트랜스포머를 활용한 자연어 처리

5장 - 텍스트 생성

6장 - 요약

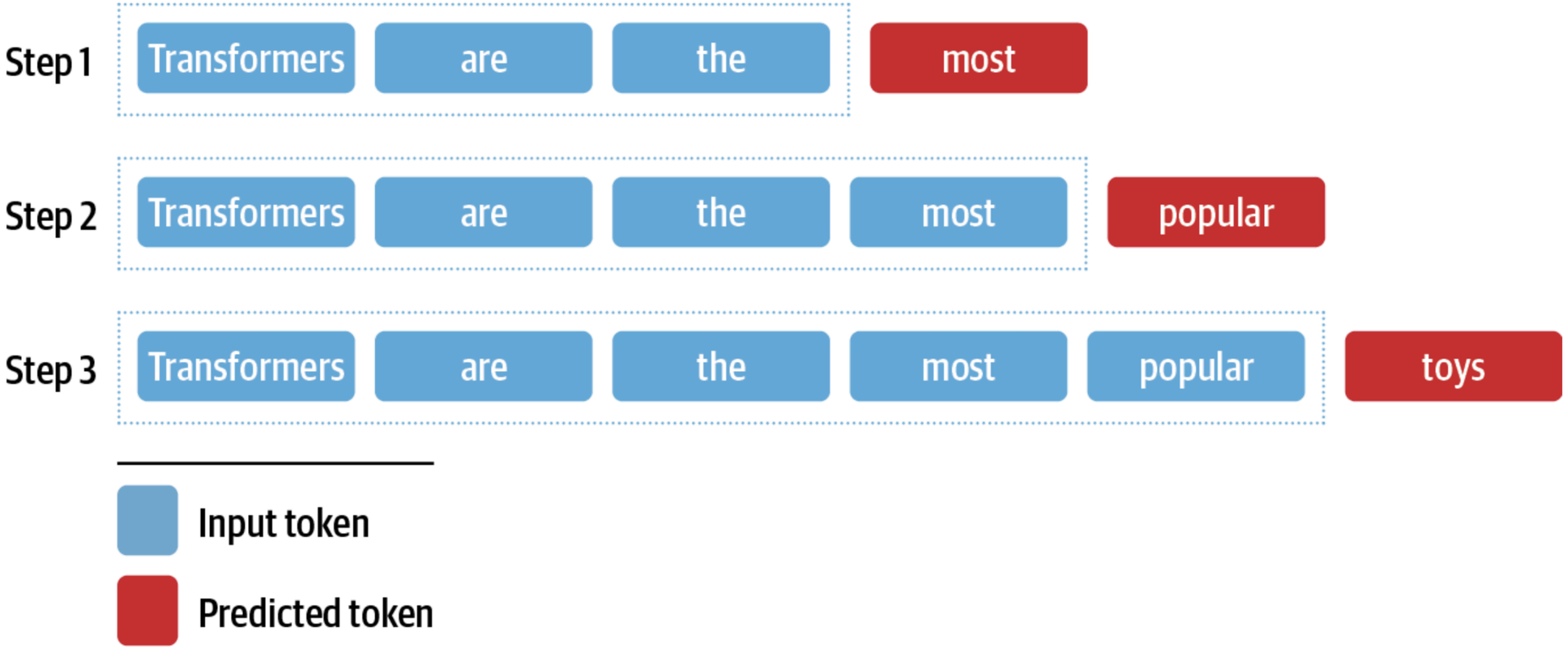
2024.7.22 최종현

5장 - 텍스트 생성

텍스트 생성

- **디코딩**을 통해 모델의 확률 출력을 텍스트로 변환
- GPT-2 모델
  - 문맥 시퀀스 $x = x_1, x_2, \ldots x_k$ 가 주어질 때 텍스트에 등장하는 $y = y_1, y_2, \ldots y_t$ 의 확률 $P(y \mid x)$를 추정 (확률의 연쇄법칙을 사용한 조건부 확률의 곱)
  - $$P(y_1, \ldots, y_t \mid x) = \prod_{t=1}^{N} P(y_t \mid y_{<t}, x)$$

텍스트 생성



조건부 텍스트 생성

텍스트 생성

- 디코딩 방법을 통해 타임스텝에서 어떤 토큰을 생성할지 결정함
- 소프트맥스 함수를 통해 다음 토큰 w_i 에 대한 확률 분포를 얻음
  $P(y_t = w_i \mid y_{<t}, x) = \text{softmax}(z_{t,i})$
- 디코딩 방법에서 $\hat{y}$ 을 선택해 확률이 가장 높은 시퀀스를 찾음 -> **근사적 방법 사용**
  $\hat{y} = \arg\max_{y} P(y \mid x)$

그리디 서치 디코딩

빔 서치 디코딩

텍스트 생성 - 그리디 서치 디코딩

- 각 타임 스텝에서 확률이 가장 높은 토큰을 선택
- $\hat{y}_t = \arg\max_{y_t} P(y_t \mid y_{<t}, x)$

```
model_name = 'gpt2-xl'
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name).to(device)
```

```
input_txt = "Transformers are the"
input_ids = tokenizer(input_txt, return_tensors='pt')["input_ids"].to(device)
iterations = []
n_steps = 8
choices_per_step = 5

with torch.no_grad():
    for _ in range(n_steps):
        iteration = dict()
        iteration["Input"] = tokenizer.decode(input_ids[0])
        output = model(input_ids=input_ids)

        # 첫 번째 배치의 마지막 토큰의 로짓을 선택해 소프트맥스를 적용
        next_token_logits = output.logits[0, -1, :]
        next_token_probs = torch.softmax(next_token_logits, dim=-1)
        sorted_ids = torch.argsort(next_token_probs, dim=-1, descending=True)

        # 가장 높은 확률의 토큰을 저장
        for choice_idx in range(choices_per_step):
            token_id = sorted_ids[choice_idx]
            token_prob = next_token_probs[token_id].cpu().numpy()
            token_choice = (f"{tokenizer.decode(token_id)} ({100 * token_prob:.2f}%)")
            iteration[f"Choice {choice_idx+1}"] = token_choice

        # 예측한 다음 토큰을 입력에 추가
        input_ids = torch.cat([input_ids, sorted_ids[None, 0, None]], dim=-1)
        iterations.append(iteration)

pd.DataFrame(iterations)
```

텍스트 생성 - 그리디 서치 디코딩

	Input	Choice 1	Choice 2	Choice 3	Choice 4	Choice 5
0	Transformers are the	most (8.53%)	only (4.96%)	best (4.65%)	Transformers (4.37%)	ultimate (2.16%)
1	Transformers are the most	popular (16.78%)	powerful (5.37%)	common (4.96%)	famous (3.72%)	successful (3.20%)
2	Transformers are the most popular	toy (10.63%)	toys (7.23%)	Transformers (6.60%)	of (5.46%)	and (3.76%)
3	Transformers are the most popular toy	line (34.38%)	in (18.20%)	of (11.71%)	brand (6.10%)	line (2.69%)
4	Transformers are the most popular toy line	in (46.29%)	of (15.09%)	, (4.94%)	on (4.40%)	ever (2.72%)
5	Transformers are the most popular toy line in	the (65.99%)	history (12.42%)	America (6.91%)	Japan (2.44%)	North (1.40%)
6	Transformers are the most popular toy line in the	world (69.27%)	United (4.55%)	history (4.29%)	US (4.23%)	U (2.30%)
7	Transformers are the most popular toy line in ...	, (39.73%)	. (30.64%)	and (9.87%)	with (2.32%)	today (1.74%)

텍스트 생성 - 그리디 서치 디코딩

```
max_length = 128
input_txt = """In a shocking finding, scientist discovered \
a herd of unicorns living in a remote, previously unexplored \
valley, in the Andes Mountains. Even more surprising to the \
researchers was the fact that the unicorns spoke perfect English.\n\n
"""
input_ids = tokenizer(input_txt, return_tensors='pt')["input_ids"].to(device)
output_greedy = model.generate(input_ids, max_length=max_length, do_sample=False)
print(tokenizer.decode(output_greedy[0]))
```

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The researchers, from the University of California, Davis, and the University of Colorado, Boulder, were conducting a study on the Andean cloud forest, which is home to the rare species of cloud forest trees.

The researchers were surprised to find that the unicorns were able to communicate with each other, and even with humans.

The researchers were surprised to find that the unicorns were able

반복적인 출력 시퀀스를 생성하는 경향

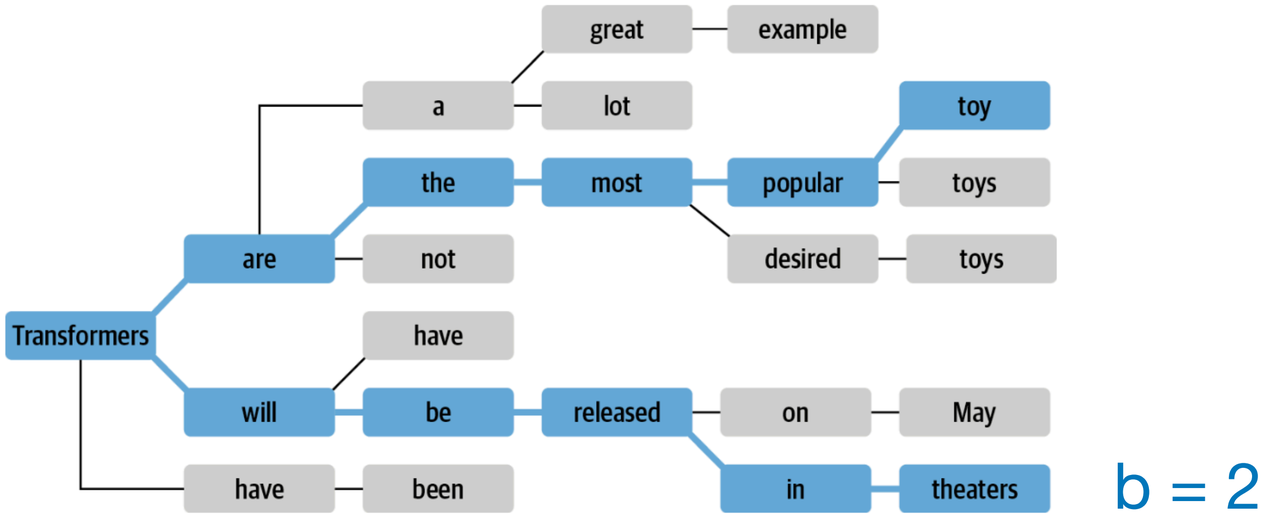
텍스트 생성 - 그리디 서치 디코딩

	Input	Choice 1
0	Transformers are the	most (8.53%)
1	Transformers are the most	popular (16.78%)
2	Transformers are the most popular	toy (10.63%)
3	Transformers are the most popular toy	line (34.38%)
4	Transformers are the most popular toy line	in (46.28%)
5	Transformers are the most popular toy line in	the (65.99%)
6	Transformers are the most popular toy line in the	world (69.26%)
7	Transformers are the most popular toy line in ...	, (39.73%)
8	Transformers are the most popular toy line in ...	and (32.51%)
9	Transformers are the most popular toy line in ...	the (11.70%)
10	Transformers are the most popular toy line in ...	Transformers (6.38%)
11	Transformers are the most popular toy line in ...	are (16.26%)
12	Transformers are the most popular toy line in ...	the (30.39%)
13	Transformers are the most popular toy line in ...	most (45.68%)
14	Transformers are the most popular toy line in ...	popular (62.52%)
15	Transformers are the most popular toy line in ...	toy (23.44%)
16	Transformers are the most popular toy line in ...	line (35.05%)
17	Transformers are the most popular toy line in ...	in (75.57%)
18	Transformers are the most popular toy line in ...	the (78.00%)
19	Transformers are the most popular toy line in ...	world (77.77%)
20	Transformers are the most popular toy line in ...	. (60.69%)
21	Transformers are the most popular toy line in ...	\n (9.01%)
22	Transformers are the most popular toy line in ...	\n (99.52%)
23	Transformers are the most popular toy line in ...	The (8.62%)
24	Transformers are the most popular toy line in ...	Transformers (11.21%)

“Transformers are the most popular toy line in the world” 반복

텍스트 생성 - 빔 서치 디코딩

- 확률이 가장 높은 상위 b개의 다음 토큰을 추적
- 최대 길이 혹은 EOS 토큰에 도달할 때까지 반복
- $$\log P(y_{1,\ldots,y_t} \mid x) = \sum_{t=1}^{N} \log P(y_t \mid y_{<t}, x)$$



b = 2

텍스트 생성 - 디코딩 방법 비교

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The researchers, from the University of California, Davis, and the University of Colorado, Boulder, were conducting a study on the Andean cloud forest, which is home to the rare species of cloud forest trees.

The researchers were surprised to find that the unicorns were able to communicate with each other, and even with humans.

The researchers were surprised to find that the unicorns were able

그리디 서치

로그 확률: -87.43

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The discovery of the unicorns was made by a team of scientists from the University of California, Santa Cruz, and the National Geographic Society.

The scientists were conducting a study of the Andes Mountains when they discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English

빔 서치

로그 확률: -55.23

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The discovery was made by a team of scientists from the University of California, Santa Cruz, and the National Geographic Society.

According to a press release, the scientists were conducting a survey of the area when they came across the herd. They were surprised to find that they were able to converse with the animals in English, even though they had never seen a unicorn in person before. The researchers were

빔 서치 (n-그램 페널티)

로그 확률: -93.12

텍스트 생성 - 샘플링

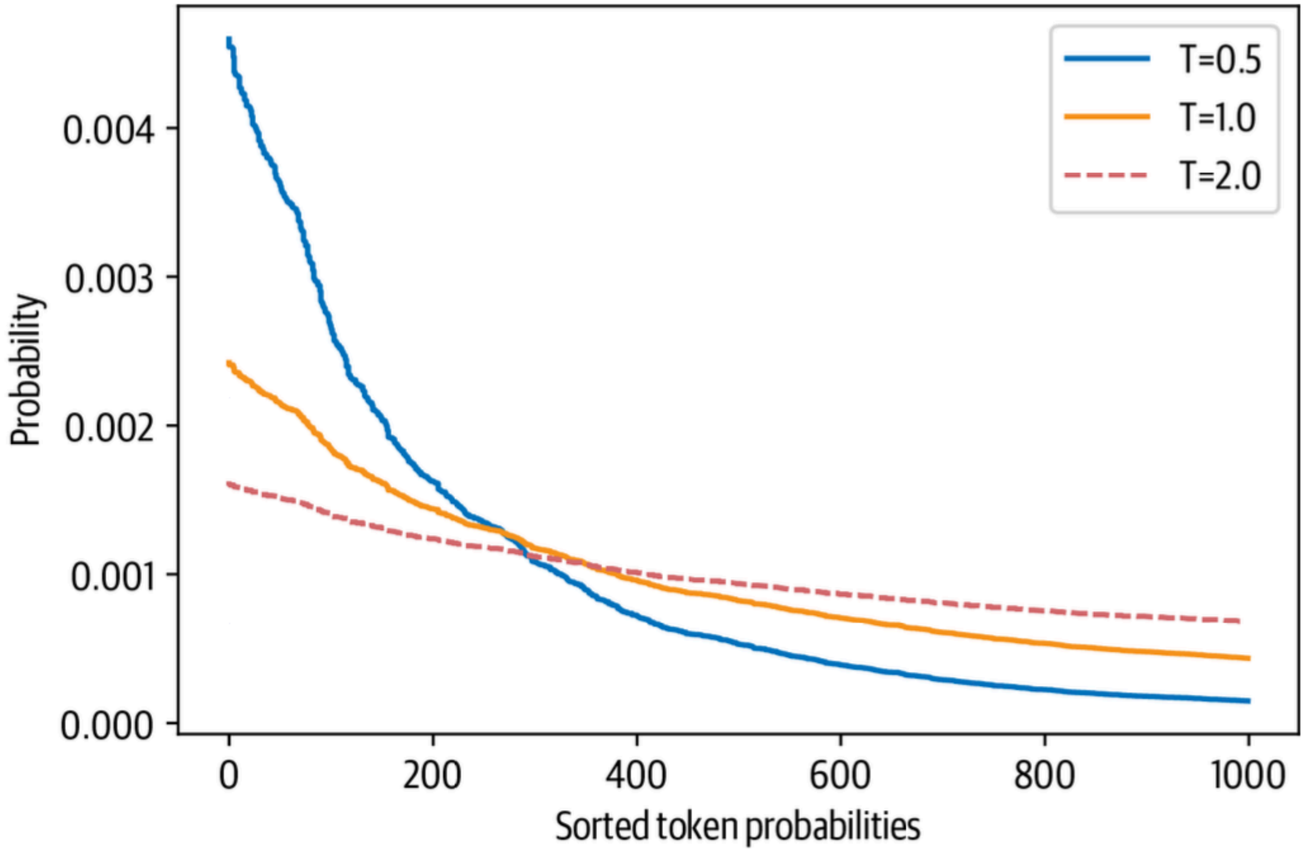
- 각 스텝의 모델에서의 어휘 사전 확률 분포에서 랜덤하게 샘플링

$$P(y_t = w_i \mid y < t, x) = \text{softmax}(z_{t,i}) = \frac{\exp(z_{t,i})}{\sum_{j=1}^{|V|} \exp(z_{t,j})}$$

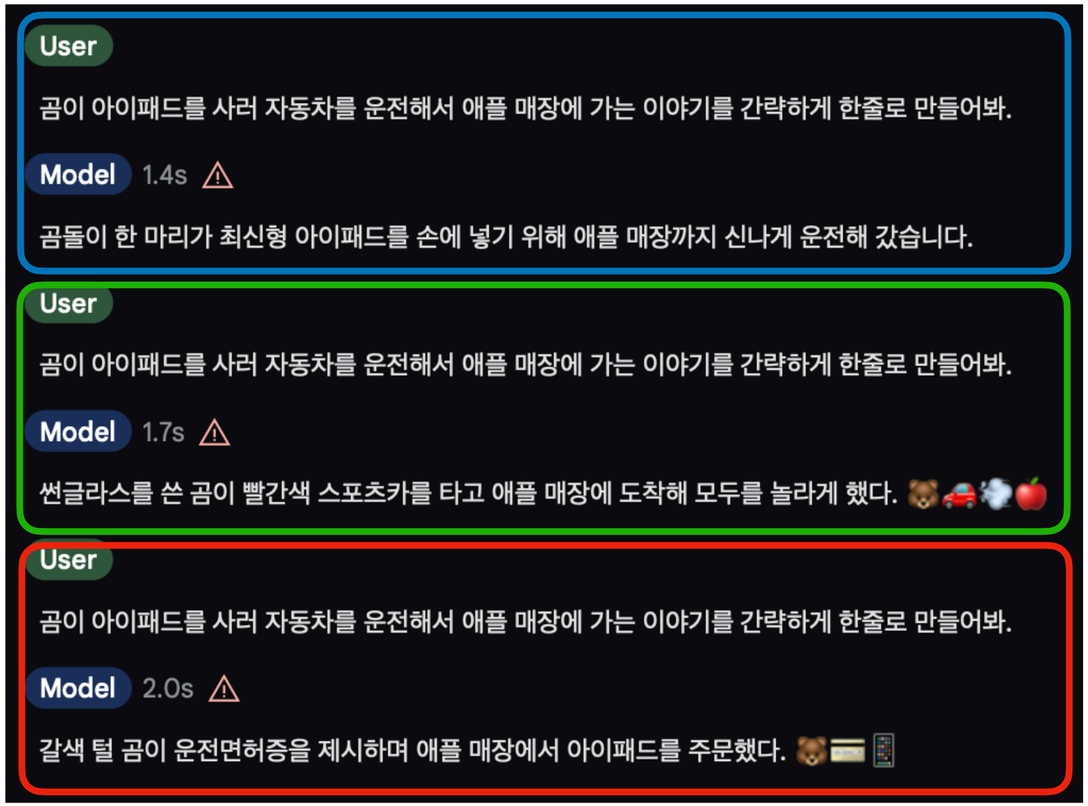
- 온도 파라미터를 추가하여 다양성 조정

$$P(y_t = w_i \mid y < t, x) = \text{softmax}(z_{t,i}) = \frac{\exp(z_{t,i}/T)}{\sum_{j=1}^{|V|} \exp(z_{t,j}/T)}$$

텍스트 생성 - 샘플링


T=0.5
T=1.0
T=2.0
0.004
0.003
0.002
0.001
0.000
Probability
0
200
400
600
800
1000
Sorted token probabilities

텍스트 생성 - 샘플링



T = 0.1

T = 1

T = 2

텍스트 생성 - 샘플링

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Beat Scores Post artworkesra 1968 lowquality460 144 Pirate murdering Hipp Hill corrupted explosion478 scalable export 450 Wind stagnFriendyl prison experimentsmax worksase adventurerhttp Andanto EU Leviathan contaminatur MDMA neurothereal ahead dumb immediatelyClose Stud I vaccsame incompatible shun complicate yesFile bbmrival farDepth convertible unaccompanied professionally psychedel curl chaotic were Ult Robots graphics witness Atmospherappropriately Warsaw Suit AkCast outcomes Tepl Prot

T = 2

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The herd of unicorns is an endangered species, but the researchers say that the creatures have been living in the valley for decades.

The researchers have been studying the valley for years, but have only recently been able to find the unicorns.

The unicorns are said to be a bright orange color, but the researchers have been unable to determine the exact color of the unicorns.

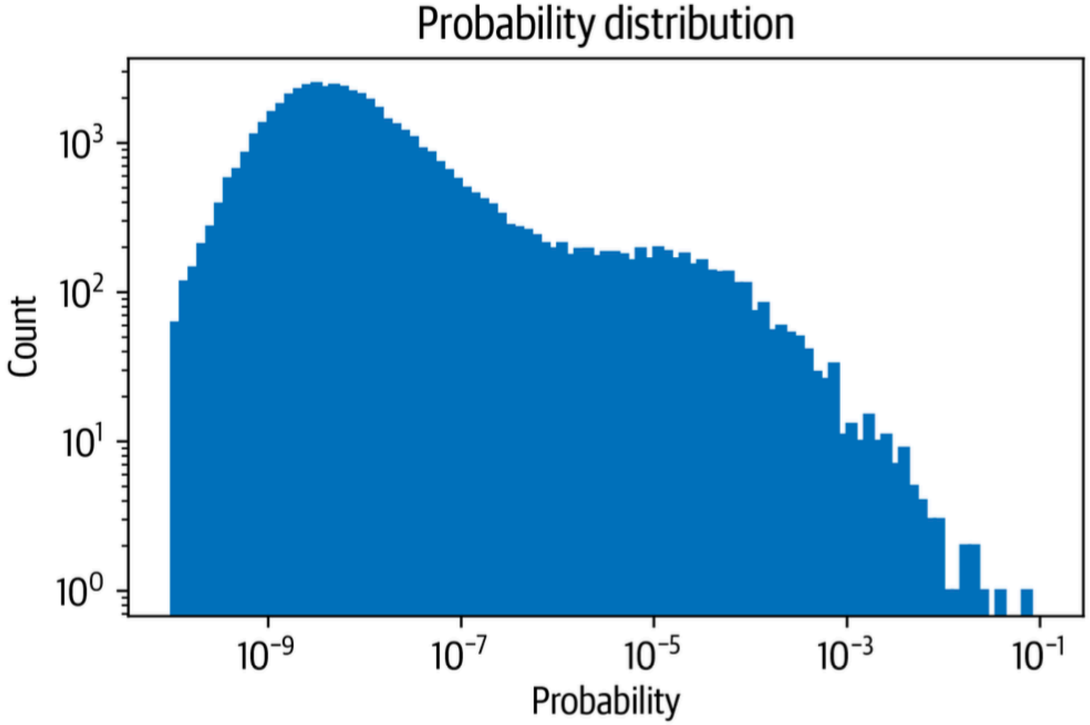
T = 0.5

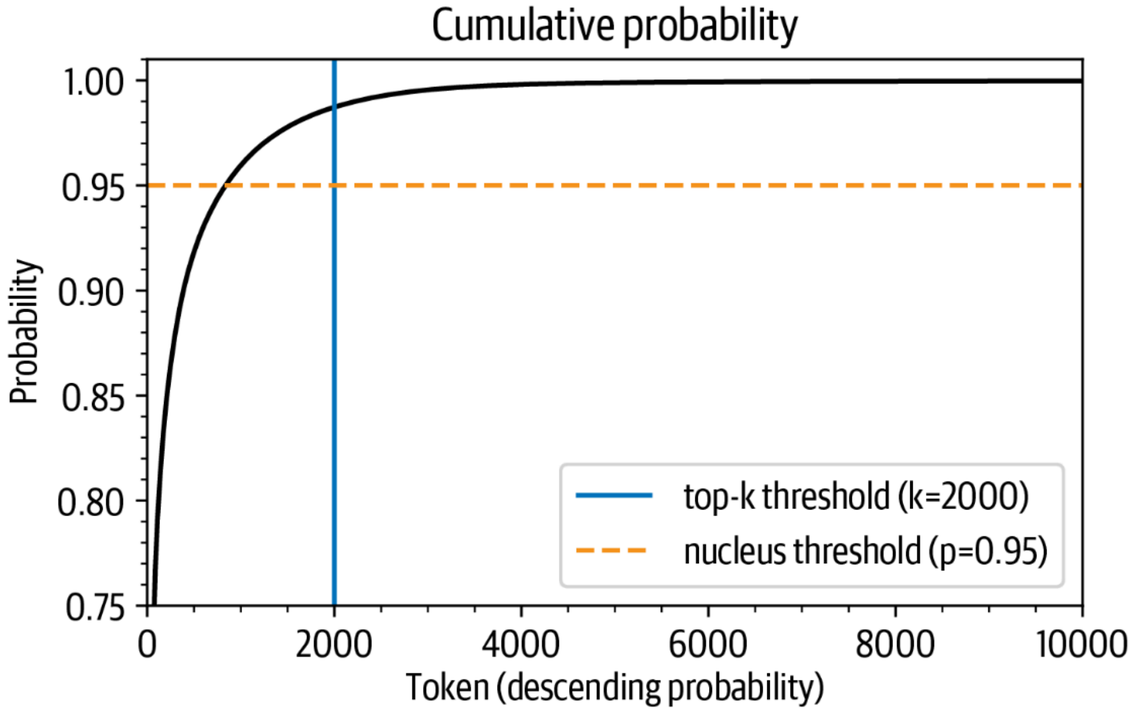
탑 - k 샘플링

뉴클리어스 샘플링 (탑-p)

텍스트 생성 - 탑-k, 뉴클리어스 (탑-p)

- 샘플링에서 사용할 토큰의 개수를 줄인다는 개념





텍스트 생성 - 탑-k, 뉴클리어스 (탑-p)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

How could such a rare creature exist? The reason for their existence was not just a result of random chance; there was a purpose. Because of the harsh environmental conditions, the unicorn was unable to travel to other areas; even if they did, the weather there was so unpredictable that other creatures would have been there to take advantage of its weakness.

The only reason to make such an attempt was for

```
output_topk = model.generate(input_ids, max_length=max_length, do_sample=True, top_k=50)
print(tokenizer.decode(output_topk[0]))
```

텍스트 생성 - 탑-k, 뉴클리어스 (탑-p)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

According to the study, the unicorns were spotted near the Chacaltaya mountains in the Peruvian Andes, and were identified by their distinctive hooves.

Professor Luis Chiriboga, who led the research expedition, is quoted as saying in an exclusive interview with MailOnline, "Our first reaction was that they must be a joke. They don't look like any of the herds

```
output_topp = model.generate(input_ids, max_length=max_length, do_sample=True, top_p=0.95)
print(tokenizer.decode(output_topp[0]))
```

6장 - 요약

요약

- 긴 단락 이해, 관련 내용 추론, 원래 문서의 주제를 통합해 텍스트를 생성
- 정교한 수준의 도메인 일반화 필요 (기사 요약과 법률 계약서 요약은 다름)
- 요약에는 인코더-디코더 트랜스포머가 적합
- 요약에 사용하는 CNN/DailyMail 말뭉치 데이터셋

요약 - CNN/DailyMail

Datasets: abisee/cnn_dailymail | like 196

Tasks: Summarization | Modalities: Text | Formats: parquet | Sub-tasks: news-articles-summarization | Languages: English | Size: 1

License: apache-2.0

Dataset card | Viewer | Files and versions | Community 10

Dataset Viewer — Auto-converted to Parquet | API | Embed | View in Dataset Viewer

Subset (3): 1.0.0 · 312k rows | Split (3): train · 287k rows

Search this dataset

article string · *lengths* 48 – 15.9k	highlights string · *lengths* 14 – 7.39k	id string · *lengths* 40 – 40
LONDON, England (Reuters) -- Harry Potter star Daniel Radcliffe gains access to a…	Harry Potter star Daniel Radcliffe gets £20M fortune as he turns 18 Monday . Young actor…	42c027e4ff9730fbb3de84c1af0d2c506e41c3e4
Editor's note: In our Behind the Scenes series, CNN correspondents share their…	Mentally ill inmates in Miami are housed on the "forgotten floor" Judge Steven Leifman…	ee8871b15c50d0db17b0179a6d2beab35065f1e9
MINNEAPOLIS, Minnesota (CNN) -- Drivers who were on the Minneapolis bridge when it…	NEW: "I thought I was going to die," driver says . Man says pickup truck was folded in…	06352019a19ae31e527f37f7571c6dd7f0c5da37
WASHINGTON (CNN) -- Doctors removed five small polyps from President Bush's colon on…	Five small polyps found during procedure; "none worrisome," spokesman says . President…	24521a2abb2e1f5e34e6824e0f9e56904a2b0e88
(CNN) -- The National Football League has indefinitely suspended Atlanta Falcons…	NEW: NFL chief, Atlanta Falcons owner critical of Michael Vick's conduct . NFL…	7fe70cc8b12fab2d0a258fababf7d9c6b5e1262a
BAGHDAD, Iraq (CNN) -- Dressed in a Superman shirt, 5-year-old Youssif held his sister's…	Parents beam with pride, can't stop from smiling from outpouring of support . Mom: "I…	a1ebb8bb4d370a1fdf28769206d572be60642d70

< Previous | 1 | 2 | 3 | ... | 2,872 | Next >

Hugging Face에서 확인한 CNN/DailyMail

요약 - CNN/DailyMail

```
(CNN) -- Usain Bolt rounded off the world championships Sunday by claiming his
third gold in Moscow as he anchored Jamaica to victory in the men's 4x100m
relay. The fastest man in the world charged clear of United States rival Justin
Gatlin as the Jamaican quartet of Nesta Carter, Kemar Bailey-Cole, Nickel
Ashmeade and Bolt won in 37.36 seconds. The U.S finished second in 37.56 seconds
with Canada taking the bronze after Britain were disqualified for a faulty
handover. The 26-year-old Bolt has n

Summary (length: 180):
Usain Bolt wins third gold of world championship .
Anchors Jamaica to 4x100m relay victory .
Eighth gold at the championships for Bolt .
Jamaica double up in women's 4x100m relay .
```

실습에서 사용할 데이터셋

요약 - CNN/DailyMail

```
import nltk
from nltk.tokenize import sent_tokenize

nltk.download('punkt')

string = "The U.S. are a country. The U.N. is an organization."
sent_tokenize(string)
```

```
[nltk_data] Downloading package punkt to /Users/andy/nltk_data...
[nltk_data]   Package punkt is already up-to-date!

['The U.S. are a country.', 'The U.N. is an organization.']
```

NLTK의 send_tokenize로 종결과 약어의 구두점을 구분

```
def three_sentence_summary(text):
    return "\n".join(sent_tokenize(text)[:3])

summaries["baseline"] = three_sentence_summary(sample_text)
```

처음 세 문장으로 *baseline* 구성

GPT-2

T5

BART

PEGASUS

요약 - 모델 비교

GPT-2	T5	BART	PEGASUS
Decoder	Encoder-Decoder	Encoder-Decoder	Encoder-Decoder
텍스트 생성 (글쓰기, 번역, 요약)	텍스트 기반 작업 (변역, 요약, 질문 답변 등)	텍스트 생성 및 이해 (요약, 질문 답변, 문장 유사도)	요약

요약 - 모델 비교

```
from transformers import pipeline, set_seed

set_seed(42)

pipe = pipeline("text-generation", model="gpt2-xl")
gpt2_query = sample_text + "\nTL;DR:\n"
pipe_out = pipe(gpt2_query, max_length=512, clean_up_tokenization_spaces=True)
summaries["gpt2"] = "\n".join(sent_tokenize(pipe_out[0]["generated_text"][len(gpt2_query) :]))
```

GPT-2

```
pipe = pipeline("summarization", model="t5-large")
pipe_out = pipe(sample_text)
summaries["t5"] = "\n".join(sent_tokenize(pipe_out[0]["summary_text"]))
```

T5

```
pipe = pipeline("summarization", model="facebook/bart-large-cnn")
pipe_out = pipe(sample_text)
summaries["bart"] = "\n".join(sent_tokenize(pipe_out[0]["summary_text"]))
```

BART

```
pipe = pipeline("summarization", model="google/pegasus-cnn_dailymail")
pipe_out = pipe(sample_text)
summaries["pegasus"] = pipe_out[0]["summary_text"].replace(" .<n>", ".\n")
```

PEGASUS

요약 - 모델 비교

GROUND TRUTH
Usain Bolt wins third gold of world championship .
Anchors Jamaica to 4x100m relay victory .
Eighth gold at the championships for Bolt .
Jamaica double up in women's 4x100m relay .

GPT2
Nesta, the fastest man in the world.
Gatlin, the most successful Olympian ever.
Kemar, a Jamaican legend.
Shelly-Ann, the fastest woman ever.
Bolt, the world's greatest athlete.
The team sport of pole vaulting

T5
usain bolt wins his third gold medal of the world championships in the men's 4x100m relay .
the 26-year-old anchored Jamaica to victory in the event in the Russian capital .
he has now collected eight gold medals at the championships, equaling the record .

요약 - 모델 비교

GROUND TRUTH
Usain Bolt wins third gold of world championship .
Anchors Jamaica to 4x100m relay victory .
Eighth gold at the championships for Bolt .
Jamaica double up in women's 4x100m relay .

BART
Usain Bolt wins his third gold of the world championships in Moscow.
Bolt anchors Jamaica to victory in the men's 4x100m relay.
The 26-year-old has now won eight gold medals at world championships.
Jamaica's women also win gold in the relay, beating France in the process.

PEGASUS
Usain Bolt wins third gold of world championships.
Anchors Jamaica to victory in men's 4x100m relay.
Eighth gold at the championships for Bolt.
Jamaica also win women's 4x100m relay .

요약 - 성능평가

- BLEU, ROGUE 등의 평가 지표

BLEU **(Bilingual Evaluation Understudy)**	**ROGUE** **(Recall-Oriented Understudy for Gisting Evaluation)**
정밀도를 근간으로 하는 지표	재현율을 근간으로 하는 지표
생성한 텍스트에 참조 텍스트와 일치하는 비율 측정	참조 텍스트가 생성된 텍스트에 포함된 비율을 측정
기계 번역 시스템의 성능 평가	텍스트 요약 시스템의 성능 평가

요약 - 성능평가 (BLEU)

snt: 문장

snt': 참조 문장

$$p_n = \frac{\sum_{\text{snt}\in C}\sum_{\text{n-gram}\in\text{snt'}}\text{Count}_{\text{clip}}(\text{n-gram})}{\sum_{\text{snt'}\in C}\sum_{\text{n-gram}\in\text{snt}}\text{Count}(\text{n-gram})}$$

$$BR = \min\left(1, e^{1-\frac{\ell_{\text{ref}}}{\ell_{\text{gen}}}}\right)$$

$$\text{BLEU-N} = BR \times \left(\prod_{n=1}^{N} p_n\right)^{1/N}$$

요약 - 성능평가 (ROGUE)

snt: 문장

snt': 참조 문장

$$\text{ROUGE-N} = \frac{\sum_{snt' \in C} \sum_{n\text{-}gram \in snt'} \text{Count}_{\text{match}}(n\text{-}gram)}{\sum_{snt' \in C} \sum_{n\text{-}gram \in snt'} \text{Count}(n\text{-}gram)}$$

$$R_{LCS} = \frac{LCS(X, Y)}{m} \qquad P_{LCS} = \frac{LCS(X, Y)}{n}$$

$$F_{LCS} = \frac{(1 + \beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta P_{LCS}}, \text{ where } \beta = P_{LCS} / R_{LCS}$$

요약 - 성능평가

예측: Korea won the soccer World Cup final.

요약: Korea won the World Cup.

ROGUE-1

$$\frac{\text{서로 겹치는 유니그램 수}}{\text{참조 요약의 유니그램 수}} = \frac{5}{5}$$

ROGUE-2

$$\frac{\text{서로 겹치는 바이그램 수}}{\text{참조 요약의 바이그램 수}} = \frac{3}{4}$$

요약 - CNN/DailyMail 평가

- ROGUE 지표로 평가

	rouge1	rouge2	rougeL	rougeLsum
baseline	0.303571	0.090909	0.214286	0.232143
gpt2	0.187500	0.000000	0.125000	0.187500
t5	0.486486	0.222222	0.378378	0.486486
bart	0.582278	0.207792	0.455696	0.506329
pegasus	0.866667	0.655172	0.800000	0.833333

요약 - 요약 모델 훈련

- SAMSum 데이터셋 - 대화와 짧은 요약으로 구성됨

```
dataset_samsum = load_dataset("samsum")
split_lengths = [len(dataset_samsum[split]) for split in dataset_samsum]

print(f"분할 크기: {split_lengths}")
print(f"특성: {dataset_samsum['train'].column_names}")
print("\n대화: ")
print(dataset_samsum["test"][0]["dialogue"])
print("\n요약: ")
print(dataset_samsum["test"][0]["summary"])
```

```
대화:
Hannah: Hey, do you have Betty's number?
Amanda: Lemme check
Hannah: <file_gif>
Amanda: Sorry, can't find it.
Amanda: Ask Larry
Amanda: He called her last time we were at the park together
Hannah: I don't know him well
Hannah: <file_gif>
Amanda: Don't be shy, he's very nice
Hannah: If you say so..
Hannah: I'd rather you texted him
Amanda: Just text him 🙂
Hannah: Urgh.. Alright
Hannah: Bye
Amanda: Bye bye

요약:
Hannah needs Betty's number but Amanda doesn't have it. She needs to contact Larry.
```

요약 - 요약 모델 훈련

- PEGASUS로 요약 파이프라인 실행

```
pipe_out = pipe(dataset_samsum["test"][0]["dialogue"])
print("요약: ")
print(pipe_out[0]["summary_text"].replace(" .<n>", ".\n"))
```

```
요약:
Amanda: Ask Larry Amanda: He called her last time we were at the park together.
Hannah: I'd rather you texted him.
Amanda: Just text him .
```

- ROGUE 평가 결과

	rouge1	rouge2	rougeL	rougeLsum
pegasus	0.296168	0.087803	0.229604	0.229514

요약 - 요약 모델 훈련

- PEGASUS로 파인튜닝
- PEGASUS와 같은 인코더-디코더 구조에서는 **티처 포싱** 적용
  - 디코더는 한 토큰이 이동된 정답을 입력으로 받음
  - 디코더는 이전 스텝의 정답 레이블만 보고 현재와 미래의 레이블은 못 봄
  - 현재와 미래의 모든 입력을 마스킹하는 Masked Self-Attention

step	decoder_input	label
1	[PAD]	Transformers
2	[PAD, Transformers]	are
3	[PAD, Transformers, are]	awesome
4	[PAD, Transformers, are, awesome]	for
5	[PAD, Transformers, are, awesome, for]	text
6	[PAD, Transformers, are, awesome, for, text]	summarization

요약 - 요약 모델 훈련

- 훈련

```
trainer.train()
score = evaluate_summaries_pegasus(
    dataset_samsum["test"], rouge_metric, trainer.model, tokenizer,
    batch_size=2, column_text="dialogue", column_summary="summary")

rouge_dict = dict((rn, score[rn].mid.fmeasure) for rn in rouge_names)
pd.DataFrame(rouge_dict, index=[f"pegasus"])
```

- 훈련 결과

	rouge1	rouge2	rougeL	rougeLsum
pegasus	0.427614	0.200571	0.340648	0.340738

요약 - 대화 요약 생성

- 임의의 대화 생성

```
custom_dialogue = """\
Thom: Hi guys, have you heard of transformers?
Lewis: Yes, I used them recently!
Leandro: Indeed, there is a great library by Hugging Face.
Thom: I know, I helped build it ;)
Lewis: Cool, maybe we should write a book about it. What do you think?
Leandro: Great idea, how hard can it be?!
Thom: I am in!
Lewis: Awesome, let's do it together!
"""
print(pipe(custom_dialogue, **gen_kwargs)[0]["summary_text"])
```

- 훈련 결과

```
Thom, Lewis and Leandro are going to write a book about transformers. Thom
helped build a library by Hugging Face. They are going to do it together.
```

Q&A