NLP - Dialogue Summarization

■ 데이터

🔶 Train

- 원본 DialogSum 데이터셋을 **한국어로 번역**한 데이터셋.

> DialogSum 벤치마크에서 상위권을 달성하 모델 및 방법론을 조사하면 좋을 듯.

- 크기: 12457

	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, 열쇠 한 묶음 보셨나요?\n#Person2#: 어떤...	#Person1#은 열쇠 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#: 이윙 씨가 우리가 컨퍼런스 센터에 오후 4시에 도착해야 한다고 ...	#Person1#과 #Person2#는 이윙 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 ...

- dialogue 규칙
  1. 발화자 규칙: `#Person[Number]#`
  2. `:` (콜론 + 띄어쓰기) 이후에 발화문 등장.
  3. `\n` (개행문자)로 발화자 간 구분

```
#Person1#: 안녕하세요, 스미스씨. 저는 호킨스 의사입니다. 오늘 왜 오셨나요?
#Person2#: 건강검진을 받는 것이 좋을 것 같아서요.
#Person1#: 그렇군요, 당신은 5년 동안 건강검진을 받지 않았습니다. 매년 받아야 합니다.
#Person2#: 알고 있습니다. 하지만 아무 문제가 없다면 왜 의사를 만나러 가야 하나요?
#Person1#: 심각한 질병을 피하는 가장 좋은 방법은 이를 조기에 발견하는 것입니다. 그러니 당신의 건강을 위해 최소한 매년 한 번은 오세요.
#Person2#: 알겠습니다.
#Person1#: 여기 보세요. 당신의 눈과 귀는 괜찮아 보입니다. 깊게 숨을 들이쉬세요. 스미스씨, 담배 피우시나요?
#Person2#: 네.
#Person1#: 당신도 알다시피, 담배는 폐암과 심장병의 주요 원인입니다. 정말로 끊으셔야 합니다.
#Person2#: 수백 번 시도했지만, 습관을 버리는 것이 어렵습니다.
#Person1#: 우리는 도움이 될 수 있는 수업과 약물들을 제공하고 있습니다. 나가기 전에 더 많은 정보를 드리겠습니다.
#Person2#: 알겠습니다, 감사합니다, 의사선생님.
```

[example]

- summary 규칙
  - 한 문장 또는 여러 문장으로 구성.
  - 발화자 또는 주요 개채명을 포함

🔶 Test

- 총 499개
  - 250 개: 공개용 평가 데이터셋 → Public 점수 계산에 사용.
  - 249 개: 비공개용 평가 데이터셋 → Final 점수 계산에 사용.

    - 499개를 대화 주제 비율에 맞춰 분할. → 249개와 250개 간 대화 주제 분포는 유사할 것.
- 데이터셋 구성
    - summary가 3개 존재: 주제, 내용, 개체명 등 어떤 관점에서 요약하느냐에 따라 요약문이 다양할 수 있어 이를 반여하여 3개 요약문으로 평가하여 종합.
- **정답 요약문 작성의 주요 기준**
    1. 대화의 가장 중요한 정보를 전달
    2. 대화 길이의 20% 이내로 간략하게 요약
    3. 대화 내에 중요한 명명된 개체를 보존 (사람 이름, 기업명 등)
    4. 관찰자의 관점에서 작성 (화자의 의도를 이해하고 작성)

        > CoT, Reasoning 기법을 통해 dialogue의 의도를 먼저 파악하는 단계를 거치는 것이 효과적일까?

    5. 은어나 약어 없이 공식적으로 사용되는 언어로 작성

■ Evaluation

🔶ROUGE (루지)

$$\text{Score} = \frac{\sum_i^N \text{ROUGE-1-F1}(\text{pred}, \text{gold}_i)}{N} + \frac{\sum_i^N \text{ROUGE-2-F1}(\text{pred}, \text{gold}_i)}{N} + \frac{\sum_i^N \text{ROUGE-L-F1}(\text{pred}, \text{gold}_i)}{N}$$

🔶토큰화

- ROUGE를 평가할 때 토큰화한 후에 평가

[Original text]
호킨스 의사는 매년 건강검진을 받는 것을 권장합니다.

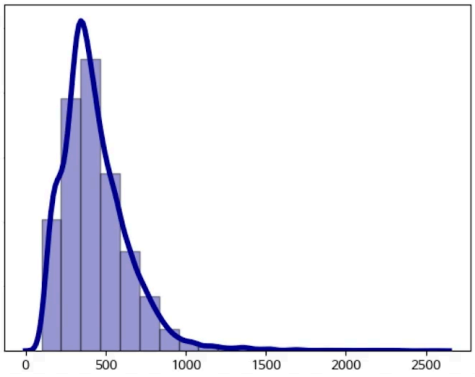
[Tokenized text]
호킨스 의사 는 매년 건강 검진 을 받 는 것 을 권장 합니다 .

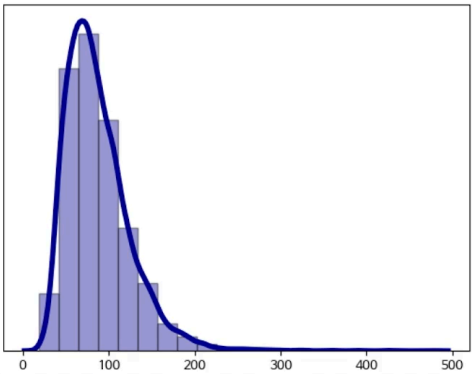
■ EDA

🔻 Baseline EDA

🔶 텍스트 길이

- dialogue, summary 모두 right-skewed(꼬리가 오른쪽으로 치우친) 분포를 가짐.





  - 좌: train 데이터셋의 dialogue 길이 분포
  - 우: train 데이터셋의 summary 길이 분포
    - summary의 분포가 전반적으로 dialogue의 20% 수준

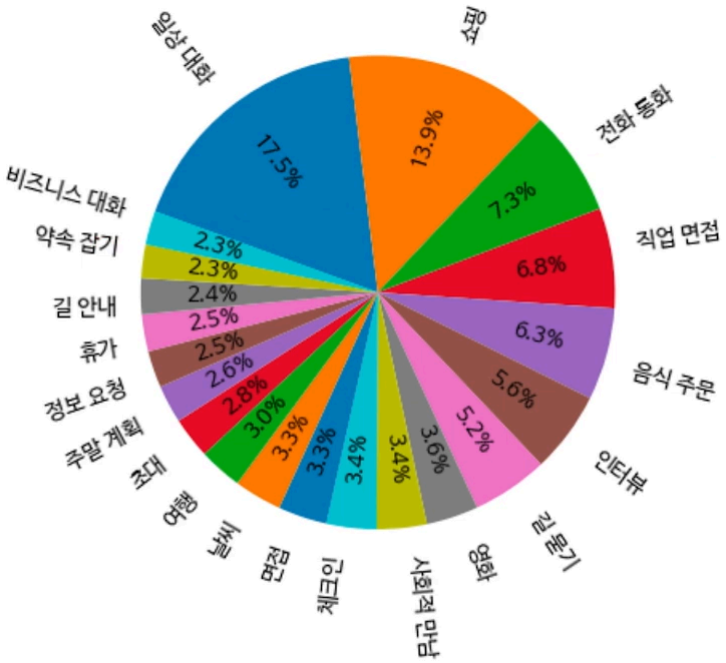
🔶 dialogue 카테고리별 특징

- 채팅 : 약어(ETA, BWT, ...)나 이모티콘(:), ^^, ...)을 최대한 줄인 데이터만 포함.

> 대회를 위해 약어, 이모티콘을 줄였다곤 하지만 여전히 존재할 수 있음. 이를 위한 처리 방법에 무엇이 있을까?

- 일상뿐만 아니라 비즈니스, 면접 대화도 있음으로, 카테고리별로 모델을 구분하거나 이에 맞는 처리를 하는 것이 필요.
- 상위 30개 주제 시각화주제 카테고리가 30개보다 많을 수 있음.

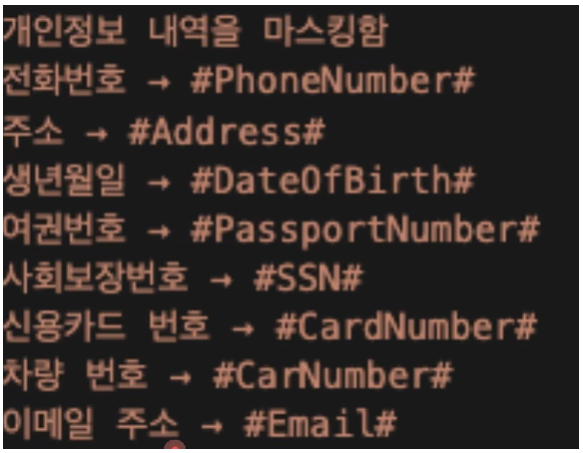
> 대회를 위해 약어, 이모티콘을 줄였다곤 하지만 여전히 존재할 수 있음. 주제를 분류하는 방법들에는 어떤 게 있을가? 주제가 잘못 분류되었을 가능성도 존재.



🔶 개인정보 마스킹

- 총 8개의 개인정보를 마스킹

> special token에 추가해야 함.



🔻DialogSum 논문

🔶데이터셋 배경

- 텍스트 요약 task에 대한 연구가 활발히 진행되고 있지만, 대화 요약은 관심이 적었다. 그 이유는 데이터셋이 충분치 않기 때문.

🔶데이터셋 특징

- Dailydialog, DREAM, MuTual 데이터셋을 합쳐 DialogSum을 구성함.
  - 온라인 대화의 경우 비격식, 약어 토큰(e.g. BTW), 이모티콘 등을 포함
  - 일상 대화는 더 격식 있는 스타일
- summary와 topic은 모두 labeler가 수동으로 제작한 것.
- AIM 보다 더 긴 길이의 대화로 이루어져 있으므로

  ⇒ 대화가 더 명확한 의사소통 패턴과 의도를 가짐.

  ⇒ 더 많은 사건과 그들 사이의 담화 관계가 존재.
- 사람을 표현할 때 주관적, 객관적, 소유격 대명사가 없고 #Person1# 같이 태그로 지칭
- 대부분 구어체

🔶데이터 전처리

- Cleansing
  - 영어만.

    - 오타 및 문법 오류 수정.
    - 텍스트 유사성 기반, 중복된 데이터 제거.
- bi-turn dialogue flow : 동일한 화자의 연속적인 발화를 하나의 발화로 묶어서 항상 A 화자가 말하면 다음은 B 화자가 말하는 식으로 구성됨.

🔶Labeler

- annotator들에게 아래 기준으로 정답을 작성하도록 지시
    - 기본 규칙
        1. 대화의 가장 중요한 정보를 전달
        2. 대화 길이의 20% 이내로 간략하게 요약
        3. 대화 내에 중요한 명명된 개체를 보존 (사람 이름, 기업명 등)
        4. 관찰자의 관점에서 작성 (화자의 의도를 이해하고 작성)
        5. 은어나 약어 없이 공식적으로 사용되는 언어로 작성
    - 시제 : dialogue를 현재 시점으로 간주. dialogue 내의 사건에 대해 적절한 시제를 선택해야 함.
    - 담화 관계 : 단어 개체간 의미론적 연결, 사건/사물 간 연결, 인과 관계 등 담화 관계가 명확한 경우 요약문도 해당 관계를 포함해야 한다.
    - 감정 : 종종 감정이 내포됨. → 요약문에 사건과 관련된 중요한 감정을 명시적으로 설명하도록 지시
    - 의도 식별 : 대화의 결과를 단순히 요약하기보단, 화자의 의도를 명확히 식별할 수 있는 경우 의도를 요약에 설명해야 됨.
        - 대화의 의도 = 화자가 대화를 시작하는 동기

        > *The intent here refers to the motivation of a speaker to initiate a conversation*

🔶DialogSum 결과 분석

Error Type	Transformer	UNILMV2$_{BASE}$
Incorrect Coref.	94%	60%
Missing Salient Inf.	64%	32%
Redundant Inf.	62%	44%
Unfactual Inf.	74%	22%
Syntactic Error	72%	22%

- 50개의 모델 생성 요약에 대한 오류 분석 결과
  1. 부정확한 coreference(지시 정보) → 모델이 상호 작용적인 정보 흐름 때문에 혼란스러워할 수 있음???

     > **발화자 정보 활용:**
     >
     > "너", "나"와 같은 1인칭, 2인칭 대명사는 발화자에 따라 지칭하는 대상이 달라지므로, 발화자 정보(#Person1#,...)를 모델 입력에 포함시키는 것이 필수적입니다.
     >
     > **QA 데이터셋으로 pre-trained**
     >
     > 된 모델로 전이학습을 시도
     >
     > **전통**
     >
     > 적인 방법 : 텍스트 쌍들을 먼저 coference에 해당하는지 파악하고 이를 feature로 반영

  2. 중복된 요약을 생성

     > 후처리 → 요약문의 각 sentence 간 유사성을 검사해 유사성이 threshold보다 높은 문장들은 가장 높은 유사성의 문장만 남기고 중복 제거한다.

- Case Study

1. Information Flow : 일반적으로 다중 발화자 대화문은 **중요한 정보가 분산**되어 있어 독백보다 요약하기 어렵다.
2. Regular Greetings : 대화 처음, 마지막의 인사말이 때때로 대화 주제에 따라 필수적인 의도를 표현하기도 한다.

> 인사말을 전처리, 후처리할 필요는 없을 듯?

3. 정보 분산의 이유 중 하나 : 대화문에서는 상대에 의한 발언의 중단이 자주 발생

> e.g. Excuse me... , Ummmm ... , Wait... 같은 텍스트
> "... " 앞에 존재하는 텍스트들을 추출하여 분석한 후, 필요하다면 불용어로 해당 텍스트들을 추가하여 영향력을 파악

4. 구어체 대화문이 많은 DialogSum에는 "지시 표현(coreference) & 생략(ellipsis)"가 많다. 이를 모델이 잘 파악하는 게 중요. →
5. 구어체 대화문에서 대화의 내용뿐만 아니라 화자의 행동을 요약할 수 있어야 한다. ⇒

> - 발화자 임베딩, 대화 구조 인코딩 ...
> - 계층적 학습 방법 : 발화(utterance) → 턴(turn) → 대화(dialogue)의 계층 구조를 학습시킨다.
> - 데이터 증강을 통해 유사한 화행을 갖는 여러 대화문을 생성하여 모델의 이해도를 높인다.