참을 수 없는
GPT의 묵직함.

김영빈 김준영 김현수

Step 1 > Step 2 > Step 3 > Step 4

검증 단계	탐색 단계	학습 단계	글쓰기 단계
모델과 데이터의 유효성을 검증	수필 생성을 위한 최적의 디코딩 전략, 훈련 전략 수립	탐색 단계에서 수립한 전략을 기반으로 모델 학습	주제에 맞는 수필 작성

Step 1 > Step 2 > Step 3 > Step 4

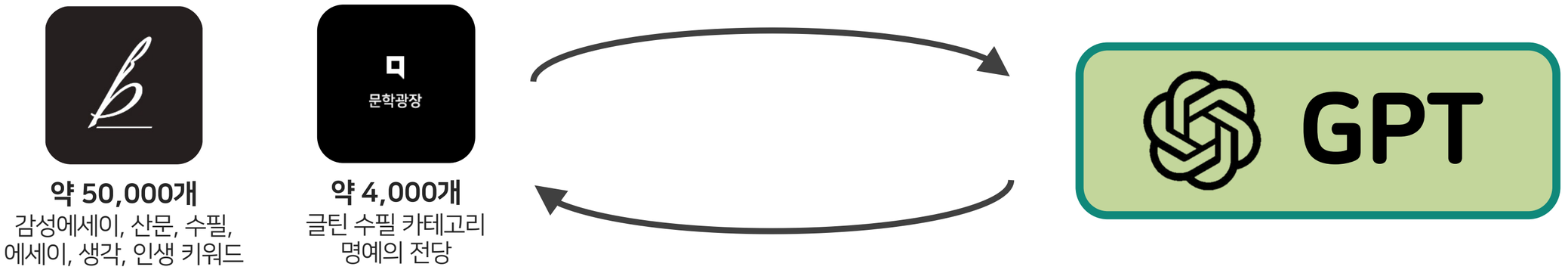
검증 단계	탐색 단계	학습 단계	글쓰기 단계

모델과 데이터의
유효성을 검증

STEP 01: 검증 단계

데이터의 수집과, 이를 모델에 훈련하는 과정을 빠르게 반복

모델이 데이터를 제대로 학습하고 있는가?
→ 모델 train/test 코드 유효성 검증



약 50,000개
감성에세이, 산문, 수필,
에세이, 생각, 인생 키워드

약 4,000개
글틴 수필 카테고리
명예의 전당

데이터의 어떠한 문제점이 모델의 성능을 저하하는가?
→ 정제된 데이터셋

STEP 01: 검증 단계

Data specific한 전처리

포스팅 단위 전처리

1. 한글 비율이 일정 이하인 데이터 제거
2. 특정 키워드(브런치, 코로나 등) 포함 데이터 제거
3. 특정 저자의 데이터 제거
4. 중국어, 일본어 포함 데이터 제거
5. 길이가 너무 짧은 데이터 제거

문장 단위 전처리

1. Email, URL, hashtag(#), mention(@) 등 제거
2. 일반적인 형태의 문장 부호로 치환 (e.g “→ ”)
3. 괄호 제거
4. 자음 혹은 모음만 있는 문자 제거
5. 반복되는 특수문자와 공백 한 개로 대치
6. 개행문자 제거

STEP 01: 검증 단계

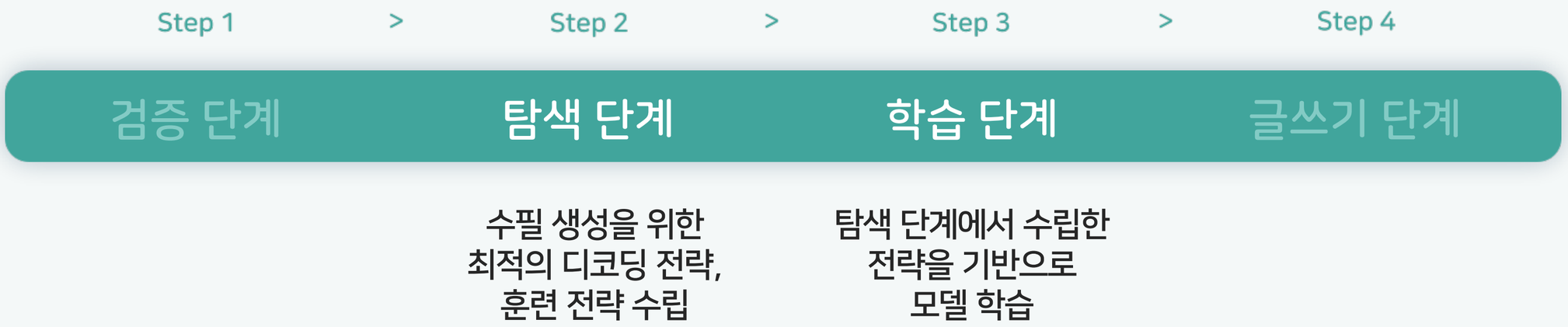
Data Specific한 전처리

문장 단위 전처리 전

아ㅏㅏㅏㅏㅏㅋㅋㅋㅋㅋㅋㅋㅋ 이거 너무 웃긴데?????
북커톤 진짜 재밌다!!!! (리얼루 (완전)) [찐좋음 {반박불가] (안 닫히는 괄호도 있음
{무슨 일이 있을지 모르니까]
지피티가 글을 잘 쓸까....??!#!!0$4%S#?!!!!---------
https://brunch.co.kr/@brunch/1
https://example.com
#지피티를 #사랑하는 @junieberry junyoung44@g.skku.edu
skku.edu
blog.naver.com/junieberry/222222222222

문장 단위 전처리 후

야 이거 너무 웃긴데? 북커톤 진짜 재밌다!
지피티가 글을 잘 쓸까...!


Step 1
>
Step 2
>
Step 3
>
Step 4
검증 단계
탐색 단계
학습 단계
글쓰기 단계
수필 생성을 위한
최적의 디코딩 전략,
훈련 전략 수립
탐색 단계에서 수립한
전략을 기반으로
모델 학습

STEP 02,03: 탐색 및 학습 단계

문제 1

Open-ended Generation

자연스럽고 유의미한 문장을
생성하는 방법이 뭘까?

문제 2

20,000자의 분량

긴 문장을 생성하는 동안 어떻게
문맥을 유지할 수 있을까?

문제 3

한정된 리소스

어떻게 하면 16GB T4에서
효과적으로 학습할 수 있을까?

STEP 02,03: 탐색 및 학습 단계

$$x_t = \arg\max_{v \in V^{(k)}} \left\{ (1-\alpha) \times \underbrace{p_\theta(v|\boldsymbol{x}_{<t})}_{\text{model confidence}} - \alpha \times \underbrace{(\max\{s(h_v, h_{x_j}) : 1 \le j \le t-1\})}_{\text{degeneration penalty}} \right\},$$

1. 자연스럽고 유의미한 문장을 생성하는 방법은 무엇일까?

Beam search

담대한 마음을 가져야 한다. 중요한 것은 꺾이지 않는 마음이다. 중요한 것은 꺾이지 않는 마음이다. 중요한 것은 꺾이지 않는 마음이다.

동일한 문장이 생성되는 동어 반복 문제 발생

Nucleus (top-p)

담대한 마음을 가져야 한다. 또한, 굳건한 마음도 가져야 한다. 사람의 마음이 곧 몸을 움직이기 때문이다. 그리고 나약한 마음을 가져야 한다.

때로 **의미적으로 일관되지 않은 문장**을 생성할 수 있음

Contrastive Search

담대한 마음을 가져야 한다. 담대하지 못한 사람은 실패를 두려워하며, 넘어져도 다시 일어날 수 있는 추진력이 없다.

문맥을 유지하면서도 자연스러운 흐름을 가지는 문장 생성

Su, Y., Lan, T., Wang, Y., Yogatama, D., Kong, L., & Collier, N. (2022). A Contrastive Framework for Neural Text Generation. *arXiv preprint arXiv:2202.06417*.
Su, Yixuan, and Nigel Collier. "Contrastive search is what you need for neural text generation." *arXiv preprint arXiv:2210.14140* (2022).

STEP 02,03: 탐색 및 학습 단계

2. 긴 문장을 생성하는 동안 어떻게 문맥을 유지할 수 있을까?

Sliding Window

행복해지는 방법에 대해 생각해본다. 나는 행복한가?라는
물음을 달고 산다. 행복하지 않은 척, 행복한 척을 하면서
살아가고 있다. 사실 그 모습이 진짜 행복은 아닐지도 모른다.
그저 내 머릿속에서만 맴돌고 있는 거겠지라고 생각하면서
오늘도 하루하루를 버티고 있다. 오늘은 어제보다 나은 삶을
살아야겠다는 생각이 들었지만 오늘이 내일보다 낫지
않을까라는 막연함에 사로잡혀 아무것도 하지 않고 멍하니
창밖을 바라보기만 했다. 그러다 어느 순간 깨달았다. 내가
원하는 삶이 무엇일까 생각해 본 적이 없다는 것을.

전체 문서를 학습할 수 있도록
문서를 정해진 길이만큼 잘라
학습 진행

Input - Target

행복해지는 방법에 대해 생각해본다. 나는
행복한가?라는 물음을 달고 산다. 행복하지
않은 척, 행복한 척을 하면서 살아가고 있다.
사실 그 모습이 진짜 행복은 아닐지도 모른다.
그저 내 머릿속에서만 맴돌고 있는 거겠지라고
생각하면서 오늘도 하루하루를 버티고 있다.
오늘은 어제보다 나은 삶을 살아야겠다는
생각이 들었지만 오늘이 내일보다 낫지
않을까라는 막연함에 사로잡혀 아무것도 하지
않고 멍하니 창밖을 바라보기만 했다.

Prompt

생성

앞선 문장을 프롬프트로 활용

Keyword

생각 오늘 행복</s>
행복해지는 방법에 대해 생각해본다. 나는 행복한가?라는
물음을 달고 산다. **행복**하지 않은 척, 행복한 척을 하면서
살아가고 있다. 사실 그 모습이 진짜 행복은 아닐지도 모른다.
그저 내 머릿속에서만 맴돌고 있는 거겠지라고 생각하면서
오늘도 하루하루를 버티고 있다. **오늘**은 어제보다 나은 삶을
살아야겠다는 생각이 들었지만 오늘이 내일보다 낫지
않을까라는 막연함에 사로잡혀 아무것도 하지 않고 멍하니
창밖을 바라보기만 했다. 그러다 어느 순간 깨달았다. 내가
원하는 삶이 무엇일까 **생각**해 본 적이 없다는 것을.

KR-WordRank를 활용해
문서의 키워드를 추출 후
프롬프트로 활용

skt/kogpt2-base-v2
125M parameters
Transformer Decoder 기반

STEP 02,03: 탐색 및 학습 단계

2.1 실험 과정

다양한 기법을 효율적으로 실험하기 위해 **크기가 작은 skt/kogpt2-base-v2 모델에 실험 진행**

- Metric
- Human evaluation

	rep-2 ↓	rep-3 ↓	rep-4 ↓	Diversity ↑	Keyword Recall ↑	Coherence ↑	Perplexity ↓
Baseline	3.3572	1.8210	1.4898	0.9520	-	0.5141	35.917
Keyword	**1.9189**	**1.2608**	**1.1378**	**0.9707**	**0.9475**	0.6803	25.600
Input – Target	3.7116	2.1753	1.8742	0.9440	-	0.3531	**21.548**
Keyword + Input – Target	2.3295	1.3471	1.2223	0.9639	0.9470	**0.6923**	26.238

STEP 02,03: 탐색 및 학습 단계

3. 어떻게 하면 16GB GPU에서 효과적으로 학습할 수 있을까?

skt/kogpt2-base-v2
125M parameters
Transformer Decoder 기반

KETI-AIR/ke-t5-base-ko
770M parameters
Transformer Encoder-Decoder 기반

skt/ko-gpt-trinity-1.2B-v0.5
1.2B parameters
Transformer Decoder 기반

1. Automatic Mixed Precision(AMP)

   float16을 사용하여 계산 과정에서 필요한 메모리 사용량 감소 & 속도 증가

2. Gradient Accumulation

   한정된 GPU 메모리에서 큰 batch size의 효과를 낼 수 있도록 gradient를 중첩시킴

STEP 02,03: 탐색 및 학습 단계

Fine tuning 단계

디코딩 전략

Contrastive Search

문맥을 유지하면서도 자연스러운 흐름을 가지는 문장 생성

학습 전략

Keyword

KR-WordRank를 활용해 문서의 키워드를 추출 후 프롬프트로 활용

모델

skt/ko-gpt-trinity-1.2B-v0.5

1.2B parameters
Transformer Decoder 기반

- Batch size 1, 최대 384 토큰까지 학습
- ~200,000step/epoch, 2 epoch, 16h/epoch
- 2epoch 학습 후 모델의 valid loss가 증가하기 시작하여 early stop

Step 1 > Step 2 > Step 3 > Step 4

검증 단계 탐색 단계 학습 단계 **글쓰기 단계**

주제에 맞는 수필 작성

담대한 (膽大) : 겁이 없고 배짱이 두둑하다.

Fine-fine-tuning

- 주제어인 “담대하다”을 학습할 수 있도록
- 브런치에서 “담대” 키워드를 포함하는 문서 약 700개의 문서를 추가로 크롤링하여 학습

STEP 04: 글쓰기 단계

생성 전략

1. 앞에 생성된 문장에서부터 **키워드를 추출**하여, 추가 프롬프트와 함께 제공함
2. 인간의 직접적인 문장 수정을 최소화하고, **키워드를 추가하는 방식으로 모델이 글을 쓰는 방향을 조절**함
3. 모델이 문장을 이어서 작성할 수 있도록 작성된 문서의 마지막 n개 문장을 제공함

이전 문장

나는 나에 대해서 공부하기로 마음 먹었다. 내 마음가짐에 따라 달라지는 나를 알 수 있는 방법이 무엇일까 궁금해서 찾아본 방법이라고 생각이 드니 실천해보자는 결심을 하게 된 계기이기도 하다. 내가 어떤 사람인지 알아보기 시작했다는 뜻이지 않을까 싶기도 하다. 그래서 난 늘 나에게 물었고 답하기를 주저하지 않았으면 좋겠다. 나를 알아야 다른 사람도 알 수 있고 더 나아갈 수 있다고 믿기 때문이다. 나란 존재에 관심을 갖기 시작할 때 나는 스스로 답을 내리지 못할 때가 많아지곤 했다. 그 질문에 대한 답을 찾는 과정이 쉽지는 않았지만 나는 결국 답을 찾으려 애썼던 것 같다.

키워드 요약

마지막 문장 입력

생성된 문장

마음 나 결심</s> 나란 존재에 관심을 갖기 시작할 때 나는 스스로 답을 내리지 못할 때가 많아지곤 했다. 그 질문에 대한 답을 찾는 과정이 쉽지는 않았지만 나는 결국 답을 찾으려 애썼던 것 같다. 그렇게 고민과 방황의 시간을 보내고 어느 날 거울에 비친 내 모습을 보며 깨달았다. 나는 실패를 두려워하는 사람이다. 나는 지금껏 실패를 두려워하기만 했다. 언제나 안정만을 추구하고, 도전을 두려워했다. 그리고 어느 순간 내 시야는 좁아졌고, 두려움은 공포가 되었다. 두려움이라는 감정은 무엇일까? 답은 간단하다. '실패'일 뿐이다.

행복으로의 내디딤

화자는 부정적인 생각에서 벗어나고 진취적인 삶을 살기 위해

자신을 성찰하면서 행복해지는 삶의 방향과 방법에 대해 고민하고,

행복을 위해 한 발자국 내딛기로 마음 먹는다.

1. 도전은 두려운 일이 아니다
2. 내 안의 불안함을 포용하자
3. 실패에 대한 두려움을 극복하자
4. 두려움을 극복하기 위해 행동하자
5. 글을 쓰자
6. 위기를 기회로

오늘도 어영부영 버티고 있는 나에게 박수를 보내고 싶다. 이 글을 읽는 당신도 그러길 바라면서. 담대하게, 포기하지 않길 바라며 응원한다.

참을 수 없는 GPT의 묵직함

Thank you.

Appendix

Metric 설명

02 문제 해결 프로세스

Step 1	>	Step 2	>	Step 3	>	Step 4
검증 단계		탐색 단계		학습 단계		글쓰기 단계
모델과 데이터의 유효성을 검증		수필 생성을 위한 최적의 디코딩 전략, 훈련 전략 수립		탐색 단계에서 수립한 전략을 기반으로 모델 학습		주제에 맞는 수필 작성

Contents

01

01 문제 정의 및 솔루션