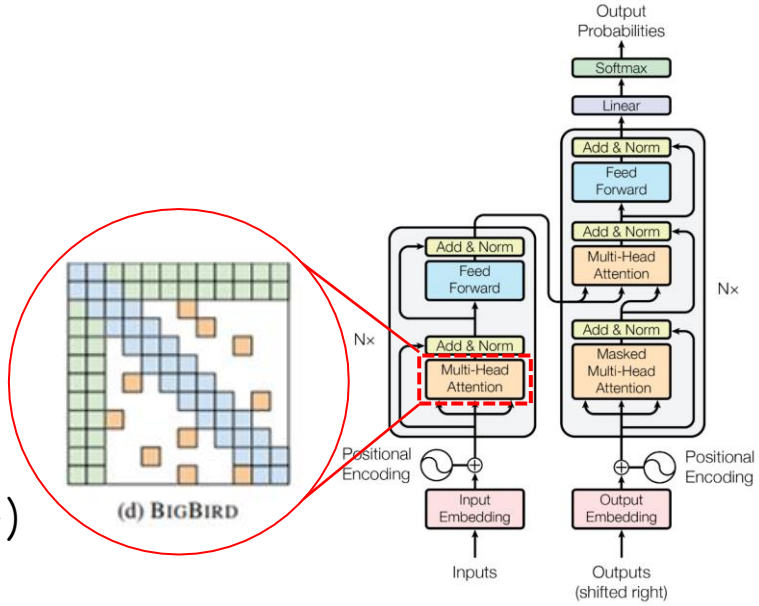


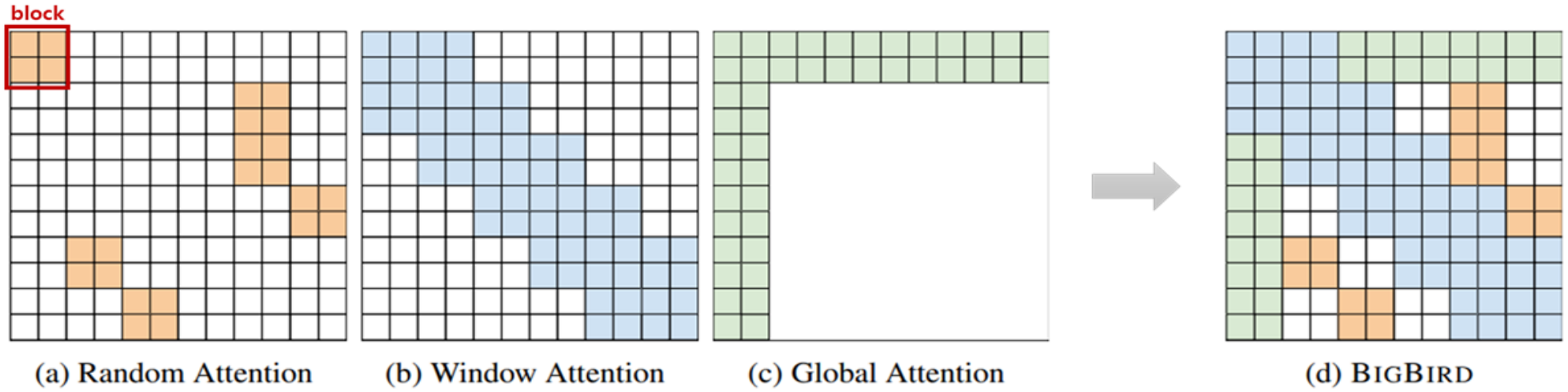


학습 파이프라인

Model – BigBART (Bigbird + BART)

- 모델 선정 이유
  - 한국어 사전 학습 모델이 존재
  - Sparse Attention
    - Longformer ➔ sliding, global attention
    - BigBird ➔ sliding, global, random attention (full attention에 근사)
  - Block 단위의 sparse attention 사용 ➔ 연산 최적화





학습 파이프라인

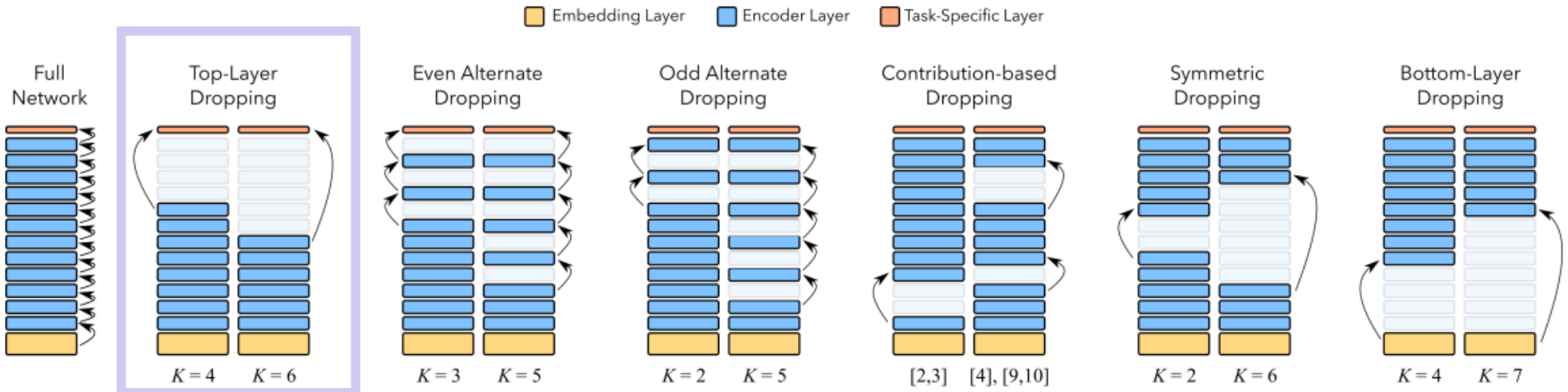
Model – BigBART (Bigbird + BART)

두 사전 학습 모델의 Word embedding 공유

Vocab size of word embedding

- Encoder(BigBird) : 32500 ← shared
- Decoder(Bart) : 30000

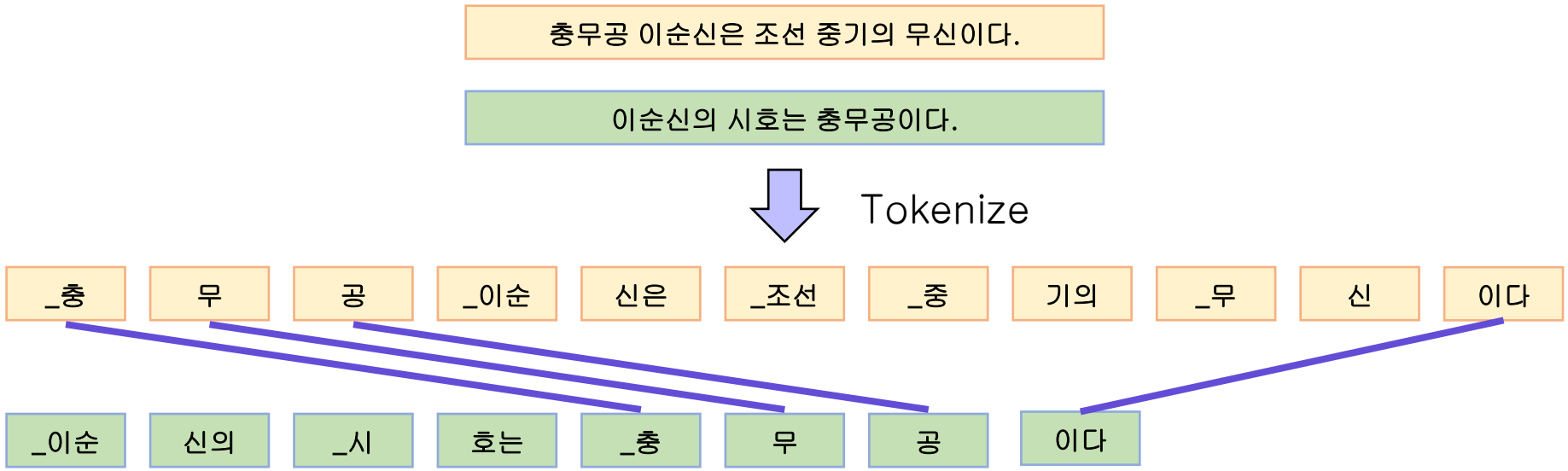
모델 사이즈 축소(파라미터 1억9천개 → 1억5천개)



학습 파이프라인

평가 방법: Rouge-L

Rouge-L: Longest common subsequence(LCS), 단일 문장 비교

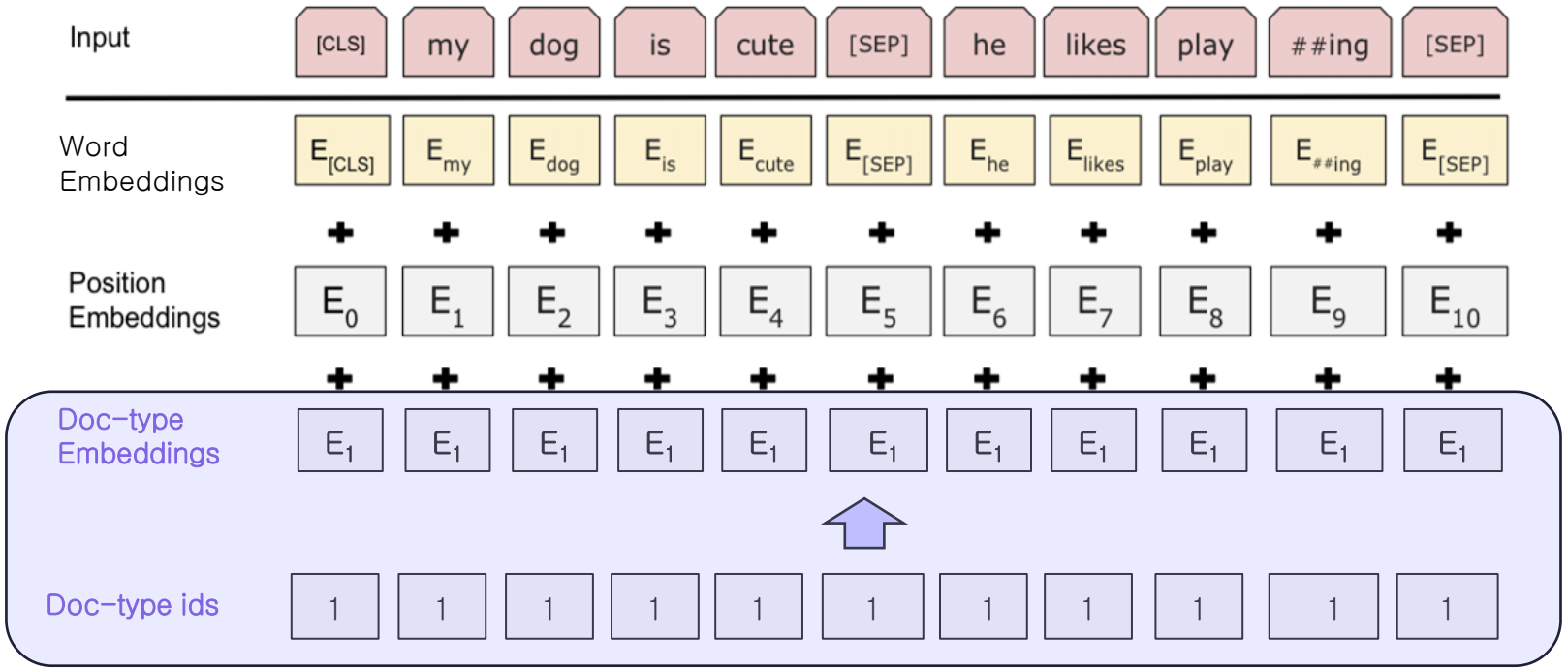


Tokenizer의 한계를 인식하고 Mecab POS(part of speech) tagging,
명사 등 usable token만 평가에 사용하는 방식으로 업데이트

3. 성능 개선

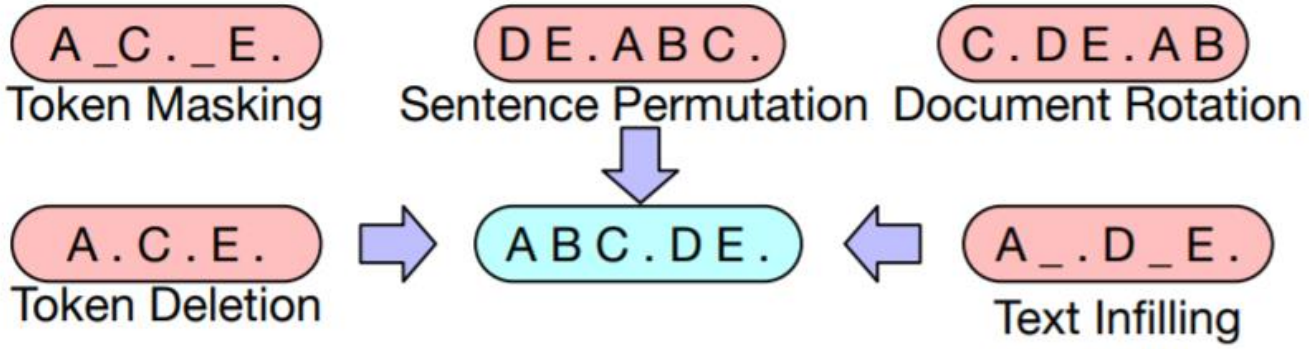
#Document_type
#Text_infilling
#R-drop
#Teacher_forcing
#결과

성능 개선
Document Type Embeddings



Document type Embedding + pretraining/Task-adaptive pretraining(Text Infilling)

성능 개선
Text Infilling



Model	SQuAD 1.1 F1	MNLI Acc	ELI5 PPL	XSum PPL	ConvAI2 PPL	CNN/DM PPL
BERT Base (Devlin et al., 2019)	88.5	**84.3**	-	-	-	-
Masked Language Model	90.0	83.5	24.77	7.87	12.59	7.06
Masked Seq2seq	87.0	82.1	23.40	6.80	11.43	6.19
Language Model	76.7	80.1	**21.40**	7.00	11.51	6.56
Permuted Language Model	89.1	83.7	24.03	7.69	12.23	6.96
Multitask Masked Language Model	89.2	82.4	23.73	7.50	12.39	6.74
BART Base						
w/ Token Masking	90.4	84.1	25.05	7.08	11.73	6.10
w/ Token Deletion	90.4	84.1	24.61	6.90	11.46	5.87
w/ Text Infilling	**90.8**	84.0	24.26	**6.61**	**11.05**	5.83
w/ Document Rotation	77.2	75.3	53.69	17.14	19.87	10.59
w/ Sentence Shuffling	85.4	81.5	41.87	10.93	16.67	7.89
w/ Text Infilling + Sentence Shuffling	**90.8**	83.8	24.17	6.62	11.12	**5.41**

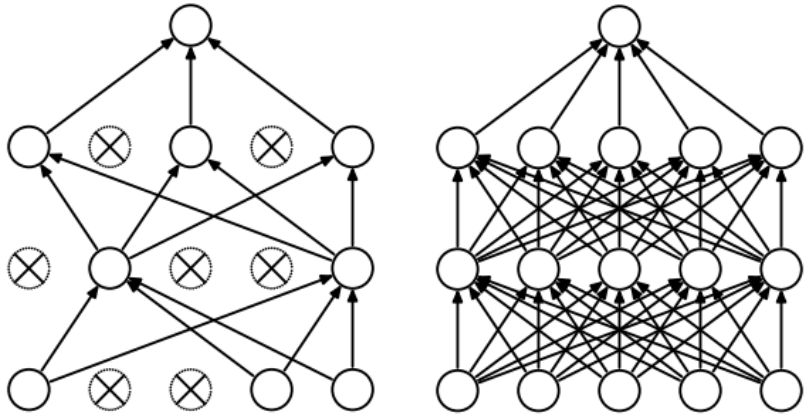
Document type Embedding + pretraining/Task-adaptive pretraining(Text Infilling)

- BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

성능 개선
Regularized Dropout(R-Drop)

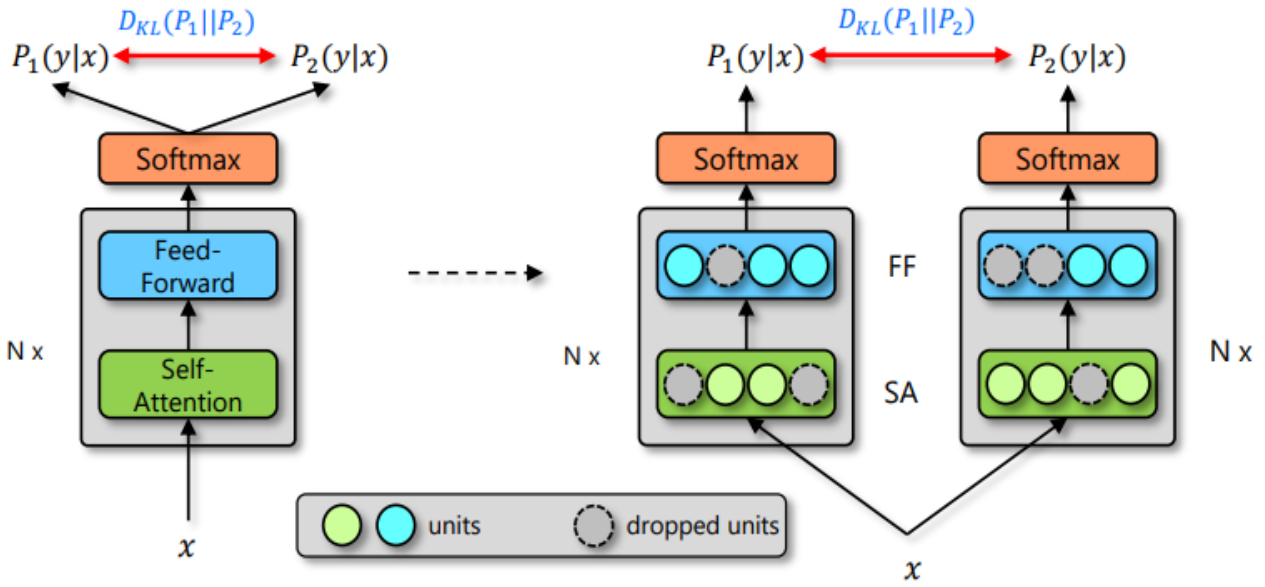
Rank	Model	ROUGE-1 ↑	ROUGE-2	ROUGE-L	Extra Training Data
1	**GLM-XXLarge**	44.7	21.4	41.4	×
2	**BART + R-Drop**	44.51	21.58	41.24	×

◀ CNN/Daily Mail 2위



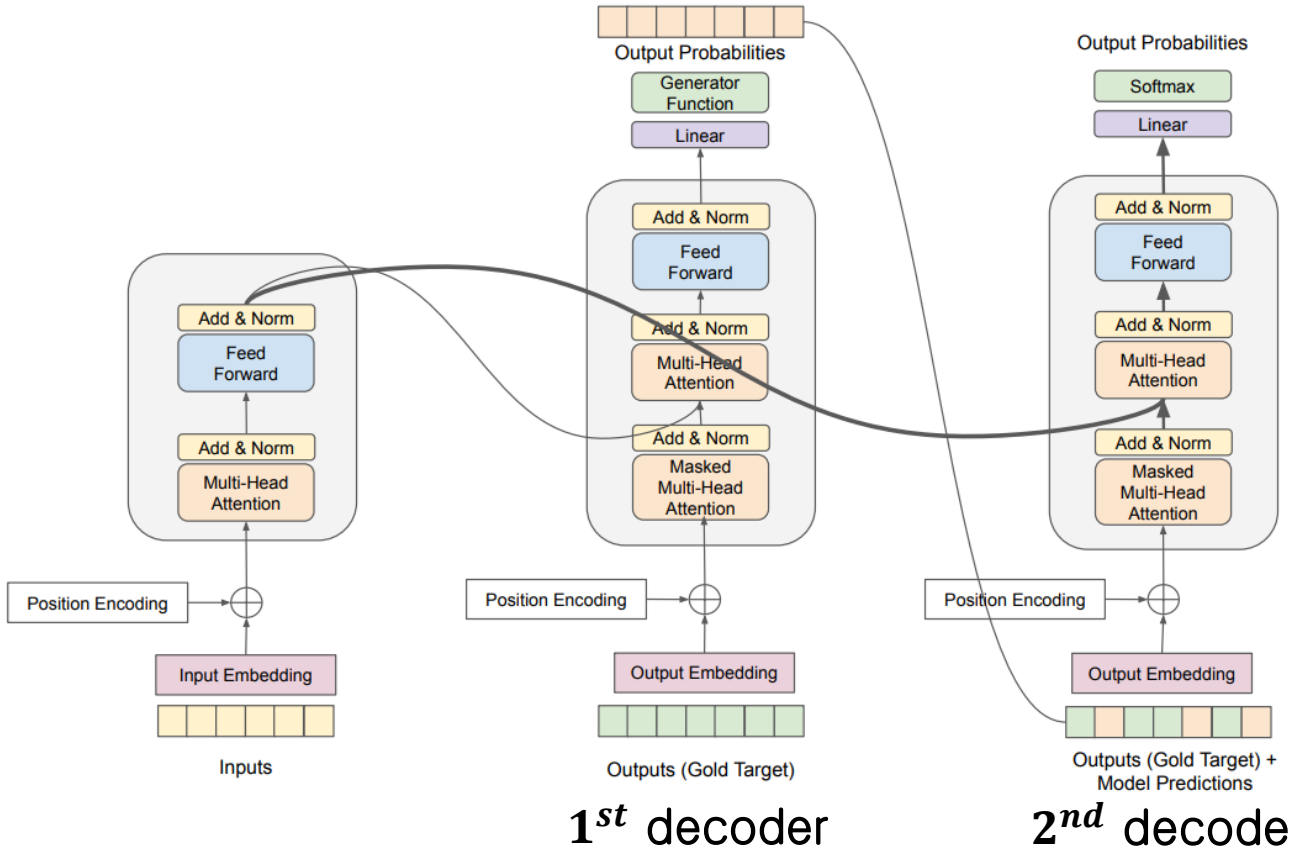
▲ Train
Dropout 적용

▲ Inference
Dropout 적용X



같은 input에 대해 다른 output이 발생하는 것을 줄이기 위해
두 output 분포 사이의 거리에 대한 loss를 추가하여 학습

성능 개선
Teacher Forcing Scheduler



Experiment	DE−EN		JA−EN	
	Dev	Test	Dev	Test
Teacher Forcing Baseline	35.05	**29.62**	18.00	19.46
No backprop				
Argmax	23.99	20.57	12.88	15.13
Top-k mix	35.19	29.42	**18.46**	20.24
Softmax mix $\alpha = 1$	35.07	29.32	17.98	20.03
Softmax mix $\alpha = 10$	35.30	29.25	17.79	19.67
Gumbel Softmax mix $\alpha = 1$	**35.36**	29.48	18.31	20.21
Gumbel Softmax mix $\alpha = 10$	35.32	29.58	17.94	**20.87**
Sparsemax mix	35.22	29.28	18.14	20.15
Backprop through model decisions				
Softmax mix $\alpha = 1$	33.25	27.60	15.67	17.93
Softmax mix $\alpha = 10$	27.06	23.29	13.49	16.02
Gumbel Softmax mix $\alpha = 1$	30.57	25.71	15.86	18.76
Gumbel Softmax mix $\alpha = 10$	12.79	10.62	13.98	17.09
Sparsemax mix	24.65	20.15	12.44	16.23

Step이 증가할 수록 stage-1 decoder을 통해 구한 prediction 사용 비율 증가 → 모델의 추론 능력 향상

성능 개선
결과

Model	Rouge-L	비고
BART ➔ 146M	34.747	
LongBART(not pretrained) ➔ 67M	27.037(-7.71)	Pretraining with Text Infilling
LongBART(pretrained) ➔ 67M	32.327(-2.42)	
BigBART(doc type and TAPT) ➔ 153M	40.987(+6.21)	Pretraining with Text Infilling
BigBART(pretrained) ➔ 153M	36.457(+1.71)	

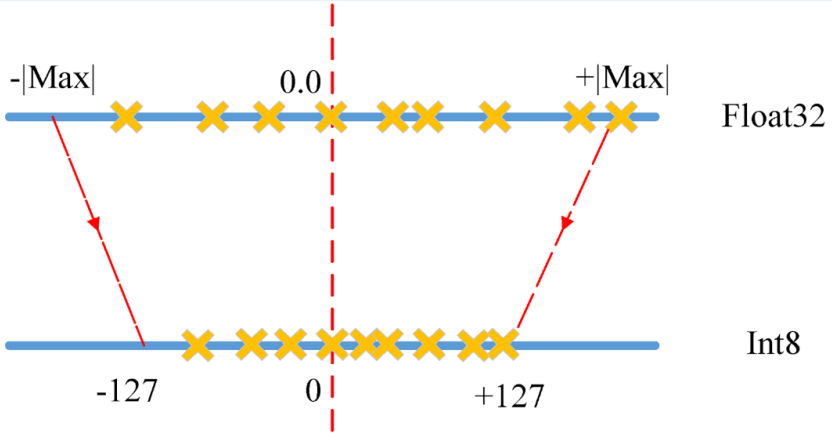
- LongBART 모델 사이즈가 가장 작지만,
  리소스 내에서 최대의 성능을 발휘하는 BigBART(doc type and TAPT) Model로 선정
-
  최종 모델: BigBART + 전처리 + R-Drop ➔ Rouge-L :41.687

4. 경량화

#Dynamic_Quantization
#Knowledge_Distillation
#경량화 결과

경량화
Dynamic Quantization

Quantization ➔ Int8

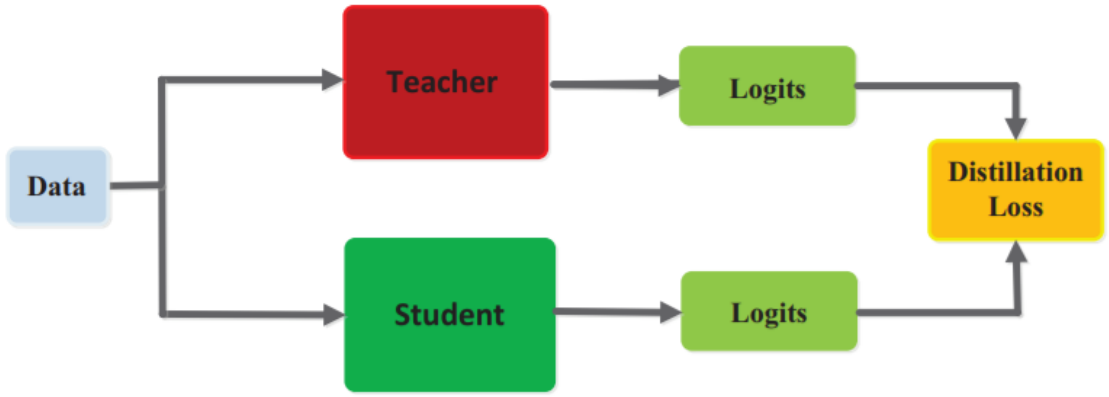


Quantization ➔ Float16

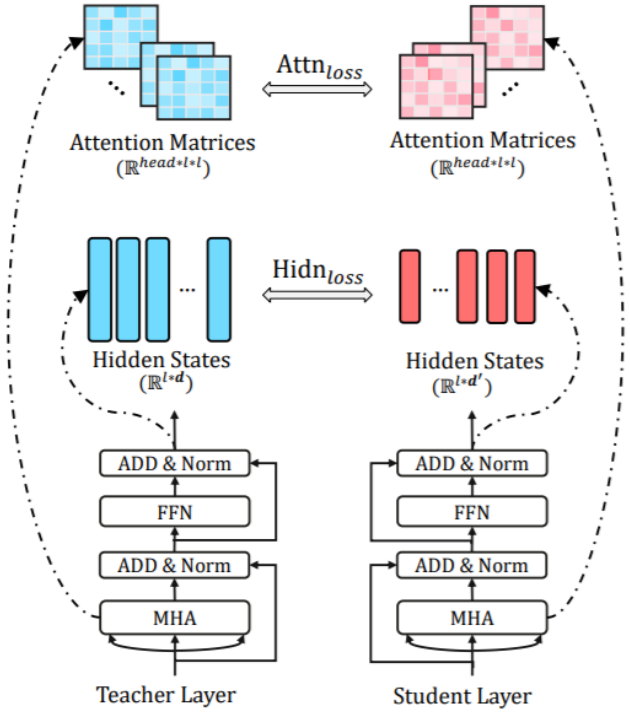
적용방법	개선사항	적용가능 Hardware
Int8	589.40MB → 223.58MB	CPU
Half	589.40MB → 294.76MB	GPU

Rouge score은 유지하였으나, latency에서의 개선사항은 없었음

경량화
Knowledge Distillation



Basic Distillation



Tiny Distillation

성능이 좋은 teacher 모델로부터 student모델이 학습

TinyBERT: Distilling BERT for Natural Language Understanding
DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter

경량화
결과

Pruning(Student Model)

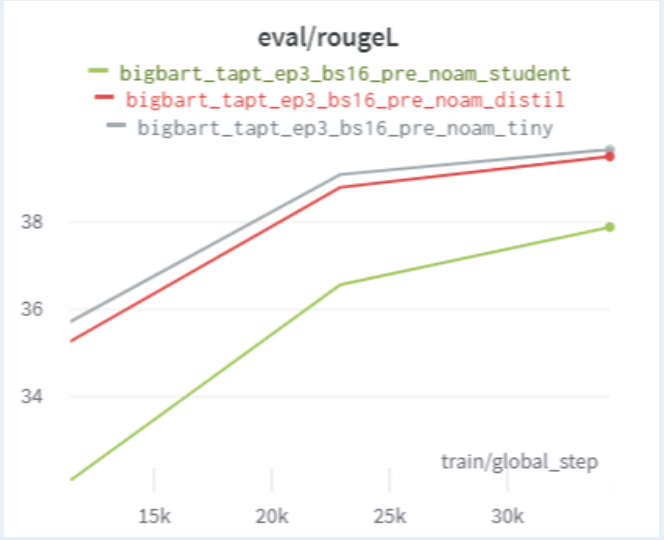
589.40MB → 400.07MB(32% 경량화)
13ms → 6.8ms(48% 감축)

Dynamic Quantization

400.07MB → 200.08MB(50% 경량화)

Distillation

적용방법	결과(Rouge-L)
Student Model	37.889
Basic Distillation	+ 1.621
Tiny Distillation	+ 1.783



5. 더 나아가기

#시각화
#ElasticSearch

더 나아가기
시각화

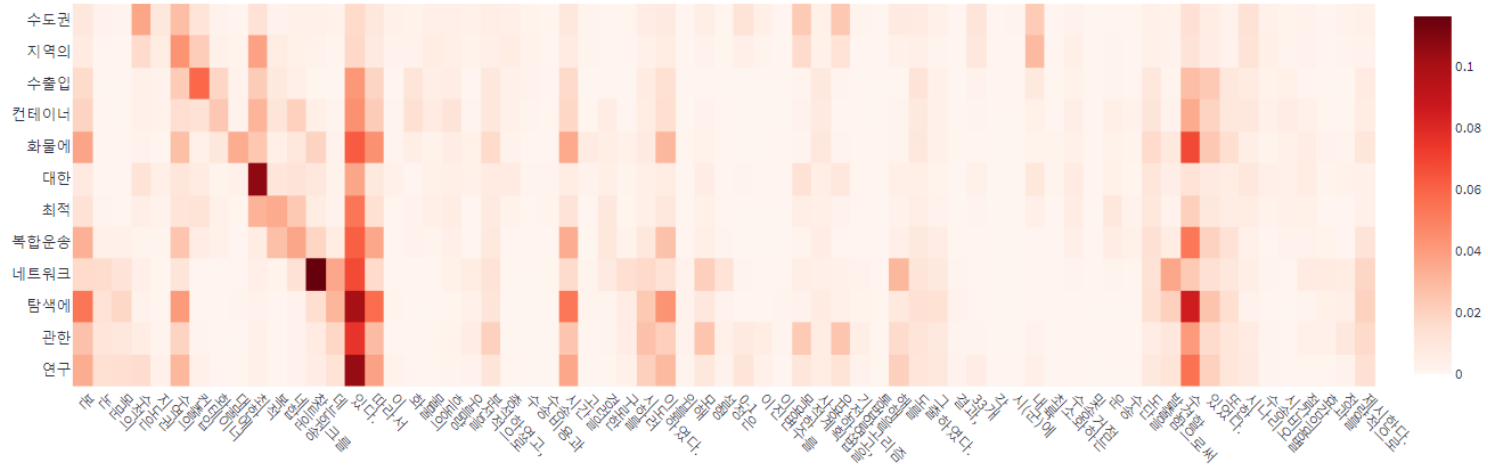
원문 제목: **유전알고리즘을 이용한 복합운송최적화모형에 관한 연구**

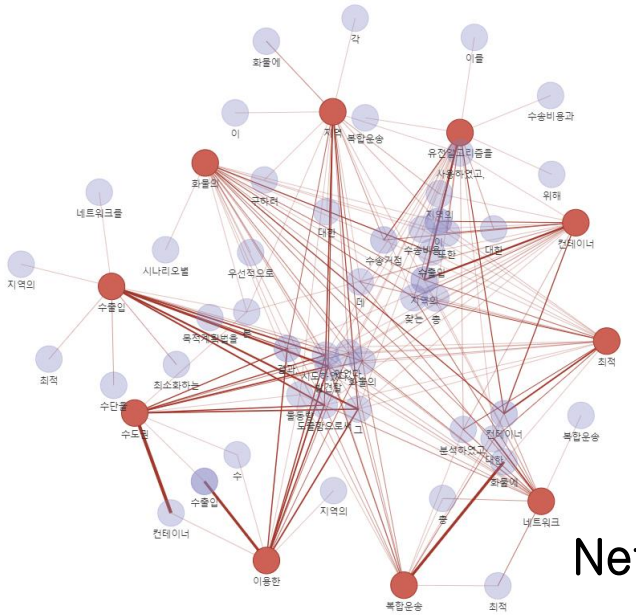
생성된 제목: **수도권 지역의 수출입 컨테이너 화물에 대한 최적 복합운송 네트워크 탐색에 관한 연구**

Text highlight

본 논문의 목적은 수도권 지역의 수출입 컨테이너 화물에 대한 최적 복합운송 네트워크를 찾는 데 있다. 따라서 이 지역의 컨테이너 화물의 물동량 흐름을 우선적으로 분석하였고, 총 수송비용과 총 수송 시간을 고려한 최적 경로를 구하려 시도하였다. 이를 위해 모형 설정은 0-1 이진변수를 이용한 목적계획법을 사용하였고, 유전알고리즘 기법을 통해 해를 도출하였다. 그 결과, 수도권 지역의 33개 각 시(군)에 대한 내륙 수송비용과 수송 시간을 최소화하는 수송거점 및 운송 수단을 도출함으로써 이 지역의 수출입 컨테이너 화물에 대한 최적 복합운송 네트워크를 발견할 수 있었다. 또한 시나리오별 수송비용 및 수송 시간의 절감 효과를 정량적으로 제시한다.

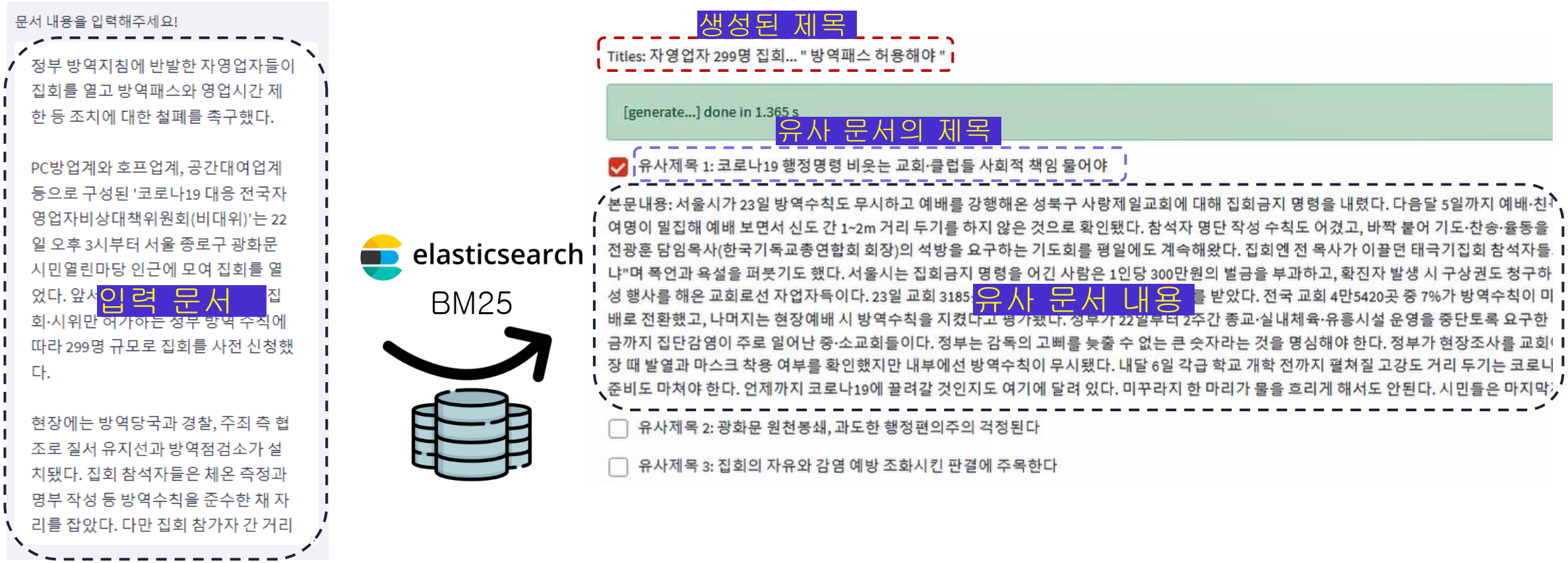
2D heatmap





Network graph

더 나아가기
ElasticSearch



보유 데이터 중 유사한 문서의 제목과 모델을 통해 생성된 제목 비교 가능

6. 회고

#회고

회고

회고

잘한 점

- 유사 데이터셋에서 좋은 성능을 보인 성능 개선 방법론에 대해 다양한 시도를 함
- 개개인의 역량 파악을 통해 각 태스크 별로 정확한 일정 관리가 이뤄짐
- 프로젝트 전 과정에 대한 기록 및 github/notion/google sheet/wandb를 통한 협업이 원활히 이뤄짐
- 이전 대회에서 각자 다뤄보지 못한 태스크(model,data,visualization 등)들을 충분히 다뤄봄으로써 개인의 역량이 강화됨

아쉬운 점

- Korean Rouge Score의 평가 방법 오류를 늦게 파악하여 전체 성능 파악이 늦어짐
- Unstructured Pruning을 BART모델에 적용해보지 못함.

개선 방향

- 데이터 증량 시 성능 개선 및 LongBART의 Pretraining이 가능할 것으로 판단됨.
- BART 이외에 다양한 모델(T5,PEGASUS, switch transformers 등) 시도를 통해 성능 개선 및 적용 범위를 확대시킬 수 있을 것이라 판단됨
- 생성 요약만 진행했지만, 추출요약을 수행 후 시도를 했더라면 더 좋은 모델이 구현됐을 것으로 판단됨 (BERTsum)

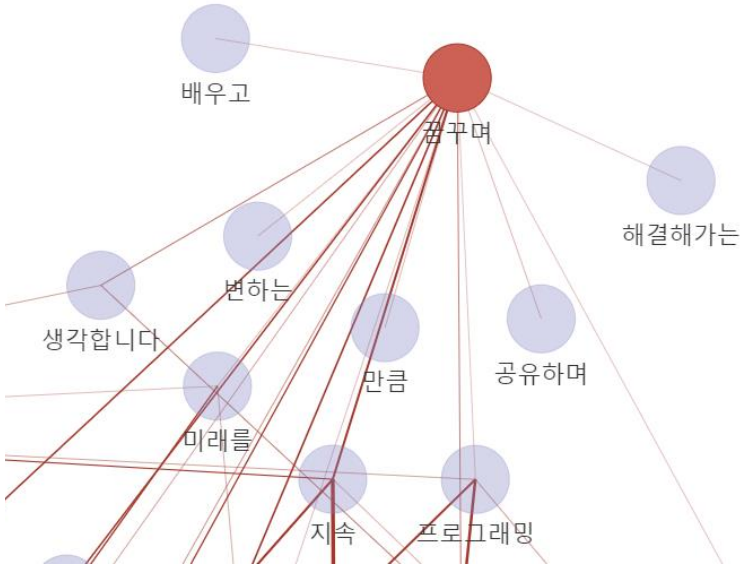
Title : 부스트캠프, 지속 가능한 개발자를 꿈꾸며

Titles: 부스트캠프, 지속가능한 개발자를 꿈꾸며

[generate...] done in 1.227 s

Visualization!

부스트캠프는 단순히 뛰어난 프로그래밍 스킬을 가진 사람이 아닌 , 사람들과 끊임없이 커뮤니케이션하고 서로에게 배우고 공유하며 문제를 해결해가는 사람 , 그렇게 더 큰 미래를 그리는 사람이 좋은 개발자이고 지속 가능한 개발자라고 생각합니다 . 좋은 경험과 습관을 가진 지속 가능한 개발자 양성을 목표로 , 최고의 전문가들이 모여 만들어낸 소프트웨어 교육 커뮤니티 . 누구에게나 열려 있는 기회 , 그러나 강도 높은 주도적인 문제 해결의 경험을 제공합니다 . 세상이 빠르게 변하는 만큼 부스트캠프는 매년 새로운 시도들로 새로운 가능성을 찾습니다 .



감사합니다

Q&A