Dialogue Summarization | 일상 대화 요약

학교 생활, 직장, 치료, 쇼핑, 여가, 여행 등 광범위한 일상 생활 중 하는 대화들에 대해 요약합니다.

#비공개대회 #AI부트캠프14기 #NLPAdvanced

30 | 2025.09.26 10:00 ~ 2025.10.15 19:00 • 진행중

데이터 개요

데이터 다운로드 링크

https://aistages-api-public-prod.s3.amazonaws.com/app/Competitions/000365/data/data.tar.gz

학습 데이터 개요

> *"데이터 건수"*

모든 데이터는 .csv 형식으로 제공되고 있으며, 각각의 데이터 건수는 다음과 같습니다.

- train : 12457
- dev : 499
- test : 250
- hidden-test : 249

> *"데이터 구성"*

데이터는 아래와 같은 형태이며, 최소2턴, 최대 60턴으로 대화가 구성되어 있습니다. 대화(*dialogue)를 보고 이에 대한 요약(*summary) 를 예측하는 것이 최종 목표입니다.

	fname	dialogue	summary
0	train_0	#Person1#: 안녕하세요, 스미스씨. 저는 호킨스 의사입니다. 오늘 왜 오셨나...	스미스씨가 건강검진을 받고 있고, 호킨스 의사는 매년 건강검진을 받는 것을 권장합니...
1	train_1	#Person1#: 안녕하세요, 파커 부인, 어떻게 지내셨나요?\n#Person2#...	파커 부인이 리키를 데리고 백신 접종을 하러 갔다. 피터스 박사는 기록을 확인한 후...
2	train_2	#Person1#: 실례합니다, 엘의 한 묶음을 보셨나요?\n#Person2#: 어떤...	#Person1#은 엘의 한 묶음을 찾고 있고, 그것을 찾기 위해 #Person2#...
3	train_3	#Person1#: 왜 너는 여자친구가 있다는 걸 말해주지 않았어?\n#Person...	#Person1#은 #Person2#가 여자친구가 있고 그녀와 결혼할 것이라는 사실...
4	train_4	#Person1#: 안녕, 숙녀분들! 오늘 밤 당신들은 정말 멋져 보여. 이 춤을 ...	말릭이 니키에게 춤을 요청한다. 말릭이 발을 밟는 것을 신경 쓰지 않는다면 니키는 ...
...	...	...	...
12452	train_12455	#Person1#: 실례합니다. 맨체스터 출신의 그린 씨이신가요?\n#Person2...	탄 링은 흰머리와 수염으로 쉽게 인식되는 그린 씨를 만나 호텔로 데려갈 예정입니다...
12453	train_12456	#Person1#: 이링 씨가 우리가 컨퍼런스 센터에 오후 4시에 도착해야 한다고 ...	#Person1#과 #Person2#는 이링 씨가 늦지 않도록 요청했기 때문에 컨퍼...
12454	train_12457	#Person1#: 오늘 어떻게 도와드릴까요?\n#Person2#: 차를 빌리고 싶...	#Person2#는 #Person1#의 도움으로 5일 동안 소형 차를 빌립니다.
12455	train_12458	#Person1#: 오늘 좀 행복해 보이지 않아. 무슨 일 있어?\n#Person2...	#Person2#의 엄마가 일자리를 잃었다. #Person2#는 엄마가 우울해하지 ...
12456	train_12459	#Person1#: 엄마, 다음 토요일에 이 삼촌네 가족을 방문하기 위해 비행기를 ...	#Person1#은 다음 토요일에 이 삼촌네를 방문할 때 가방을 어떻게 싸야 할지 ...

- fname : 대화 고유번호 입니다. 중복되는 번호가 없습니다.
- dialogue : 최소 2명에서 최대 7명이 등장하여 나누는 대화 내용입니다. 각각의 발화자를 구분하기 위해 #Person"N"#: 을 사용하며, 발화자의 대화가 끝나면 \n 으로 구분합니다. 이 구분자를 기준으로 하여
- summary : 해당 대화를 바탕으로 작성된 요약문입니다.

> *"데이터 다운로드"*

wget 명령어를 통해 데이터셋을 본인의 작업 환경에 다운로드 합니다.

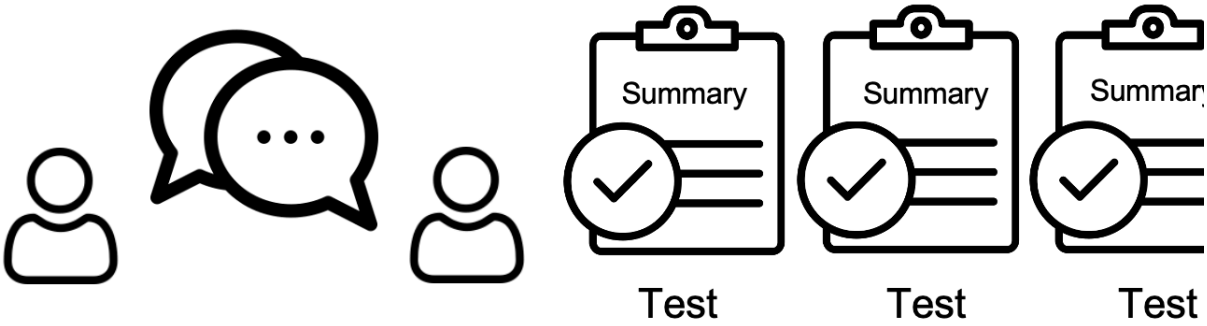
```
- 데이터 : wget [URL 첨부]
ex) wget https://aistages-api-public-prod.s3.amazonaws.com ~ /data.tar.gz
```

평가 데이터 개요

평가 데이터의 대화를 바탕으로 요약문을 생성하고 이를 평가합니다.

다만, 모델이 대화의 다양한 부분을 초점 삼아 요약문을 생성할 수 있는데, 학습할 때 처럼 1개의 요약문으로만 평가하게 된다면 실제로 사람이 봤을 때는 옳게 생성한 요약문인데, score 는 낮을 수 있습니다.

그렇기에 평가시에는 하나의 대화에서 다양한 관점으로 작성된 3개의 요약문으로 평가합니다.



대화의 핵심 주제를 어떤 것으로 보느냐에 따라 다른 요약문이 나올 수 있는 부분을 반영하여 채점하게 되겠죠? 모델의 다양성을 고려하여 좀 더 일반적으로 측정하기 위해 구성되었습니다.

이 세 개의 요약문과 모델이 예측한 요약문을 비교하여 ROUGE metric으로부터 산출된 점수를 바탕으로 score 가 계산됩니다. 자세한 내용은 평가 방법을 참고해주세요!

평가 데이터는 총 499개이나, 평가 데이터는 여러분께 공개되어 점수가 계산되는 공개 평가 데이터와, 여러분들께 공개되지 않으나 점수에는 계산되는 비공개 평가 데이터로 나뉩니다. 이 두 평가 데이터는 50:50 .

데이터 노이즈

학습, 검증 및 평가 데이터는 다양한 형태의 노이즈를 포함할 수 있습니다.

예를 들어, 맞춤법 오류, 문장 부호의 누락 또는 과다 사용, 발화 구분 기호의 불일치, 화자 표기의 불명확함 등이 존재할 수 있습니다. 아래는 대표적인 예시입니다.

- 예시 1. newline character `"\n"` 가 `"\\n"` 형태로 표현된 경우

```
#Person1#: 저, 불만이 있어요. 열 분 동안 테이블에서 기다렸는데, 웨이터가 드디어 와서 주문을 받았어요. 그런데 나온 음식이 제가 주문한 게 아니더라고요.Wn#Person2#: 정말 죄송합니다. 오늘 밤은
```

- 예시 2. newline character 대신 HTML tag `<br>`이 포함된 경우

```
#Person1#: 요즘 잘 지내고 있어요?<br>#Person2#: 제 코치가 제 혈압을 체크해 달라고 부탁했어요.<br>#Person1#: 전에 고혈압 있다고 들은 적 있나요?<br>#Person2#: 고혈압 증상은 없어요.<br>#Pers
```

이러한 노이즈는 학습과 평가에 영향을 줄 수 있으므로, 모델 학습/평가 시 이를 적절히 처리하거나 노이즈에 강건한 구조를 설계하는 것이 중요합니다.