Dialogue Summarization I 일상 대화 요약

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

#비공개대회 #AI부트캠프14기 #NLPAdvanced

평가방법

모델 성능 평가

Dialogue Summarization task에서는 여러 인물들이 나눈 대화 내용을 요약하는 문제입니다. 예측된 요약 문장을 3개의 정답 요약 문장과 비교하여 metric의 평균 점수를 산출합니다. 본 대회에서는 ROUGE-1-F1, ROUGE-2-F1, ROUGE-L-F1, 총 3가지 종류의 metric으로부터 산출된 평균 점수를 더하여 최종 점수를 계산합니다.

해당 평가지표를 활용한 이유는 다음과 같습니다. DialogSum 데이터셋은 Multi-Reference Dataset으로 multi-reference에 대한 average를 보는 것이 중요합니다. 따라서 데이터셋의 특성에 맞추어 최종 점수 산출도 평균을 활용했습니다.

따라서, 3개의 정답 요약 문장의 metric 평균 점수를 활용하기에 metric 점수가 100점이 만점이 아니며, 3개의 정답 요약 문장 중 하나를 랜덤하게 선택하여 산출된 점수가 약 70점 정도임을 말씀드립니다.

더 자세한 Metric에 대한 설명은 아래와 같습니다.

Evaluation Metric

- ROUGE는 텍스트 요약, 기계 번역과 같은 태스크를 평가하기 위해 사용되는 대표적인 metric입니다. 모델이 생성한 요약본 혹은 번역본을 사람이 만든 참조 요약본과 비교하여 점수를 계산합니다.
  - ROUGE-Recall: 참조 요약본을 구성하는 단어들 중 모델 요약본의 단어들과 얼마나 많이 겹치는지 계산한 점수입니다.
  - ROUGE-Precision: 모델 요약본을 구성하는 단어들 중 참조 요약본의 단어들과 얼마나 많이 겹치는지 계산한 점수입니다.
- ROUGE-N과 ROUGE-L은 비교하는 단어의 단위 개수를 어떻게 정할지에 따라 구분됩니다.
  - ROUGE-N은 unigram, bigram, trigram 등 문장 간 중복되는 n-gram을 비교하는 지표입니다.

  - ROUGE-1는 모델 요약본과 참조 요약본 간에 겹치는 unigram의 수를 비교합니다.
  - ROUGE-2는 모델 요약본과 참조 요약본 간에 겹치는 bigram의 수를 비교합니다.
- ROUGE-L: LCS 기법을 이용해 최장 길이로 매칭되는 문자열을 측정합니다. n-gram에서 n을 고정하지 않고, 단어의 등장 순서가 동일한 빈도수를 모두 세기 때문에 보다 유연한 성능 비교가 가능합니다.

- ROUGE-F1은 ROUGE-Recall과 ROUGE-Precisioin의 조화 평균입니다.

ROUGE-N

$$\text{Recall} = \frac{\text{Gold와 Pred의 겹치는 N-gram의 수}}{\text{Gold의 N-gram의 수}} \qquad \text{Precision} = \frac{\text{Pred와 Gold의 겹치는 N-gram의 수}}{\text{Pred의 N-gram의 수}}$$

ROUGE-1

$$\text{Recall} = \frac{N_{(\text{the, cat, was, under, the, bed})}}{N_{(\text{the, cat, was, under, the, bed})}} = \frac{6}{6}$$

$$\text{Precision} = \frac{N_{(\text{under, the, bed, was, the, cat})}}{N_{(\text{under, the, bed, there, was, the, cat})}} = \frac{6}{7}$$

Gold (정답) : the cat was under the bed
Pred (모델) : under the bed there was the cat

ROUGE-2

$$\text{Recall} = \frac{N_{((\text{the, cat}),\ (\text{under, the}),\ (\text{the, bed}))}}{N_{((\text{the, cat}),\ (\text{cat, was}),\ (\text{was, under}),\ (\text{under, the}),\ (\text{the, bed}))}} = \frac{3}{5}$$

$$\text{Precision} = \frac{N_{((\text{under, the}),\ (\text{the, bed}),\ (\text{the, cat}))}}{N_{((\text{under, the}),\ (\text{the, bed}),\ (\text{bed, there}),\ (\text{there, was}),\ (\text{was, the}),\ (\text{the, cat}))}} = \frac{3}{6}$$

Gold (정답) : the cat was under the bed
Pred (모델) : under the bed there was the cat

ROUGE-L

$$\text{Recall} = \frac{\text{가장 긴 공통부분 문자열의 길이}}{\text{Gold의 1-gram의 수}} = \frac{N_{(\text{under the bed})}}{N_{(\text{the, cat, was, under, the, bed})}} = \frac{3}{6}$$

$$\text{Precision} = \frac{\text{가장 긴 공통부분 문자열의 길이}}{\text{Pred의 1-gram의 수}} = \frac{N_{(\text{under the bed})}}{N_{(\text{under, the, bed, there, was, the, cat})}} = \frac{3}{7}$$

Gold (정답) : the cat was under the bed
Pred (모델) : under the bed there was the cat

ROUGE-F1

$$F1 = 2 \times \frac{\text{ROUGE recall} \times \text{ROUGE precision}}{\text{ROUGE recall} + \text{ROUGE precision}}$$

$$\begin{aligned}\text{Final Score} &= \max_{i}^{N} \text{ROUGE-1-F1}(\text{pred}, \text{gold}_i) \\ &+ \max_{i}^{N} \text{ROUGE-2-F1}(\text{pred}, \text{gold}_i) \\ &+ \max_{i}^{N} \text{ROUGE-L-F1}(\text{pred}, \text{gold}_i)\end{aligned}$$

문장 토큰화

한국어 데이터 특성 상 정확한 ROUGE score 산출하기 위하여 문장 토큰화를 진행한 후 평가합니다.
한국어 형태소 분석기를 통해 의미를 갖는 최소한의 단위인 형태소 단위로 문장을 쪼갠 뒤 모델이 생성한 문장과 정답 문장을 비교하여 ROUGE score를 산출합니다.

- 문장 토큰화 예시

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

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

upstage AI Stages

contact@upstage.ai
경기도 용인시 수지구 광교중앙로 338, A815 (상현동, 광교우미뉴브)
upstage.ai

한국어 English

Privacy policy Terms