트랜스포머를 활용한 자연어 처리

6장 요약

6.1 CNN/DailyMail 데이터셋

CODE

```
sample = dataset["train"][1]
print(f"""기사 (500개 문자 발췌, 총 길이: {len(sample['article'])}):""")
print(sample["article"][:500])
print(f"\n요약 (길이: {len(sample['highlights'])}):")
print(sample["highlights"])
```

6.1 CNN/DailyMail 데이터셋

CODE

```
for sample in dataset["train"]:
    if sample["article"].startswith("(CNN)  -- Usain Bolt rounded"):
        print(sample["article"])
        break
```

6.1 CNN/DailyMail 데이터셋

OUTPUT

```
기사 (500개 문자 발췌, 총 길이: 4051):
Editor's note: In our Behind the Scenes series, CNN correspondents share their experiences in covering news and analyze the stories behind the events. Here, Soledad O'Brien takes users inside a jail where many of the inmates are mentally ill. An inmate housed on the "forgotten floor," where many mentally ill inmates are housed in Miami before trial. MIAMI, Florida (CNN) -- The ninth floor of the Miami-Dade pretrial detention facility is dubbed the "forgotten floor." Here, inmates with the most s

요약 (길이: 281):
Mentally ill inmates in Miami are housed on the "forgotten floor" Judge Steven Leifman says most are there as a result of "avoidable felonies" While CNN tours facility, patient shouts: "I am the son of the president" Leifman says the system is unjust and he's fighting for change .
```

6.1 CNN/DailyMail 데이터셋

SUMMARY

마이애미의 정신 질환 수감자들은 "잊혀진 층"에 수용되어 있습니다.
스티븐 레이프만 판사는 대부분 "피할 수 있는 중범죄"의 결과로 그곳에 있다고 말합니다.
CNN이 시설을 둘러 보는 동안 환자는 외칩니다: "나는 대통령의 아들입니다“
레이프만은 시스템이 부당하며 변화를 위해 싸우고 있다고 말합니다.

6.1 CNN/DailyMail 데이터셋

SUMMARY

CNN 특파원 Soledad O'Brien이 Steven Leifmen 판사와 교도소를 방문하여 취재한 내용입니다.
Miami Dade 재판 전 **구금 시설의 9층은 "잊혀진 층"** 이라고 불립니다.
여기에는 심각한 **정신 질환을 앓고 있는 범죄자들이 법정에 출두하기 전까지 수감**됩니다.
수감자들은 **자해하지 않도록** 민소매 옷을 입고 있으며, 신발도 신고 있지 않습니다.
교도소의 수감자 중 약 3분의 1은 정신질환자이며, 좁고 시끄러운 감방에 갇혀 있습니다.
과거에는 정신질환자는 **혐의가 없어도** 사회에 부적합한 존재로 간주되어 **감옥에 갇혔습니다.**
그들이 곧 감옥에서 병원으로 옮겨졌으나, 많은 정신병원이 문을 닫았고 **노숙자 신세**가 되었습니다.
1955년 주립 정신 병원에는 50만 명 이상의 환자가 있었으나 현재는 4~5만 명 뿐입니다.
판사는 이를 **바꾸기 위해 노력 중**이며, 처벌이 아닌 치료를 위해 수감자를 정신 건강 시설로 보낼 예정입니다.
환자는 치료받고, 정부는 재소자를 반복해서 재수감하지 않음으로 비용을 절감하는 **윈윈 솔루션**입니다.

6.2 텍스트 요약 파이프라인

CODE

```
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)
```

OUTPUT

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

6.2 텍스트 요약 파이프라인

CODE

```
from nltk.tokenize.punkt import PunktSentenceTokenizer, PunktParameters

string = "Fig. 2 shows a U.S.A. map."
print(sent_tokenize(string))

punkt_param = PunktParameters()
abbreviation = ['u.s.a', 'fig']
punkt_param.abbrev_types = set(abbreviation)
tokenizer = PunktSentenceTokenizer(punkt_param)
print(tokenizer.tokenize(string))
```

OUTPUT

```
['Fig.', '2 shows a U.S.A. map.’]
['Fig. 2 shows a U.S.A. map.']
```

6.2 텍스트 요약 파이프라인

T5

- Text-to-Text Transfer Transformer (T5)
- Encoder – Decoder 구조
- NLU와 NLG 작업을 text-to-text(seq-to-seq) 작업으로 변환해 통합
- SuperGLUE 작업과 masked language modeling으로 사전 훈련

6.2 텍스트 요약 파이프라인

BART

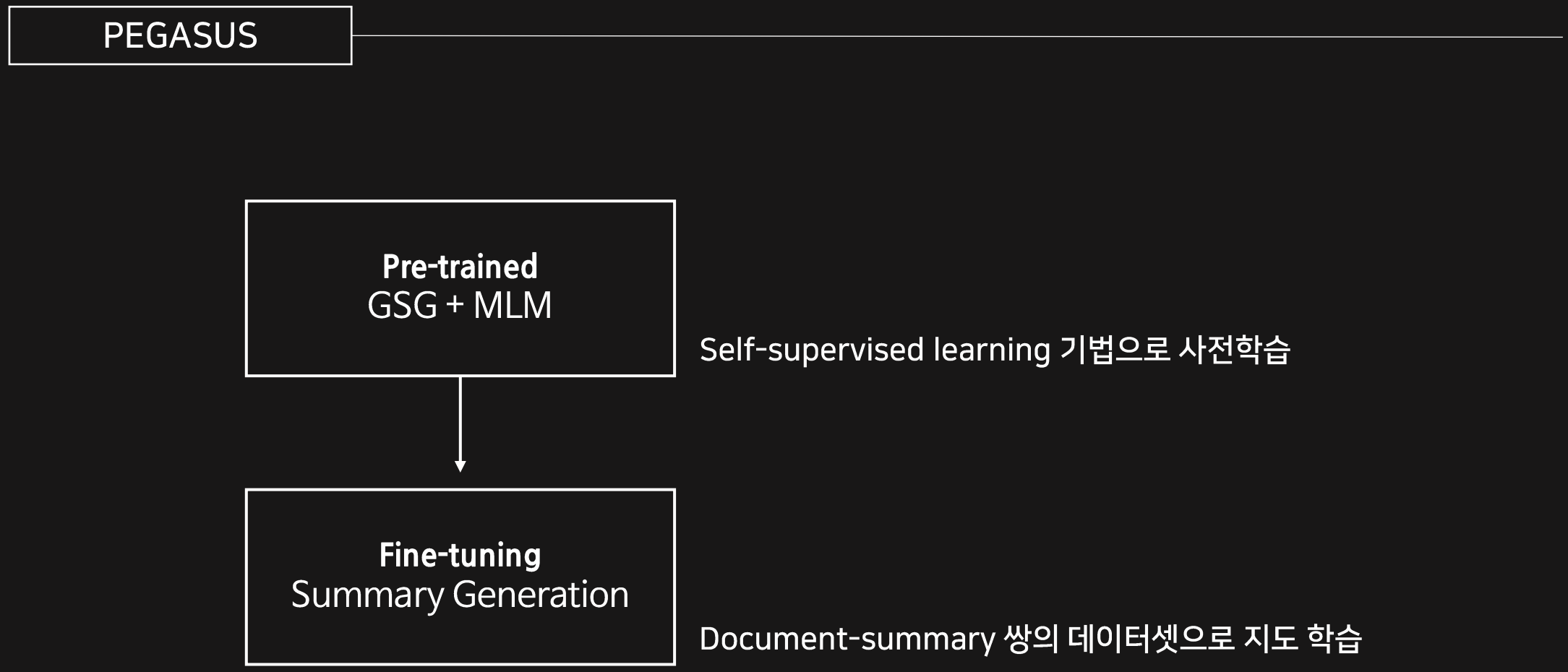
- Bidirectional + Auto-Regressive Transformer
- Encoder – Decoder 구조
- BERT와 GPT의 사전 훈련 과정을 결합
  - 입력 시퀀스는 마스킹, 문장 섞기, 토큰 삭제, 문서 순환 등의 변환을 거침
  - 디코더는 원본 시퀀스를 재구성
- NLU와 NLG 작업에 모두 사용

6.2 텍스트 요약 파이프라인

PEGASUS

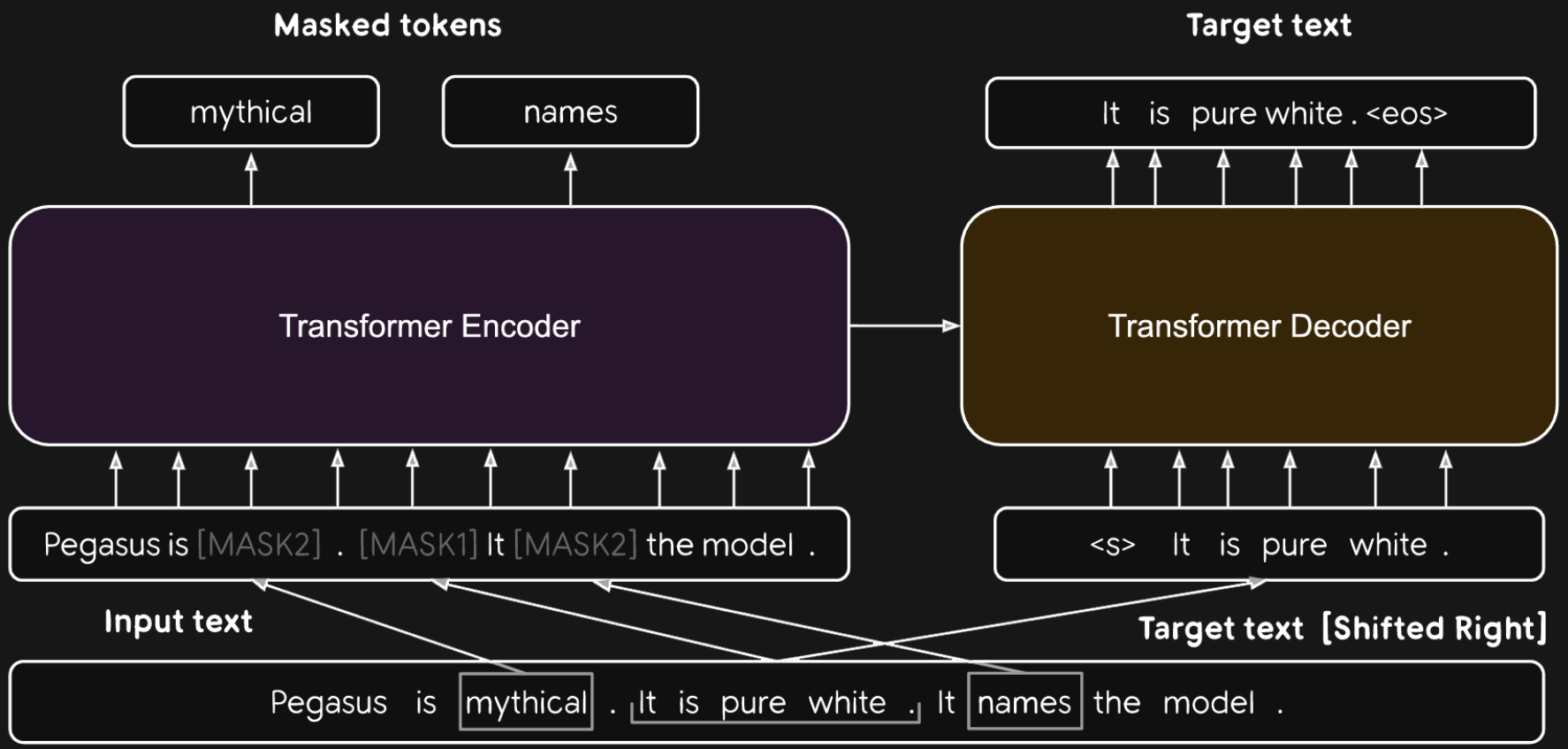
- Pre-training with Extracted Gap-sentences for Abstractive Summarization
- Encoder – Decoder 구조
- 여러 문장으로 구성된 텍스트에서 마스킹된 문장을 예측하는 것을 목표로 사전 훈련
- 일반적인 언어 모델링보다 요약에 특화
- 주변 문단의 내용을 대부분 담은 문장을 자동으로 식별

6.2 텍스트 요약 파이프라인



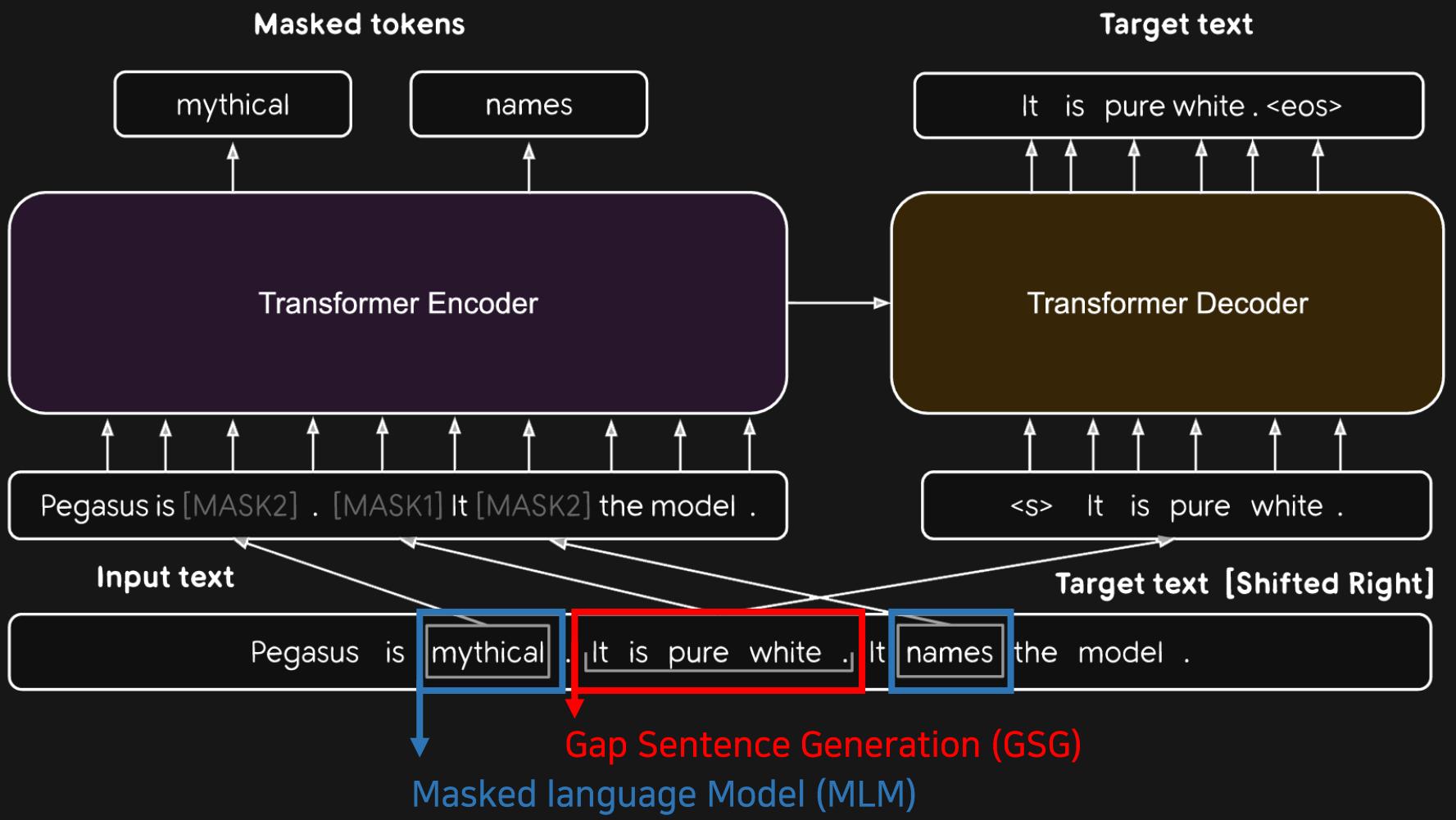
6.2 텍스트 요약 파이프라인

PEGASUS



6.2 텍스트 요약 파이프라인

PEGASUS



6.2 텍스트 요약 파이프라인

PEGASUS

Gap Sentence Generation (GSG)

- Random, Lead, Principal 3개의 전략 사용
- Principal은 Ind(independently)와 Seq(sequentially)의 두 가지 기법 사용
- ROUGE1-F1 점수를 계산할 때, n-gram을 한 번만 카운트하는 Uniq와 기존과 같은 방법인 Orig 두 가지 기법 사용
- Gap Sentence Ratio(GSR)은 30%를 사용

Maksed Language Model (MLM)

- GSG에 선택되지 않은 문장에는 MLM 기법을 적용
- 15% 토큰을 선택하여 80%는 [MASK2] 토큰으로 대체, 10%는 랜덤 토큰으로 대체, 10%는 그대로 두었음
- 다운스트림 태스크에서는 성능에 영향을 거의 미치지 않아서 $\text{PEGASUS}_{\text{LARGE}}$ 모델에는 사용하지 않음

6.3 요약 결과 비교하기

CODE

GROUD TRUTH

Mentally ill inmates in Miami are housed on the "forgotten floor" Judge Steven Leifman says most are there as a result of "avoidable felonies" While CNN tours facility, patient shouts: "I am the son of the president" Leifman says the system is unjust and he's fighting for change .

BASELINE

Editor's note: In our Behind the Scenes series, CNN correspondents share their experiences in covering news and analyze the stories behind the events. Here, Soledad O'Brien takes users inside a jail where many of the inmates are mentally ill. An inmate housed on the "forgotten floor," where many mentally ill inmates are housed in Miami before trial. MIAMI, Florida (CNN) -- The ninth floor of the Miami-Dade pretrial detention facility is dubbed the "forgotten floor."

GPT2

Here's a more "interesting" story CNN.com was unable to access. This story in the Miami Herald is even more disturbing. A video report at 6:00 of Leifman's office shows a video of a mentally ill inmate breaking through the bars of a jail cell and making a break for it. The Miami Herald headline above tells you everything

6.3 요약 결과 비교하기

CODE

T5

mentally ill inmates are housed on the ninth floor of a florida jail . most face drug charges or charges of assaulting an officer . judge says arrests often result from confrontations with police . one-third of all people in Miami-dade county jails are mental ill .

BART

Mentally ill inmates are housed on the "forgotten floor" of Miami-Dade jail. Most often, they face drug charges or charges of assaulting an officer. Judge Steven Leifman says the arrests often result from confrontations with police. He says about one-third of all people in the county jails are mentally ill.

PEGASUS

Mentally ill inmates in Miami are housed on the "forgotten floor"<n>The ninth floor is where they're held until they're ready to appear in court. Most often, they face drug charges or charges of assaulting an officer. They end up on the ninth floor severely mentally disturbed .

6.4 생성된 텍스트 품질 평가하기

참조 텍스트 : the cat is on the mat

생성 텍스트 : the the the the the the

$$p_{vanilla} = \frac{6}{6}$$

$$p_{mod} = \frac{2}{6}$$

6.4 생성된 텍스트 품질 평가하기

$$p_n = \frac{\sum_{n-gram \in snt'} Count_{clip}(n-gram)}{\sum_{n-gram \in snt} Count_{(}n-gram)}$$

n-그램으로 확장 (snt는 생성된 텍스트, snt'는 참조 텍스트)

6.4 생성된 텍스트 품질 평가하기

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

말뭉치 C에 있는 모든 샘플에 대한 식

6.4 생성된 텍스트 품질 평가하기

$$BR = min(1, e^{1-\ell_{ref}/\ell_{gen}})$$

짧은 시퀀스일수록 유리한 문제가 발생 Brevity Penalty 도입

6.4 생성된 텍스트 품질 평가하기

참조 텍스트 : the cat is on the mat

생성 텍스트 : cat

Precision이 1이 되어버리는 문제가 발생

6.4 생성된 텍스트 품질 평가하기

Good Morning

1. 안녕하세요!
2. 좋은 아침이에요!
3. 아침이 밝았어요!

>>> 안녕하세요! 좋은 아침이 밝았어요!

6.4 생성된 텍스트 품질 평가하기

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

6.4 생성된 텍스트 품질 평가하기

BLEU SCORE의 한계

1. 의미를 고려하지 않음
2. 문장 구조를 고려하지 않음
3. 형태학적으로 풍부한 언어를 잘 처리하지 못함
4. 사람의 판단과는 잘 맞지 않음

https://towardsdatascience.com/evaluating-text-output-in-nlp-bleu-at-your-own-risk-e8609665a213

6.4 생성된 텍스트 품질 평가하기

1. 의미를 고려하지 않음

참조 번역: I ate the apple.

생성된 번역

1. I consumed the apple.
2. I ate an apple.
3. I ate the potato.

https://towardsdatascience.com/evaluating-text-output-in-nlp-bleu-at-your-own-risk-e8609665a213

2. 문장 구조를 고려하지 않음

참조 번역:

Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.

Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.

Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.

Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

생성된 번역:

1. Appeared calm when he was taken to the American plane, which will to Miami, Florida.
2. which will he was, when taken appeared calm to the American plane to Miami, Florida.

3. 형태학적으로 풍부한 언어를 잘 처리하지 못함

영어
Her village is large.

페루의 시피보어
Jawen jemara ani iki.
(실제로 가본 적이 있음)

Jawen jemaronki ani iki.
(다른 사람에게 들었음)

4. 사람의 판단과는 잘 맞지 않음

BLEU 점수와 human evaluation과의 상관관계가 크지 않음

단순히 F1 점수와 상관관계를 계산했을 때보다 떨어질 때도 있음

https://towardsdatascience.com/evaluating-text-output-in-nlp-bleu-at-your-own-risk-e8609665a213

6.4 생성된 텍스트 품질 평가하기

CODE

```
from datasets import load_metric

bleu_metric = load_metric("sacrebleu")
```

CODE

```
from evaluate import load

bleu_metric_new = load("sacrebleu")
```

6.4 생성된 텍스트 품질 평가하기

CODE

```
bleu_metric.add(
    prediction="the the the the the the", reference=["the cat is on the mat"])
results = bleu_metric.compute(smooth_method="floor", smooth_value=0)
results["precisions"] = [np.round(p, 2) for p in results["precisions"]]

df = pd.DataFrame.from_dict(results, orient="index", columns=["Value"])
df["Description"] = df.index.map(descriptions)
```

6.4 생성된 텍스트 품질 평가하기

OUTPUT

	Value	Description
score	0.0	최종 BLEU 점수
counts	[2, 0, 0, 0]	매칭된 n-그램의 개수
totals	[6, 5, 4, 3]	가능한 n-그램의 총 개수
precisions	[33.33, 0.0, 0.0, 0.0]	각 n-그램에 대한 정밀도
bp	1.0	브레비티 페널티 값
sys_len	6	생성된 텍스트의 길이
ref_len	6	참조 텍스트의 길이

6.4 생성된 텍스트 품질 평가하기

$$p_n = \frac{\text{n-grams count in reference text} + \text{smooth value}(=0.1)}{\text{n-grams in generated text}}$$

`smooth_value`를 사용하면 분자에 상수 값을 추가

`smooth_method="floor"`일 경우 `smooth_value`는 0.1

6.4 생성된 텍스트 품질 평가하기

CODE

```
results = bleu_metric.compute(smooth_method="floor")
```

OUTPUT

	Value	Description
score	4.854918	최종 BLEU 점수
counts	[2, 0, 0, 0]	매칭된 n-그램의 개수
totals	[6, 5, 4, 3]	가능한 n-그램의 총 개수
precisions	[33.33, 2.0, 2.5, 3.33]	각 n-그램에 대한 정밀도
bp	1.0	브레비티 페널티 값
sys_len	6	생성된 텍스트의 길이
ref_len	6	참조 텍스트의 길이

6.4 생성된 텍스트 품질 평가하기

$$\frac{0.1}{5} = 0.02$$

겹치는 bigram이 한 개도 없지만,

smooth value에 의해 precision이 0.02가 됨

6.4 생성된 텍스트 품질 평가하기

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

6.4 생성된 텍스트 품질 평가하기

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

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

https://www.youtube.com/watch?v=TMshhnrEXlg

6.4 생성된 텍스트 품질 평가하기

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

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

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

6.5 CNN/DailyMail 데이터셋에서 PEGASUS 평가하기

CODE

```
summaries = model.generate(input_ids=inputs["input_ids"].to(device),
                           attention_mask=inputs["attention_mask"].to(device),
                           length_penalty=0.8, num_beams=8, max_length=128)
```

6.5 CNN/DailyMail 데이터셋에서 PEGASUS 평가하기

CODE

```
model_ckpt = "google/pegasus-cnn_dailymail"
model = AutoModelForSeq2SeqLM.from_pretrained(model_ckpt).to(device)
model.generate
```

OUTPUT

```
<bound method GenerationMixin.generate of PegasusForConditionalGeneration(
        (model): PegasusModel(
```

6.5 CNN/DailyMail 데이터셋에서 PEGASUS 평가하기

length_penalty (float, optional, defaults to 1.0)

Exponential penalty to the length that is used with beam-based generation. It is applied as an exponent to the sequence length, which in turn is used to divide the score of the sequence. Since the score is the log likelihood of the sequence (i.e. negative), length_penalty > 0.0 promotes longer sequences, while length_penalty < 0.0 encourages shorter sequences.

https://huggingface.co/docs/transformers/v4.33.0/en/main_classes/text_generation#transformers.GenerationMixin

6.6 요약 모델 훈련하기

CODE

```
def convert_examples_to_features(example_batch):
    input_encodings = tokenizer(example_batch["dialogue"], max_length=1024,
                                truncation=True)

    with tokenizer.as_target_tokenizer():
        target_encodings = tokenizer(example_batch["summary"], max_length=128,
                                     truncation=True)

    return {"input_ids": input_encodings["input_ids"],
            "attention_mask": input_encodings["attention_mask"],
            "labels": target_encodings["input_ids"]}
```