

제3회 한국가스공사 (KOGAS)

빅데이터·인공지능 스타트업 경진대회

가스·에너지분야 문서요약 모델개발

경진대회 수행내용 보고

이야기연구소 주식회사

- 팀장 : 오명교
- 팀원 : 박인현



한국가스공사



1. 배경 및 개요

제3회 한국가스공사 (KOGAS)

빅데이터·인공지능
스타트업 경진대회

가스·에너지분야 문서요약 모델개발



• 배경

- 본 발표 자료는 한국가스공사에서 주최 및 주관하고, 데이콘(주)에서 운영하는 “가스·에너지분야 문서요약 모델개발” 경진대회에 참가한 “이야기연구소 주식회사” 팀의 2차 평가용 자료임
- 경진대회 간 수행한 내용을 본 자료로 공유하는 주요 목적은 다음과 같음
 - 대회에서 수행한 실험 과정 및 결과의 신뢰성 함양
 - 추후 자료 및 코드의 공개를 통해 관련 분야에 대한 신진 연구자들의 접근 장벽 완화

• 개요

- 당사의 경진대회 1차 평가 테스트셋에 대한 최종 순위 및 결과는 다음과 같음
 - [최고점수] ROUGE-1: 0.3678 / ROUGE-2: 0.1803 / ROUGE-N: 0.2800
 - [최종선택] ROUGE-1: 0.3662 / ROUGE-2: 0.1811 / ROUGE-N: 0.2731
- 모델 가중치를 제외한 모든 실험 산출물(코드, 로그, 발표 자료)은 대회 종료 이후 공개 예정
 - Github(*.py): (현재 비공개 상태)
 - DACON(*.ipynb): (추후 추가 예정)
 - 훈련 로그: <https://tensorboard.dev/experiment/q59x1CV9RlydqWWTWWWuXQ/>



2. 사용 데이터

• 문서요약 텍스트(v1.2)

- 소개

- AI가 텍스트를 이해하고 핵심 내용을 자동으로 요약 또는 생성하는 기술개발을 위한 텍스트 데이터
- [AI Hub](#)에 공개되어 있으며, 별도의 사용신청이 승인된 이후 다운로드 가능

- 적용 내용

- 공개된 [법률], [사설잡지], [신문기사] 데이터 중 공식 데이터인 [신문기사]만 훈련&검증용으로 사용
- 데이터 세트 개요(이상치/결측치 제거, 개행문자 포함)

구분	훈련용	검증용	평가용	합계
신문기사	문서 수	271,088	30,122	6,596
	단어 수	59,789,951	6,860,891	1,780,607
	글자 수	648,460,445	74,725,970	18,746,612
	문서당 글자 수	2,392	2,481	2,842
요약문	문서 수	271,088	30,122	-
	단어 수	7,747,510	891,325	-
	글자 수	85,172,592	9,812,719	-
	문서당 글자 수	314	326	-

- 이외에 외부 사용 데이터 없음



2. 사용 데이터

- 탐색적 데이터 분석(Exploratory Data Analysis; EDA)

- 수행 목적

- 데이터 분포를 직접 확인함으로써, 모델링 전 데이터에 대한 좋은 인사이트(insight) 포착
 - 도메인(언론사)별 맞춤 전처리, 특별히 텍스트 정제(text cleaning) 과정 구현
 - 훈련 및 추론에 사용되는 인자(argument)의 적절성 검토 및 근거 마련

- 전제 조건 및 기본 가설

- EDA에 사용되는 데이터는 오직 훈련용 데이터 세트로만 한정
 - 과적합(overfitting) 및 데이터 누수(data leakage) 방지
 - 검증용 및 추론용 데이터 세트 집합의 분포는 훈련용 데이터 세트 집합과 같지 않을 수 있다고 가정
 - 도메인 분포가 같으면 가장 이상적이지만, 사업화 등 실제 응용 단계에서의 원본(raw) 입력 또한 고려해야 함



2. 사용 데이터

- 탐색적 데이터 분석(Exploratory Data Analysis; EDA)

- 언론사(key="media_name")별 데이터 확인

- 절차

- 훈련용 데이터에 대해, 언론사별로 데이터 분할 및 JSON 저장 → [건설경제] 등 42개 하위 도메인 생성
 - 공통으로 적용 또는 도메인별 적용을 위한 텍스트 정제 작업 구현 → 정규 표현식(regular expression; regex) 사용

- 민감정보 제거 과정 구현

- [신문기사] 데이터에 주로 포함될 것으로 기대되는 [기자 정보] 등은 미정제시 민감정보 이슈 발생 가능성 존재
 - 그 이외에 다량 포함된 것으로 확인된 전화번호, URL, 이메일 등의 개인정보 또한 훈련 전 제거 필수

- 언론사별 맞춤 정제 작업 구현

- 광고 등 본문 외 데이터

- [id=359838215] 부산일보 문서 중 “▶ 네이버에서 부산일보 구독하기 클릭!” 등의 광고 문구 제거
 - [id=333035450] 이데일리 문서 중 “이데일리 채널 구독하면 [방탄소년단 실물영접 기회가▶]” 등의 광고 문구 제거
 - [id=361593984] 매일경제 문서 중 “스탁론”이라는 문구 포함 시 해당 문서 전체가 광고를 목적으로 제작되었다고 판단, 전체 제거

- 언론사에 따라 불필요하게 반복적으로 등장하는 단어

- [id=351089794] 전남일보 문서 중 “뉴시스”, “편집애디터” 등의 반복적 단어 삭제
 - [id=350993467] 제주일보 문서 중 “제주신보” 등의 반복적 단어 삭제
 - [id=329588480] 충청일보 문서 중 “온라인충청일보” 등의 반복적 단어 삭제



2. 사용 데이터

• 탐색적 데이터 분석(Exploratory Data Analysis; EDA)

- 훈련용 데이터의 입출력(신문기사&요약문) 길이에 대한 관계 분석 및 인사이트 도출

- 절차

- 훈련용 데이터 세트의 모든 입출력에 대해 토크나이즈(tokenize) 진행
 - 입력 길이 대비 출력 길이 시각화를 통해 이상치 검출 및 관계 분석

- 이상치 검출

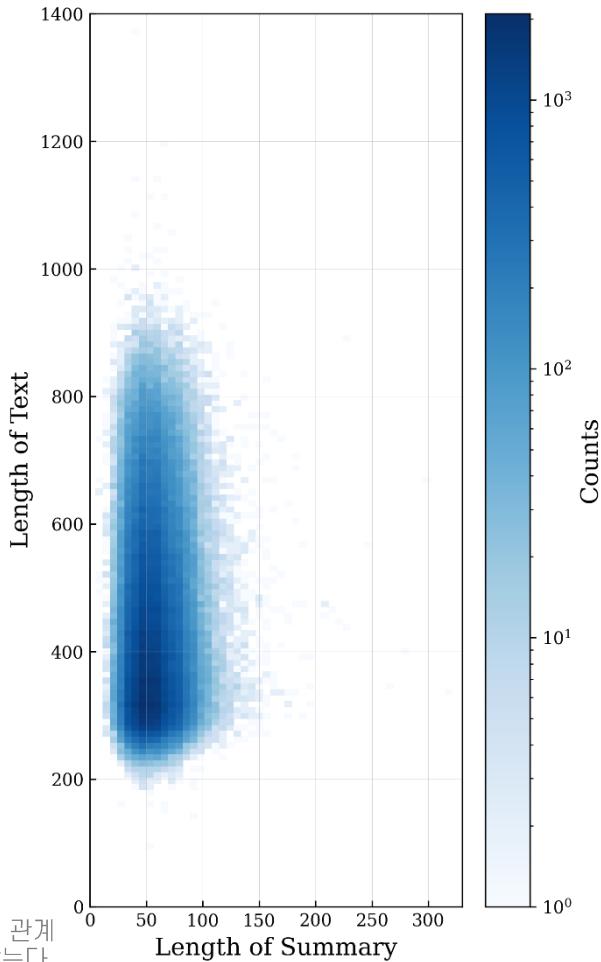
- 구분자(separator) 오류로 인해 다수의 문서가 하나의 텍스트로 취급된 이상치 처리
 - [id=338143341] 길이 14,792자의 본문 존재 (12개의 문서가 병합)

- 입출력 길이 관계 분석

- 훈련 간 최대 입력값(max_position_embeddings)은 기본값 1024에서 변경할 사항 없음
 - 추론 간 최대 출력값(max_length)은 기본값 256에서 조정할 필요가 없다고 판단
 - 값이 지나치게 작을 경우, 문장이 완성되지 못함
 - 값이 지나치게 클 경우, 복문으로 구성되며, 추론 시간이 지나치게 길어짐

- 배치(batch) 단위 패딩(padding)의 도입 근거 마련

- 패딩을 고정된 값이 아닌 배치 단위의 최대 길이로 설정 시 훈련 및 추론 속도의 향상 도모 가능



▶ 토크나이징된 입력(신문기사) 및 출력(요약문) 사이의 관계
대다수의 데이터는 그 입력 길이가 1024, 출력 길이가 256을 넘지 않는다.



2. 사용 데이터

- 전처리 (preprocessing)

- 수행 목적

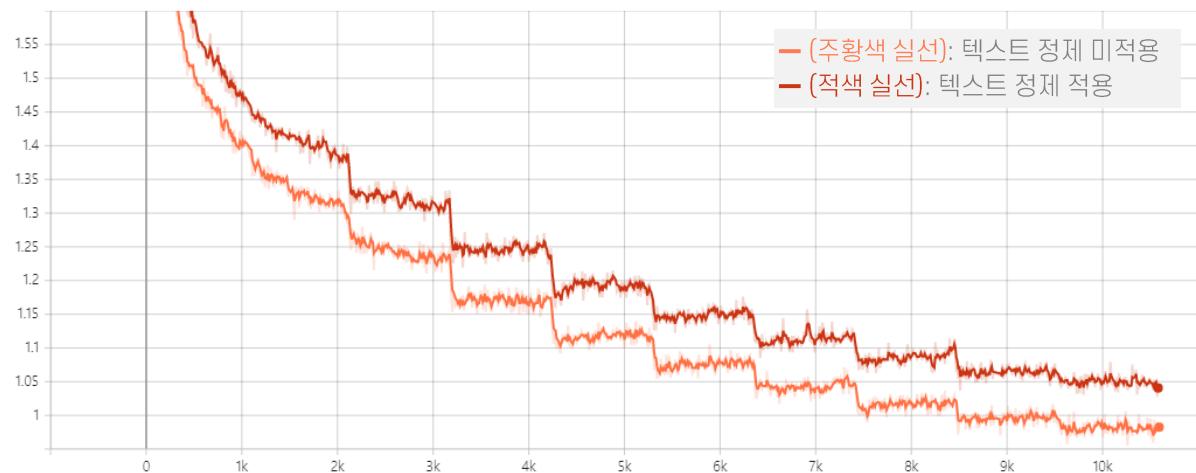
- 훈련에 부정적인 영향을 끼칠 것으로 예상되는 일부 문자열 제거 (remove) 및 교체 (replace)
 - 다만, [훈련에 부정적인 영향]은 언제까지나 개발자의 주관이므로, 훈련 결과 확인 후 기능 추가 여부 검토 필요
 - 데이터 저장 및 출력이 용이한 데이터프레임 (DataFrame) 형태로 저장

- 텍스트 정제

- 언론사별로 정제 작업을 수행했지만, 성능 (rouge score)이 좋지 않아 최종적으로는 적용하지 않음
 - 향후 사업화 단계에서 심층 구현 및 재적용 예정

- 단일 데이터 세트로 저장

- CSV (comma separated value) 대신 TSV (tab separated value) 확장자 선택
 - 문서 내 콤마 (,) 존재 가능성 고려
 - 사전 정제 작업 간 템 (wt) 공백기호 제거



3. 사용 모델기술

제3회 한국가스공사 (KOGAS)

빅데이터·인공지능
스타트업 경진대회

가스·에너지분야 문서요약 모델개발



• 개발 환경

- 기술 목적

- 실험 결과의 재현성 검증에 중대한 영향을 끼칠 수 있는 하드웨어 및 라이브러리 정보 공지

- 제원(specifications)

- OS: Ubuntu 16.04.7 LTS
 - RAM: 177GB
 - GPU: NVIDIA Tesla V100 32GB * 2장
 - CUDA Version: 11.3
 - Python: 3.8.5
 - 딥러닝 프레임워크: PyTorch 1.10.0+cu113



3. 사용 모델기술

• 데이터 입력 파이프라인

• Dataset

- 데이터를 가지고 있는 객체로, 문자열(string) 입력을 미리 토크나이징하여 List[int] 타입으로 변환
- 입력으로 인덱스(index)를 받은 뒤, 이에 따라 Dict[str, List[int]] 타입의 적절한 신문기사&요약문 데이터 쌍 반환
- 토크나이즈가 이루어진 입력 데이터의 길이에 따라 데이터를 미리 내림차순 정렬
 - 유사한 길이의 입력끼리 배치 단위로 묶어도록 유도함으로써 패딩 최소화
 - 내림차순 정렬을 통해 사전에 VRAM의 Out-Of-Memory 여부 확인 가능

• Sampler

- 데이터를 인출하기 위한 인덱스 생성 및 반환
- 입력 데이터의 길이에 따라 데이터를 정렬 후, 약간의 무작위성(randomness)을 도입하여 shuffling 효과 기대
 - HuggingFace Seq2SeqTrainingArguments 클래스에 [sortish_sampler](#) 인자 추가

• Collator

- 배치 단위의 병렬(parallel) 연산을 위해 가변 길이의 입력을 고정 길이로 변환
- 모델 입력에 대한 부가 정보(attention mask) 생성 및 토큰(<BOS>, <EOS>, <PAD>) 병합



3. 사용 모델기술

- **사전 학습 모델**(pretrained language model; PLM)
 - 사용 목적
 - 대량의 말뭉치(corpus)로 사전 훈련된 모델을 미세조정(fine-tuning) 하는 과정을 통해 시간, 컴퓨팅 파워 등 자원 절약
 - **BART**(Bidirectional and Auto-Regressive Transformers)
 - 입력 텍스트 일부에 노이즈를 추가하여 이를 다시 원문으로 복구하는 오토인코더(auto-encoder) 구조
 - 대량의 말뭉치에 대해 비지도 학습으로 사전 학습되며, 문장 생성(NLG)을 위한 테스크로 다운스트림(downstream)
 - 기본적으로 영어로 학습된 모델이므로, **한국어로 학습된 모델인 KoBART** 구조를 일반적으로 사용
- **공개된 모델 및 코드**(HuggingFace Hub 또는 Github)
 - 사용 모델 및 참고 코드
 - [gogamza / kobart-base-v1](#)
 - [seujung / kobart-summarization](#)
 - 동일/유사 데이터세트로 훈련된 일부 사전 학습 모델은 후보군에서 제외(data leakage issue)
 - [ainize / kobart-news](#)

3. 사용 모델기술

제3회 한국가스공사 (KOGAS)

빅데이터·인공지능
스타트업 경진대회

가스·에너지분야 문서요약 모델개발



- 훈련용 하이퍼 파라미터(hyper-parameter)
 - 개요
 - 훈련 과정에서 사용되는 인자에 대해 설명
 - 시간 절약을 위해 훈련용 인자는 대부분 고정한 채 추론용 인자에 대해 다양하게 탐색(grid search)
 - Engine
 - HuggingFace의 Seq2SeqTrainingArguments 및 Trainer 적용
 - HuggingFace Hub에 게시된 사전학습 모델의 미세조정 과정을 간단하게 구현 및 테스트 가능
 - Multi-GPU 사용, 혼합정밀도(mixed precision) 적용, 텐서보드(tensorboard) 로그 관리 등 용이
 - Optimizer & Learning Rate Scheduler
 - AdamW Optimizer (기본값)
 - Weight Decay: 1e-2
 - Linear Warmup Decay Scheduler (기본값)
 - Warm-up: 0.2 (20% iterations of total)



3. 사용 모델기술

- 훈련용 하이퍼 파라미터(hyper-parameter)

- 학습 전략(training strategy)
 - 혼합 정밀도(mixed precision)
 - 사용 중인 NVIDIA Tesla V100 GPU는 컴퓨팅 능력(computing capacity)이 7.0 이상으로, cuda core로 인한 연산 가속화 가능
 - Multi-GPU 학습
 - 가용한 GPU 2개 모두를 훈련에 사용 → Trainer에서 가용한 모든 GPU 인식 및 자동으로 학습에 사용
 - Gradient Accumulation Steps
 - 미니 배치(mini-batch)에 대한 gradient를 일정 개수만큼 모아서 한 번에 손실 계산 및 모델 가중치 업데이트
 - 값을 8로 설정함으로써, 마치 배치 사이즈가 8배 커진 효과 → 훈련 속도 향상
 - $\therefore \text{총 배치 사이즈} = 16_{(\text{per replica batch size})} * 2_{(\# \text{GPUs})} * 8_{(\text{gradient accumulation steps})} = 256$



3. 사용 모델기술

- 추론용 하이퍼 파라미터

- 개요

- 추론 과정에서 사용되는 인자에 대해 설명
 - 시간 절약을 위해 훈련용 인자는 대부분 고정한 채 추론용 인자에 대해 다양하게 탐색

- 문장 생성 인자

- Beam Search Size = 5

- 현재 단어로부터 다음 단어를 추론할 때, 탐욕적(Greedy)으로 매 순간 최대의 확률값을 택하는 것이 아닌 일정 개수만큼의 후보군을 두어 이전의 잘못된 선택을 복구할 수 있도록 하는 소프트웨어 공학 기법

- 생성된 요약문의 최소 길이 = 64

- 생성된 요약문의 최대 길이 = 256

- 512 등 그 값이 너무 클 경우, 요약문이 복문으로 이루어질뿐더러 추론 시간이 지나치게 길어짐

- Length Penalty = [0.8 | 1.0 | 1.2]

- 생성된 요약문의 길이에 따라 패널티를 부여함으로써 그 길이가 짧거나 길게 만들어지도록 유도
 - 1보다 큰 값을 줄 경우 긴 문장을, 1보다 작은 값을 줄 경우 짧은 문장을 지향

- Trigram Blocking (no_repeat_ngram_size = 3)

- 동일 단어가 3번 이상 반복해서 등장하는 것을 방지



4. 모델링 결과

• 1차 평가 테스트 데이터에 대한 실험 결과 정리

구분	내용									
	(A)	(A)	(A)	(B)	(A)	(C)	(C)	(A)	(D)	
Model Name	(A)	(A)	(A)	(B)	(A)	(C)	(C)	(A)	(D)	
Clean	X	X	X	0	X	X	X	X	X	
Best/Ep. (*)	7/10 (7)	7/10 (7)	7/10 (7)	5/10 (5)	7/10 (7)	5/10 (5)	5/10 (5)	7/10 (10)	12/26 (12)	
Run Time (H)	10.7	10.7	10.7	10.7	10.7	7.9	7.9	10.7	20.5	
Train Loss	1.047	1.047	1.047	1.184	1.047	1.108	1.108	1.047	1.129	
Valid Loss	1.247	1.247	1.247	1.338	1.247	1.227	1.227	1.247	1.233	
LP	1.2	1.0	0.8	0.8	0.8	0.8	0.8	0.8	0.8	
Min/Max	64/256	64/256	64/256	64/256	64 \geq / \leq 256	64/256	128/512	64/256	64/256	
Rouge-1	0.3671	0.3672	<u>0.3678</u>	0.3640	0.3575	0.2846	0.3662	0.3446	0.3613	
Rouge-2	0.1801	0.1798	0.1803	<u>0.1830</u>	0.1760	0.1442	0.1811	0.1689	0.1771	
Rouge-N	0.2778	0.2785	<u>0.2800</u>	0.2711	0.2589	0.2048	0.2731	0.2522	0.2688	
Note				Variable Summary	Variable Padding	Variable Padding	Stop at 26/50			



4. 모델링 결과

• 1차 평가 테스트 데이터에 대한 실험 결과 정리

구분	내용				
Model Name	(E)	(E)	(A)	(A)	(A)
Clean	△	△	△	△	X
Best/Ep. (*)	7/10 (7)	7/10 (7)	7/10 (7)	7/10 (7)	7/10 (7)
Run Time (H)	7.8	7.8	10.7	10.7	10.7
Train Loss	1.045	1.045	1.047	1.047	1.047
Valid Loss	1.231	1.231	1.247	1.247	1.247
LP	0.8	0.8	0.8	0.5	0.6
Min/Max	64/256	64 \geq / \leq 256	64/256	64/256	64/256
Rouge-1	0.3634	0.3617	0.3439	0.3474	0.3469
Rouge-2	0.1781	0.1768	0.1660	0.1721	0.1718
Rouge-N	0.2701	0.2660	0.2480	0.2543	0.2538
Note		Variable Summary	Unigram Blocking		

• 용어 해설

• Clean

- 0: 민감정보 제거 및 언론사별 정제 수행
- △: 민감정보만 제거 (=최소한으로 수행)
- X: 텍스트 정제 작업 미적용

• Best/Ep. (*)

- Best: 검증 손실이 가장 좋았던 에폭
- Ep.: 에폭
- (*): 실제 추론에 사용한 가중치의 에폭

• LP: Length Penalty

• Min/Max

- Min: 생성된 요약문의 최소 길이
- Max: 생성된 요약문의 최대 길이

• Variable Summary

- 배치 단위 입력의 평균 길이에 따라 그 요약문의 최소/최대 길이를 10% 내외로 가변적 조정



```
min_length = max(64, int(avg_len_per_batch * 0.05))
max_length = min(256, int(avg_len_per_batch * 0.15))
```

4. 모델링 결과

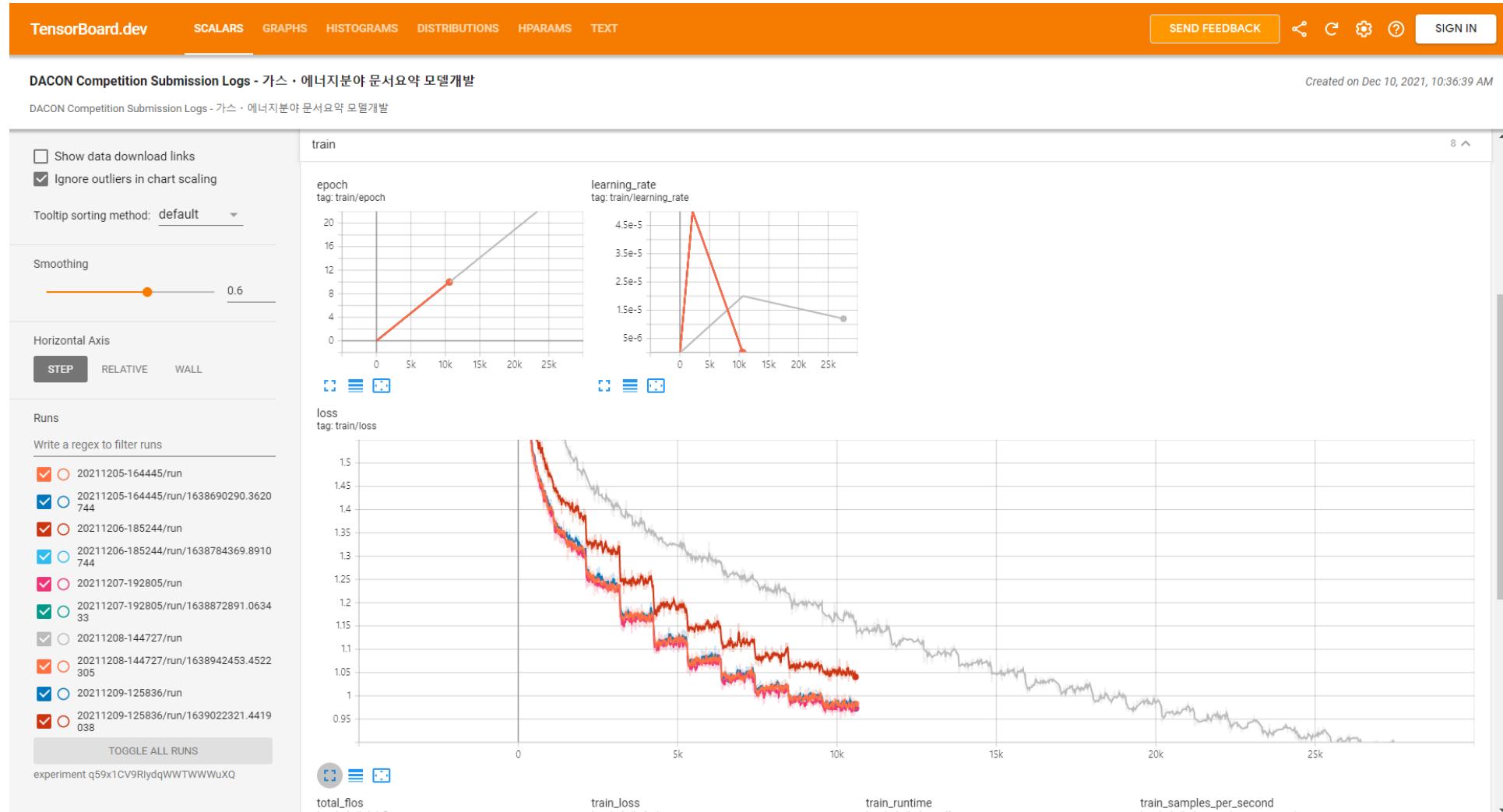
제3회 한국가스공사(KOGAS)

빅데이터·인공지능
스타트업 경진대회

가스·에너지분야 문서요약 모델개발



- 실험에 대한 모든 훈련 로그는 [TensorBoard.dev](#)에 공개적으로 게시





5. 특장점(차별성/우수성)

• 배치 단위 가변길이 패딩

• 설명

- 길이별로 정렬된 데이터에 대해, 고정된 최댓값 대신 배치 단위의 최대 크기만큼 패딩을 적용
- 패딩 추가를 최소화함으로써 입력 크기를 줄이고, 결과적으로 훈련 및 검증 속도를 향상시킴

• 예시

- 단일 배치에 다음과 같은 **3개의 문장**이 입력으로 들어왔으며, **최대 입력 크기는 15**라고 가정
 - 안녕하세요. 이야기연구소 주식회사 팀입니다.
 - 가스 · 에너지분야 문서요약 모델개발 대회
 - 날씨가 많이 추운데, 건강 유의하세요.
- 토크나이징 후 **[고정길이 패딩]**을 입힌 기존 결과는 다음과 같음(적색 음영은 garbage 값) → **입력 크기: (3, 15)**

Index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	_안녕하	세요.	_이야기	연구소	_주식	회사	_팀	입니다.	<PAD>						
1	_가스	.	에너지	분야	_문서	요	약	_모델	개발	_대회	<PAD>	<PAD>	<PAD>	<PAD>	<PAD>
2	_날씨가	_많이	_추	운데	_건강	_유의	하	세요.	<PAD>						



5. 특장점(차별성/우수성)

• 배치 단위 가변길이 패딩

- 설명

- 길이별로 정렬된 데이터에 대해, 고정된 최댓값 대신 배치 단위의 최대 크기만큼 패딩을 적용
- 패딩 추가를 최소화함으로써 입력 크기를 줄이고, 결과적으로 훈련 및 검증 속도를 향상시킴

- 예시

- 단일 배치에 다음과 같은 **3개의 문장**이 입력으로 들어왔으며, **최대 입력 크기는 15**라고 가정
 - 안녕하세요. 이야기연구소 주식회사 팀입니다.
 - 가스 · 에너지분야 문서요약 모델개발 대회
 - 날씨가 많이 추운데, 건강 유의하세요.
- 토크나이징 후 **[가변길이 패딩]**을 입힌 개선 결과는 다음과 같음(적색 음영은 garbage 값) → **입력 크기: (3, 10)**

Index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	_안녕하	세요.	_이야기	연구소	_주식	회사	_팀	입니다.	<PAD>	<PAD>	-	-	-	-	-
1	_가스	.	에너지	분야	_문서	요	약	_모델	개발	_대회	-	-	-	-	-
2	_날씨가	_많이	_추	운데	_건강	_유의	하	세요.	<PAD>	<PAD>	-	-	-	-	-



5. 특장점(차별성/우수성)

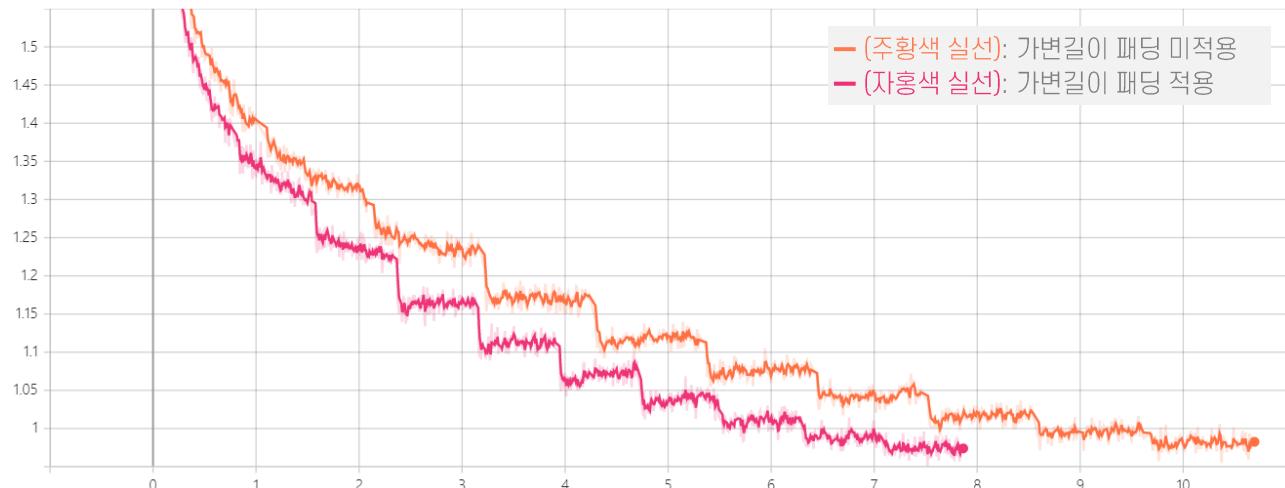
- 배치 단위 가변길이 패딩
 - 코드 구현



```
def _pad(self, sentences: List[List[int]], token_id: int) -> np.ndarray:
    ## We will pad as max length per batch, not "inp_max_len(=1024, etc)".
    max_length_per_batch = max([len(i) for i in sentences])

    ## Stack as dimension 0 (batch dimension).
    ## "token_id" can be "tokenizer.pad_token_id(=3)" or "ignore_index(=-100)"
    return np.stack([i + [token_id] * (max_length_per_batch - len(i)) for i in sentences], axis=0)
```

- 가변길이 패딩 적용 전/후 훈련 속도 비교 (10에폭 훈련 기준)
 - 적용 전: 10시간 41분
 - 적용 후: 7시간 51분
 - 훈련 시간의 약 25% 감소





6. 발전(사업화)계획

• 모델 고도화 계획

• 텍스트 정제 코드 개선

• 계획

- 연구적 목적이 아닌 상업화를 위해서는, 향후 비정형 데이터에 포함되어 있는 개인정보에 대해 적절한 가명처리가 요구됨
 - '이루다' 사건에 대한 개인정보위 결정의 의미와 시사점 (박광배 등, 2021.05)
- 더욱이, 훈련된 해당 모델에 대한 적대적 공격(adversarial attack)으로 인해 훈련에 사용된 데이터 추출/탈취 가능성 존재
 - [금보원2021-1Q] 전자금융과 금융보안 제23호 – 인공지능(AI) 기술의 보안 위협 및 대응 방안 (금융보안원, 2021.02)
 - 텍스트 기반 모델의 회원 추론 공격을 사전에 막기 위한 위험도 측정에 관한 연구 (나승호 등, 2021.11)

• 필요성

- 주어진 데이터에 대한 세부적인 검수 작업을 다시 진행 → 상업화 목적뿐 아니라, 성능상의 이점 또한 도모 가능

• 데이터 세트에 포함된 메타데이터의 추가 활용

• 필요성

- 데이터 구축 단계에서부터, 훈련 간 다양한 목적으로 활용될 수 있게끔 필수적인 원문 기사, 요약문뿐 아니라 부가 정보도 포함
 - 문서요약 텍스트 AI 데이터-비플라이소프트-데이터댐 구축 성과보고회 우수사례 (비플라이소프트, 2020.12)

• 계획

- [title] 정보를 활용하여, 입력 문장이 지나치게 긴 경우 중요도별로 순차 선별
 - 기존에 본문이 지나치게 길어질 경우 앞에서부터 일정 길이(1024자)만큼 자르는 단점을 보완할 수 있음 → 향후 사업화를 염두
 - 단, 문장 간 유사도를 계산하는 과정이 선행되어야 함
- [document_quality_scores] 정보를 활용하여, 요약문장별 손실에 가중치를 부여
 - 문서 점수는 4가지 항목에 대해 1~5점으로 구성: 가독성(readable), 정확성(accurate), 정보성(informative), 진실성(trustworthy)
 - 예측(요약)하기 쉬운 문장, 즉 문서의 평균 점수가 높은 신문기사는 더 큰 손실 패널티 부여 → 경제지 등
 - 예측(요약)하기 어려운 문장, 즉 문서의 평균 점수가 낮은 신문기사는 더 작은 손실 패널티 부여 → 모델하우스 분양 광고, 일기예보 등

6. 발전(사업화)계획

제3회 한국가스공사 (KOGAS)

빅데이터·인공지능
스타트업 경진대회

가스·에너지분야 문서요약 모델개발



• 사업화 계획: 콘텐츠 매니저 “BORA”

• 개요

- 개인이 수집한 웹 링크나 보유한 PDF 파일 등 **소장하고 있는 데이터에 대한 키워드와 요약 정보를 제공**
- 이로부터 일정 시간이 흐른 뒤 축적된 다양한 자료의 빠른 확인이 필요할 때 수월하게 데이터 확인 및 활용 보조

• 핵심 서비스 내용

• 데이터 요약 및 다중 문서 정리

- 사용자가 수집한 데이터(웹 링크, PDF 문서 등)를 기반으로 요약 데이터를 제공
→ 사용자가 원문을 직접 확인하지 않고도 그 정보에 대해 쉽게 접근 가능
- 수집이 끝난 비슷한 카테고리의 여러 데이터에 대해 관계 분석
→ 중복되는 내용을 제거하고, 합쳐진 데이터를 기준으로 요약 내용을 제공함으로써 보유한 자료에 대해 효율적인 관리 진행

• 데이터 미리보기

- 스크롤 다운과 같은 효과를 적용한 데이터 미리보기 제공
→ 원본을 확인하지 않고도 대략적인 이미지로 데이터 확인 가능
- 세부 내용의 확인이 필요한 경우, 별도의 설치 없이 본문 내용을 확인할 수 있는 뷰어(viewer) 제공 또는 원본 페이지 연결

• 키워드 추출 및 사용자 히스토리 분석

- 사용자가 보유 중인 데이터의 종류와 키워드를 미리 제공함으로써, 자신의 키워드를 평소에 인지 가능
- 검색어가 떠오르지 않거나 잊고 있던 키워드를 미리 상기시켜 줌으로써, 좀 더 효율적이고 원활한 검색 환경 제공
- 사용자가 보유하고 있지 않은 관련 자료들을 추천해주는 시스템을 마련
→ 동일 카테고리에서 미처 확인하지 못한 자료 검색 지원



6. 발전(사업화)계획

• 사업화 계획: 프로토 타입

검색 결과 (7 개) react

타일 구조:

- 스크린샷
- 문서 제목

검색 결과 (0 개) input search text

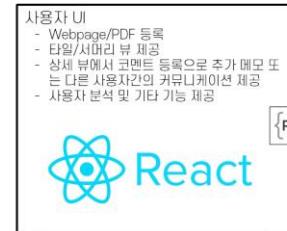
[MongoDB in Action] 6장 정리 :: 개발일지

정적 사이트 생성기 Gatsby :: Outsider's Dev Story

[React] velog, 해시태그 만들기!

기본 아키텍처

- Front-End: React
- Back-End: Flask, Node
- DB: MongoDB



서머리 구조:

- 스크린샷
- 문서 제목
- 요약문
- 카테고리&키워드
- 등록일 등

6. 발전(사업화)계획

제3회 한국가스공사 (KOGAS)

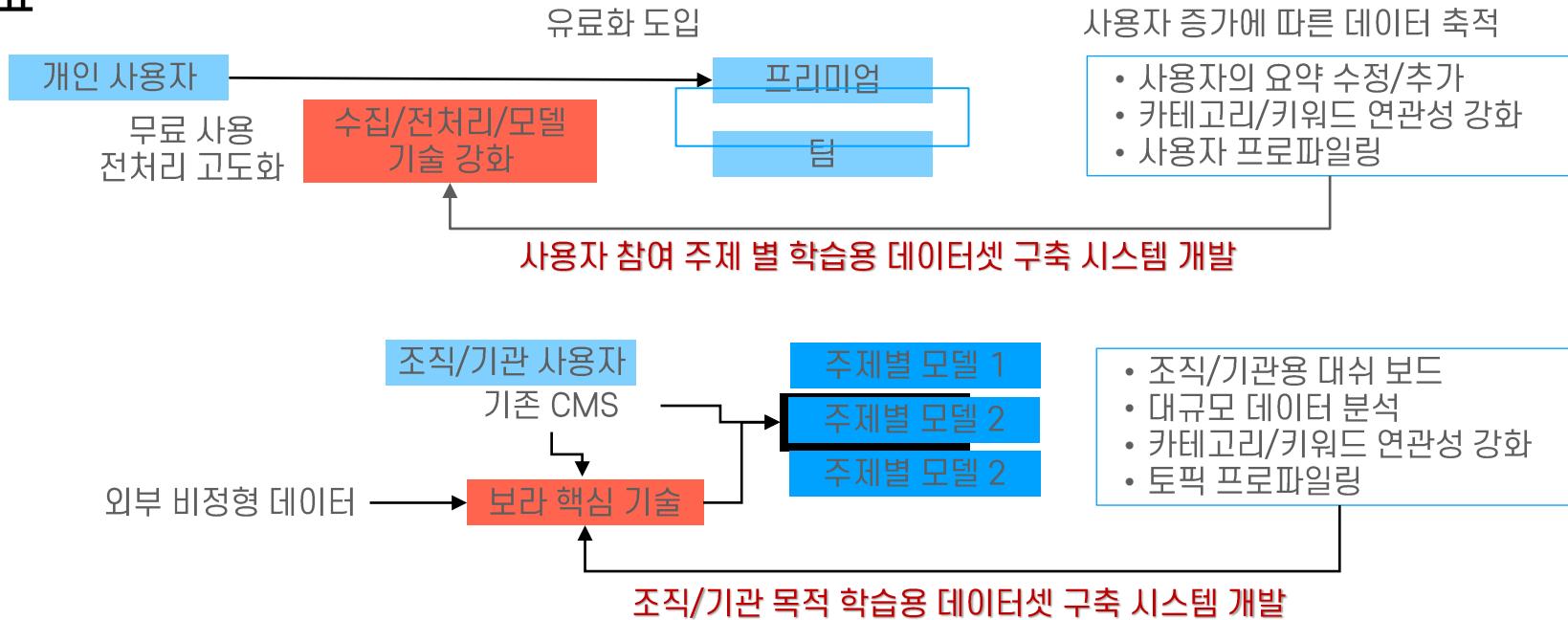
빅데이터·인공지능
스타트업 경진대회

가스·에너지분야 문서요약 모델개발



• 사업화 계획: 비즈니스 모델 & 로드맵

• 개요



• 기대 효과

- 웹링크와 같은 외부 데이터의 효과적인 관리 및 전문 자료 정리
- 자연어 처리 모델 발전에 대비한 훈련 데이터 축적 및 정제
- 사용자 및 토픽 관련 프로파일링으로 흐름과 변화에 대한 인사이트 제공
- 새로운 정보 검색 방법 제안

7. 결론(정리)

제3회 한국가스공사(KOGAS)

빅데이터·인공지능
스타트업 경진대회

가스·에너지분야 문서요약 모델개발



• 결언

- 당해 팀은 AI Hub의 [문서요약 텍스트] 데이터 세트를 활용하여 모델 훈련을 진행
 - 이외에 사용한 외부 데이터 없음
- 다양한 사전학습 모델 중 gogamza의 kobart-base-v1 모델을 미세 조정하여 사용
- 당사의 경진대회 1차 평가 테스트셋에 대한 최종 순위 및 결과는 다음과 같음
 - [최고점수] ROUGE-1: 0.3678 / ROUGE-2: 0.1803 / ROUGE-N: 0.2800
 - [최종선택] ROUGE-1: 0.3662 / ROUGE-2: 0.1811 / ROUGE-N: 0.2731
- 본 팀 및 구현 모델에 대한 특장점은 다음과 같음
 - 훈련 및 검증 단계에서 가변길이 패딩을 도입함으로써, 속도 면에서 이점 보유
 - 훈련 시간: 10시간 41분 → 7시간 51분 (약 25% 감소)
 - 검증 시간: 180초/에폭 → 96초/에폭 (약 47% 감소)
 - 구체적인 고도화 및 사업화 계획이 수립되어 있어 향후 발전 가능성 충분
 - 텍스트 정제 과정을 개선하고, 데이터 세트의 메타정보를 활용하여 학습 고도화 예정
 - 웹 페이지 및 문서(PDF 등)에 대한 요약 기능을 제공하는 콘텐츠 매니저 “BORA”의 프로토타입 제작 중

8. 팀원 역할 및 참여도

제3회 한국가스공사 (KOGAS)

빅데이터·인공지능
스타트업 경진대회

가스·에너지분야 문서요약 모델개발



• 이야기연구소 주식회사(스타트업 파트)

• 팀 구성 및 참여 내용

• 오명교(팀장)

- 이야기연구소 주식회사 연구원
- 딥러닝 기반 자연어 생성 요약 모델 개발

• 박인현(팀원)

- 이야기연구소 주식회사 대표
- 개발 방향 검토 및 향후 사업화 방안 수립

• 팀 역량

• 오명교(팀장)

- 백엔드 및 웹 크롤링/전처리 엔진 개발
- 딥러닝 기반 자연어 처리 모델 연구 및 개발
- (2020.09) [컴퓨터 비전 학습 경진대회](#) - 1위
- (2021.11) [제5회 금융보안원 논문공모전](#) - 우수상 (금융보안원 원장상)

• 박인현(팀원)

- KITRI 차세대 보안리더 양성 프로그램 멘토
- 안산대 인공지능소프트웨어학과 겸임교수
- (前) 더존비즈온 포렌식센터 컨설팅 팀장
 - 국내외 디지털 포렌식 및 E-Discovery 컨설팅 및 구축

9. 감사의 말

제3회 한국가스공사 (KOGAS)

빅데이터·인공지능
스타트업 경진대회

가스·에너지분야 문서요약 모델개발



본 연구 및 개발은 정보통신산업진흥원에서 진행한
“2021년 인공지능 고성능 컴퓨팅 지원 사업”에 선정되어 지원받은
컴퓨팅 자원을 이용하여 수행되었음을 밝힙니다.

Appendix A. 시행착오



• 추론 실패

• 개요

- 결측치 미처리로 인해 줄($line$)이 밀린 채 학습을 진행하여, 모델 추론 결과값이 모든 입력에 대해 유사하게 반복

• 발생 원인

- 훈련용 데이터세트의 사용 간 요약문(abstractive answer)이 존재하지 않는 결측치를 처리하지 않음
 - [id=362852732] 등
 - 이로부터 {신문기사, 요약문} 쌍(pair)이 불일치하여 잘못된 정답이 맵핑(mapping)
 - 개인적으로는, 모델이 입력에 대한 출력의 손실을 최소화하기 위해 모든 입력에 대해 그 생성 문장(단어)을 자명(trivial)하게, 즉 동일하게 뱉는 쪽으로 학습이 이루어졌다고 추측 중

• 결과

• 문장 별 동일한 단어 반복

- “당진”, “농업인” 등

- 단일 문장 내 동일한 어구 반복

- 실습, 실습, 실습, 실습, ...

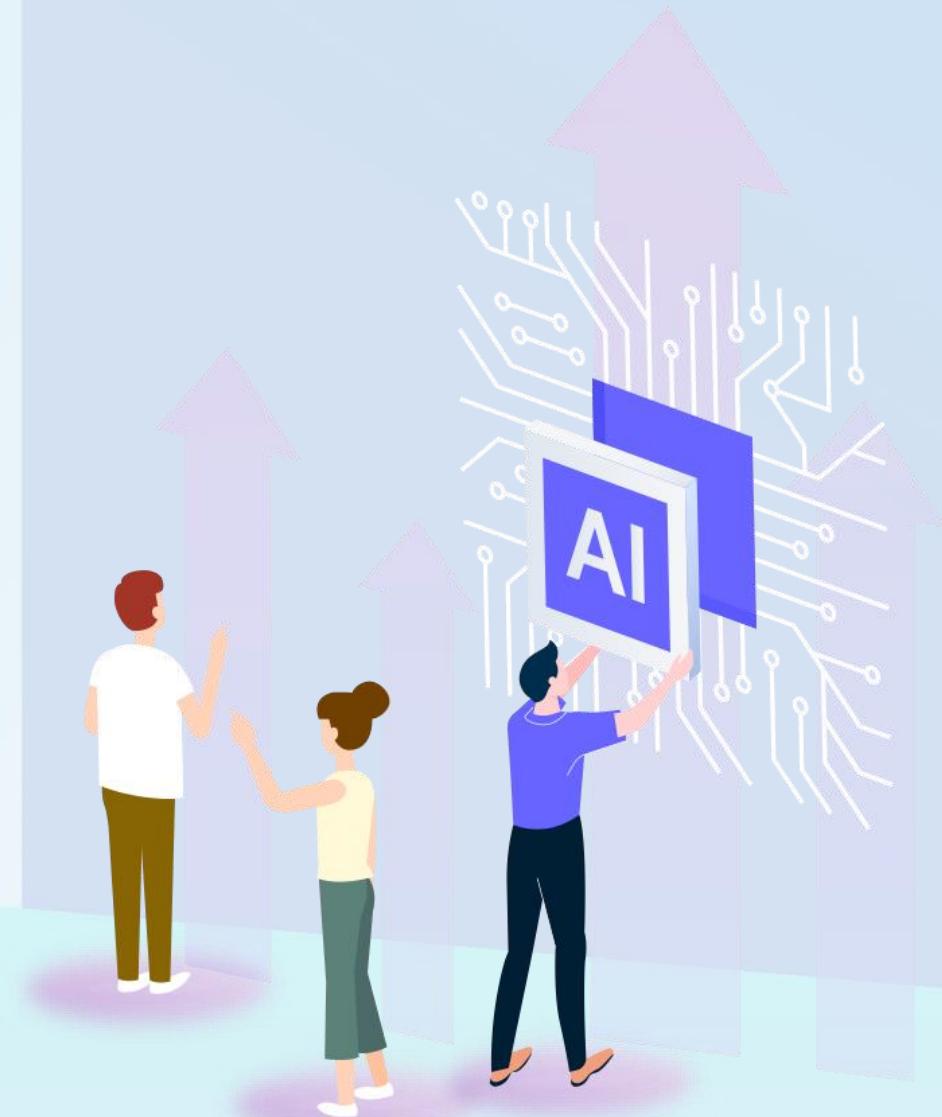
▶ 공개 평가용 테스트 데이터 세트에 대한
추론 결과에서 동일 단어 및 문구 반복현상 발생

제3회 한국가스공사 (KOGAS)

빅데이터·인공지능 스타트업 경진대회

가스·에너지분야 문서요약 모델개발

Q&A





제3회 한국가스공사 (KOGAS)

빅데이터·인공지능
스타트업 경진대회

감사합니다