AI 기반 회의 녹취록 요약 경진 대회 연습 참가

백화요란 : 2022.06.13~2022.06.19

百花擾亂

목차

대회 개요

- AI 기반 회의 녹취록 요약 경진 대회
- 카테고리 : LG | 자연어 | 생성요약
- 기간 : 2021.09.27 ~ 2021.10.25
- 참여자 : 863명
- 심사 기준 : ROUGE-N 에 대한 F1 Score

참가 목적

- JSON Reform / EDA를 통한 Augmetation 구현 및 학습
- 기간 : 2022.06.13 ~ 2022.06.19
- Tool :
  Python / JSON / NLTK / HuggingFace / Wandb
- 참여자 :

  박정현 : 프로젝트 총괄

  권남우 : 전처리, 모델링, 보고서 및 PPT 작성

  전영욱 : 전처리

  이성준 : 모델링

百花擾亂

참가 목표

대회 수상작 보다 더 나은 성능의 모델 개시

Text Summarization (ROUGE-N에 대한 F1 Score)

JSON

Data Frame

Augmentation

- Random Insertion
- Random Replacement
- Random Deletion
- Random Swap

Fine Tuning

ainize/kobart-news

百花擾亂

Text Summarization - 데이터 셋

문서화된 회의 녹취록

- 데이터에 대한 요약문을 라벨링한 데이터 셋
- JSON
- id : 회의록 id
- region : 회의 지역
- num_agenda : 안건 수
  - AGENDA_1 : 안건 1
  - AGENDA_2 : 안건 2
- label : 안건별 요약문
  - AGENDA_1 : 안건 1의 요약문
    - evidence : 요약 근거
    - summary : 요약문 (정답)

Train Data Set

Train : Row 2994 / Column 3

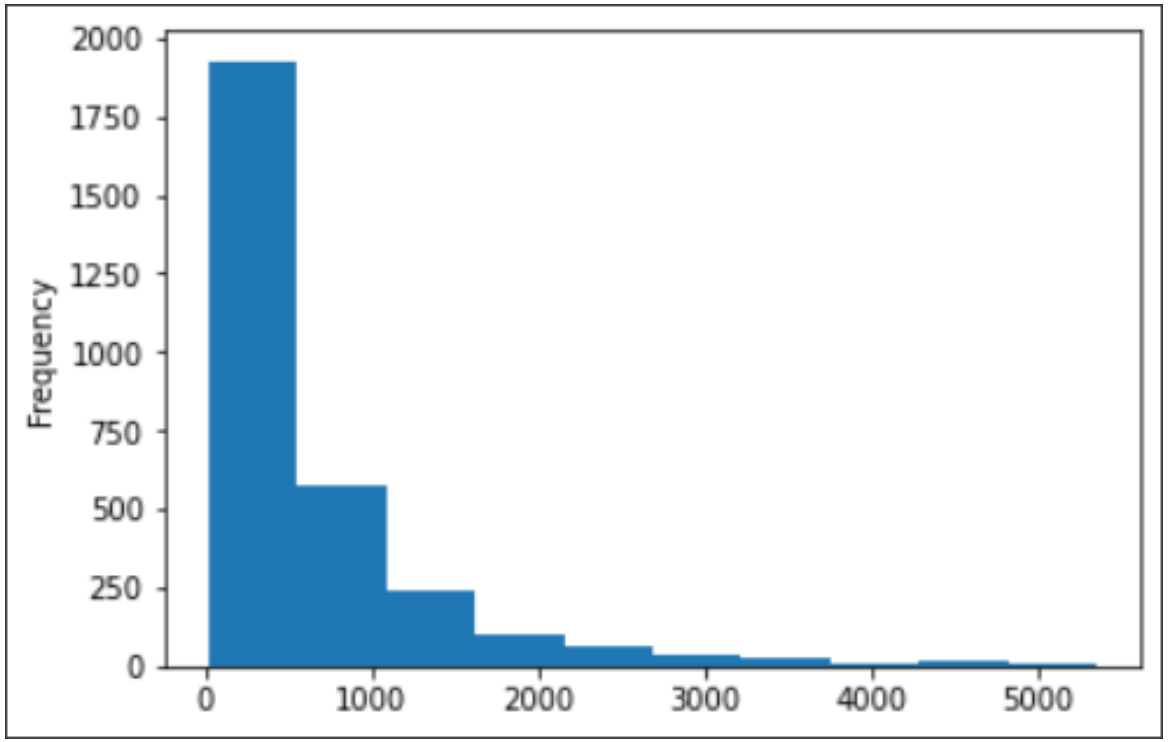
Column

- context : 회의록
- summary : 요약문
- evidence : 요약 근거

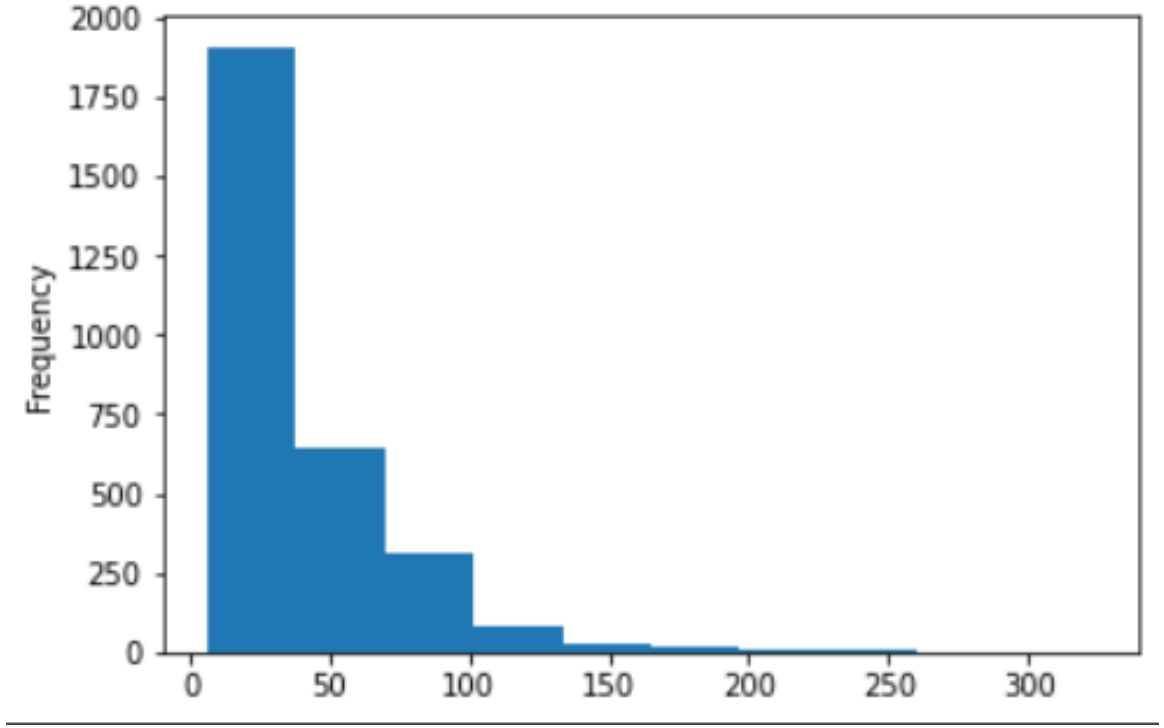
百花擾亂

Text Summarization - EDA Context ・Summary 문장 비교

Context Length



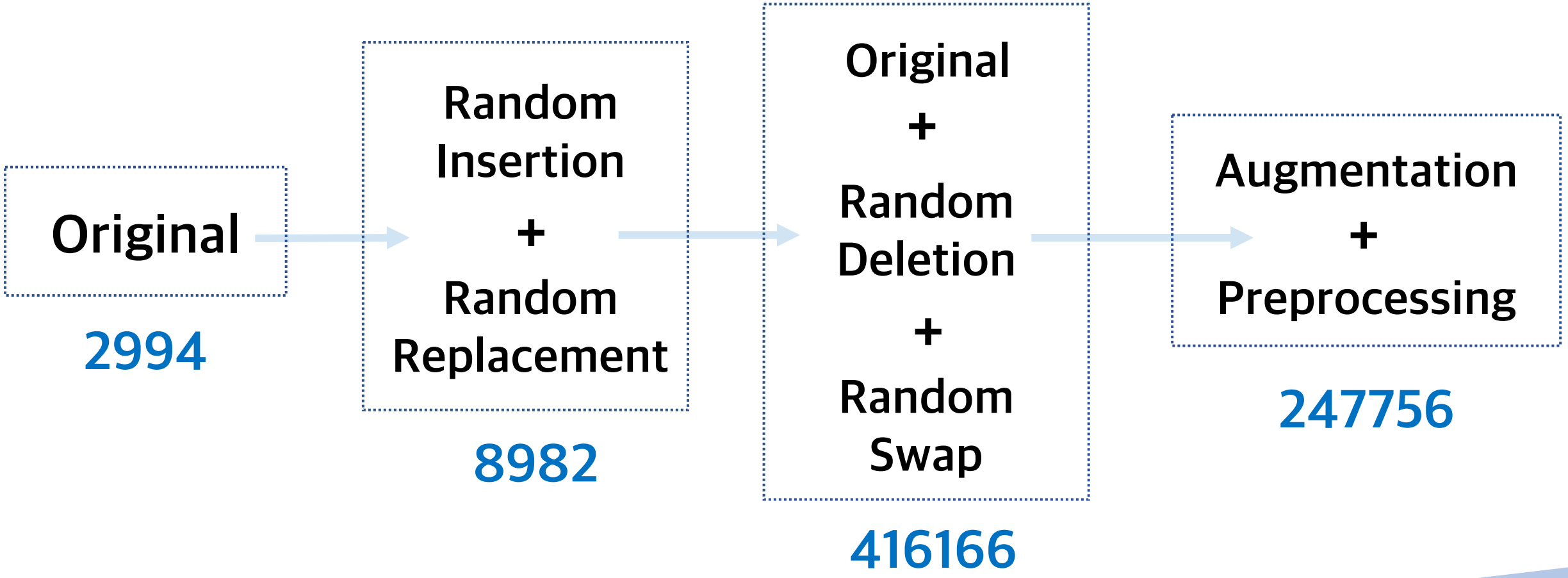
Summary Length



ainize/kobart-news

- Kobart
- Pre - Trained Model
- AI Hub 문서요약 텍스트/신문기사
- Tokenizer
  - Model : Max position embeddings : 1026
  - Max Input Length : 1026
  - Max Target Length : 514

Text Summarization - Augmentation Process



Text Summarization - Augmentation Example

Random Insertion

문장 내 무작위 위치에
단어를 삽입

원본 : 다음은 완주군수로부터 제출된 안건입니다

증식 : 다음은 완주군수로부터 공식 제출된 안건입니다

Random Replacement

문장 내 임의의 단어 하나를
비슷한 의미의 단어로 교체

원본 : 그럼 두 분 위원님께서는 본 회기동안 수고하여 주시기 바랍니다

증식 : 그럼 두 명의 의원님께서는 본 회귀동안 수고하여 주시기 바랍니다

百花擾亂

Text Summarization - Augmentation Example

Random Deletion

문장 내 임의의 단어 하나를 삭제

원본 : 의석을 정돈하여 주시기 바랍니다

증식 : 의석을 정돈하여 바랍니다

Random Swap

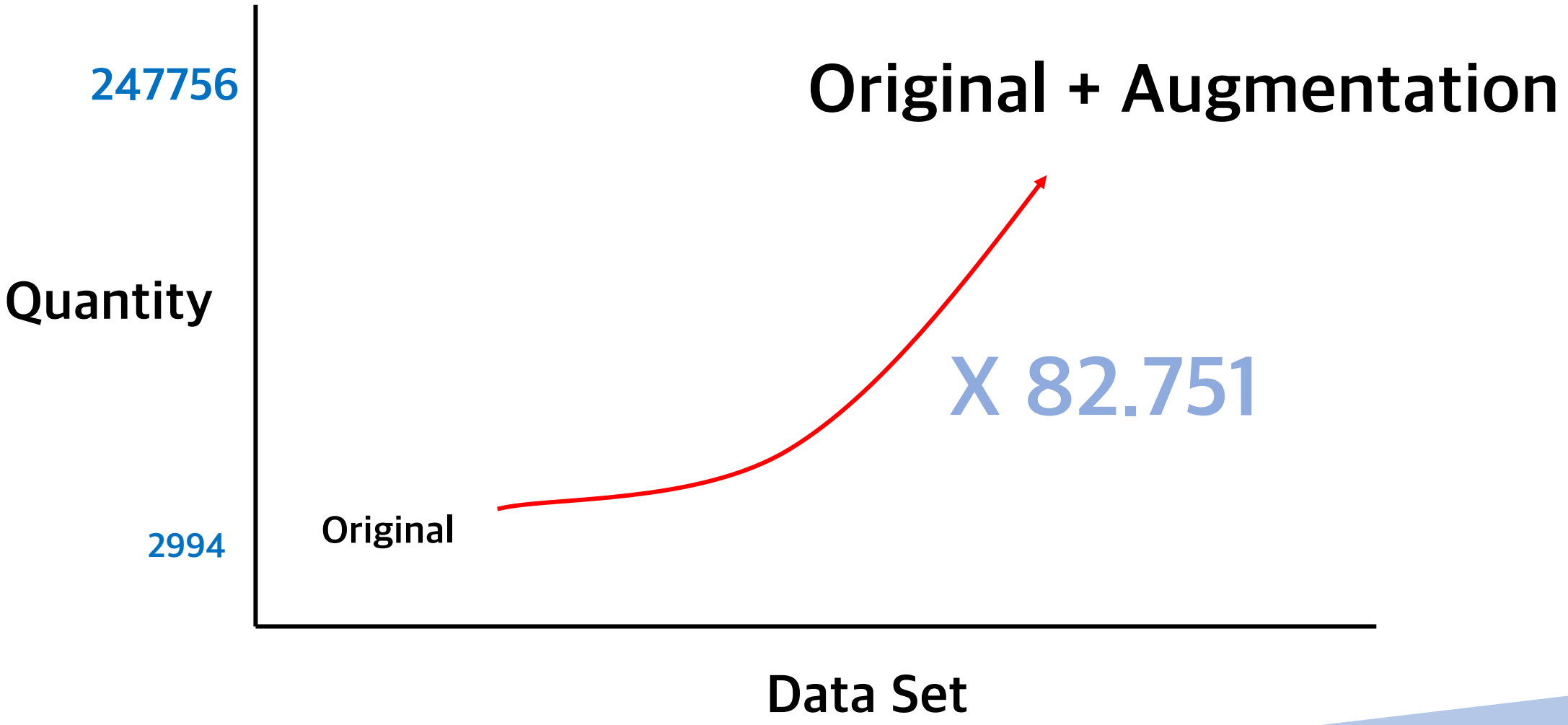
문단 내 임의의 두 문장의 위치를 변환

원본 :
다음은 의사일정 제2항 … 선출의 건을 상정합니다. 회의록 서명의원으로는 … 의원 여러분 이의 있습니까? 이의가 없으므로 가결되었음을 선포합니다. …

증식 :
이의가 없으므로 가결되었음을 선포합니다. 회의록 서명의원으로는 …
의원 여러분 이의 있습니까? 다음은 의사일정 제2항 …
선출의 건을 상정합니다. …

百花擾亂

Text Summarization - Augmentation Result



Text Summarization - 모델링 Hyper Parameter

Learning Rate

2e-5

LR Scheduler

Cosine

Weight Decay

0.01

Warm up Ratio

0.1

Epoch

비슷한 데이터로 인해 Epoch을 반복하는 효과가 발생

1

Train Batch Size

Val Batch Size

결과론적 이유

1

Eval Steps

247756/46

5368

Generation Max Length

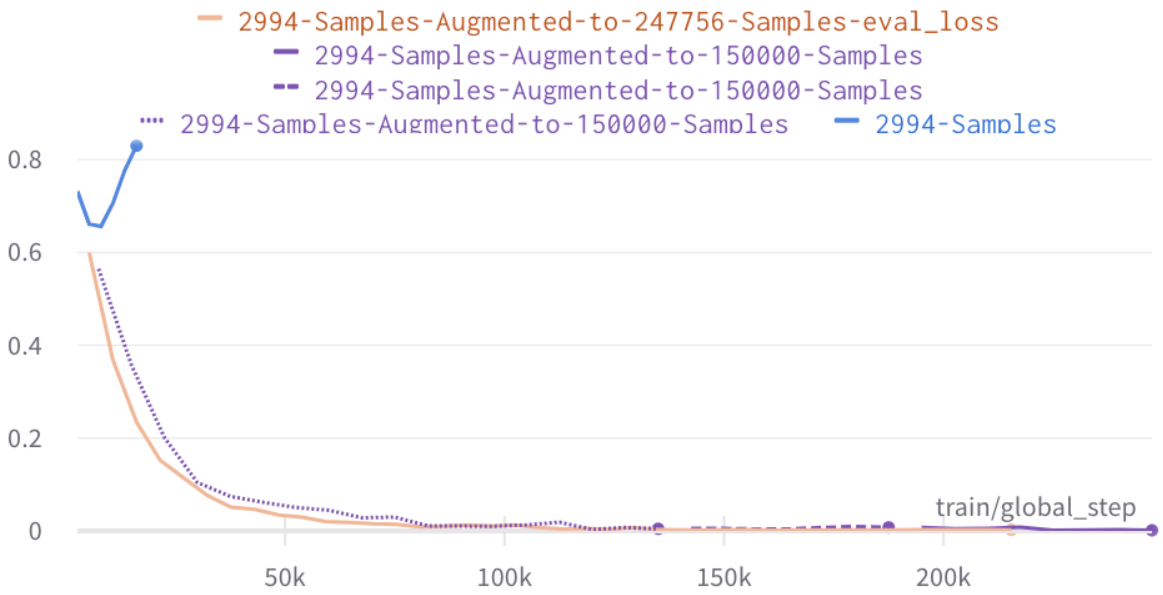
Generation Num Beams

512 / 5

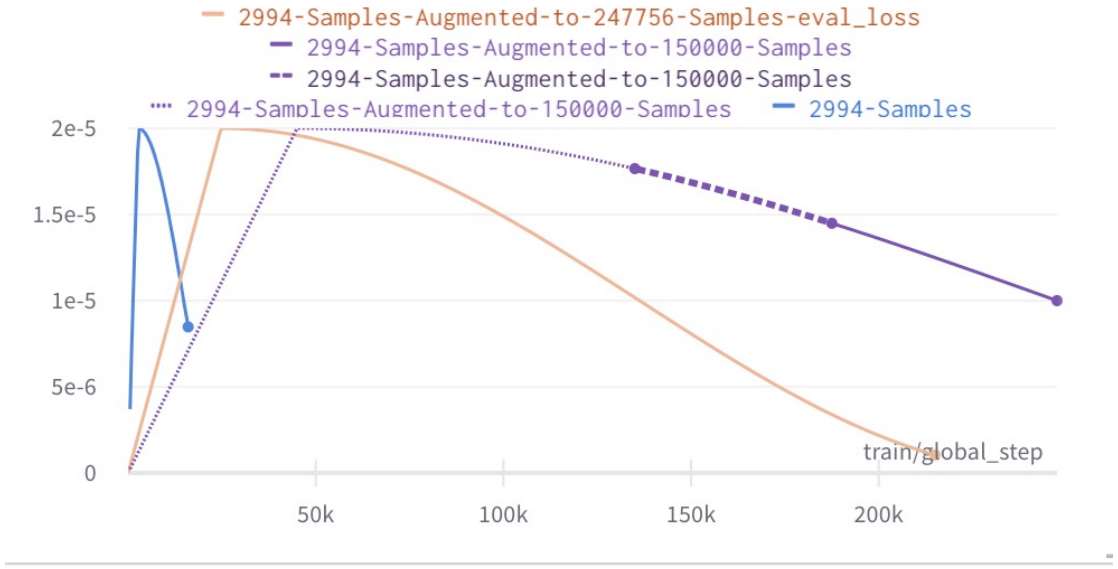
百花擾亂

Text Summarization - 모델링 Training

Eval Loss



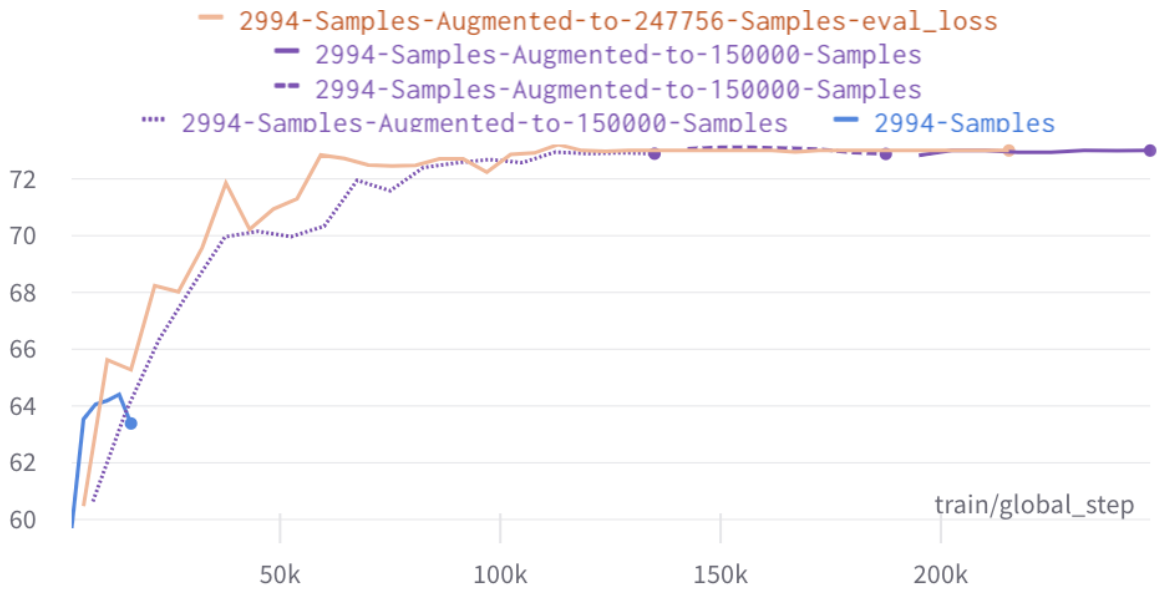
Learning Rate



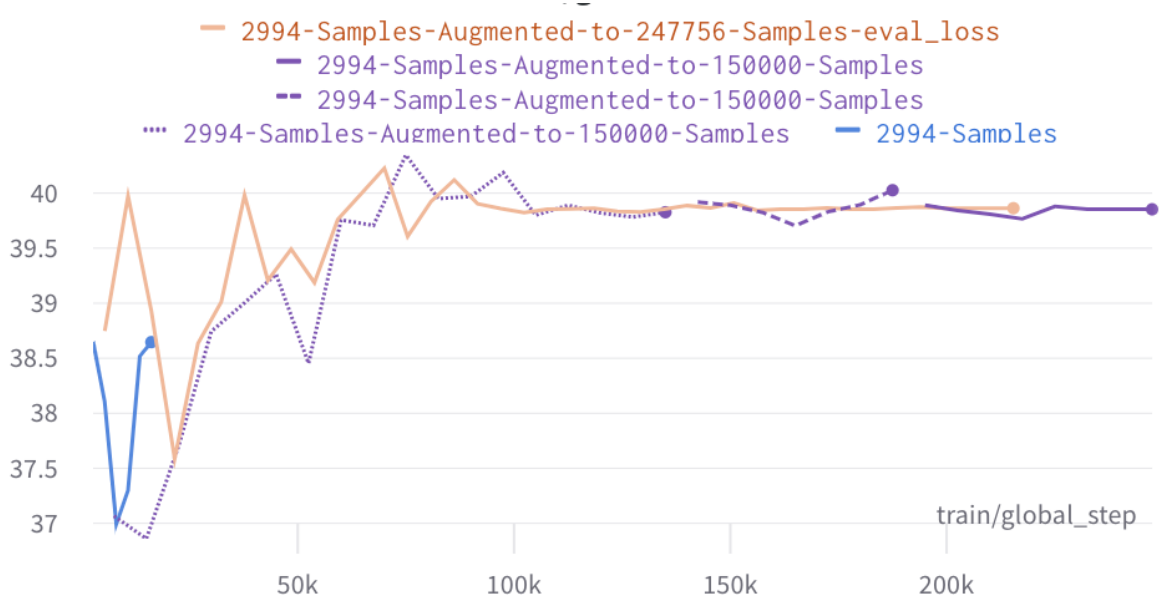
百花擾亂

Text Summarization - 모델링 Training

Eval Rouge 1



Eval Gen Length



百花擾亂

Text Summarization - 모델링 Performance

Step	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
5386	0.926200	0.599154	60.469000	41.365500	60.348900	60.508200	38.746700
10772	0.582000	0.368559	65.621800	46.318000	65.324400	65.497100	39.973300
16158	0.381500	0.234490	65.276400	47.309900	65.257100	65.219600	38.930000
21544	0.255800	0.152114	68.235700	48.495900	68.291400	68.172500	37.586700
26930	0.176900	0.113680	68.021600	48.918000	68.076600	67.936400	38.633300
32316	0.117800	0.076253	69.574200	49.657700	69.404800	69.376100	39.010000
37702	0.110300	0.051037	71.854800	53.028200	71.703000	71.774900	39.980000
43088	0.086000	0.046541	70.221400	50.916000	70.271000	70.150000	39.206700
48474	0.072500	0.034603	70.941100	51.643700	70.986100	70.949400	39.490000
172352	0.007000	0.001737	73.000000	55.000000	73.222200	73.111100	39.866700
177738	0.010600	0.001856	73.000000	55.000000	73.222200	73.111100	39.853300
183124	0.007600	0.001716	73.000000	55.000000	73.222200	73.111100	39.853300
188510	0.008500	0.002337	73.000000	55.000000	73.222200	73.111100	39.866700
193896	0.009000	0.002581	73.000000	55.000000	73.222200	73.111100	39.873300
199282	0.007000	0.002595	73.000000	55.000000	73.222200	73.111100	39.863300
204668	0.004700	0.002545	73.000000	55.000000	73.222200	73.111100	39.863300
210054	0.006300	0.003026	73.000000	55.000000	73.222200	73.111100	39.863300
215440	0.006000	0.002342	73.000000	55.000000	73.222200	73.111100	39.863300

Context

의사일정 제4항, 음성군 예산절감 및 예산낭비사례 공개 등에 관한 조례안을 상정합니다.본 의원이 대표발의한 본 안건에 대한 제안설명은 서류로 갈음하고자 하며, 사전 의원간담회에서 충분히 협의된바 바로 의결하고자 합니다.의사일정 제4항, 음성군 예산절감 및 예산낭비사례 공개 등에 관한 조례안을 원안대로 의결하고자 하는데, 의원 여러분! 이의 없으십니까? ...

Label

음성군 예산절감 및 예산낭비사례 공개 등에 관한 조례안은 가결됨.

Predict

음성군 예산절감 및 예산낭비사례 공개 등에 관한 조례안은 가결됨.

Public 7등

(430명 中)

Q & A

감사합니다