Paper summary

1. re-evaluating evalustion in text summarization

**기존 평가지표:**

* ROUGE, BERTScore, MoverScore 등

**기존 평가지표의 한계:**

* 자동 측정 지표는 인간 판단과 일치하지 않을 수 있으며, 특히 요약의 질적 측면을 평가하는 데 한계가 있다.
* 자동 측정 지표는 시스템 간 성능 차이를 정확하게 반영하지 못할 수 있다.
* 자동 측정 지표는 요약의 다양성과 창의성을 평가하는 데 한계가 있다

**한계 극복 방법**:

* 자동 평가지표과 함께 인간 판단 데이터 셋을 활용해 메타-평가 지표 계산

1. SummEval: Re-evaluating Summarization Evaluation

**기존 평가 데이터 셋:**

TAC(Text Analysis Conference), DUC(Document Understanding Conference)

**평가 데이터 셋의 문제점:**

현재 요약 시스템에서 사용되는 점수보다 낮은 척도로 모델 출력에 대한 인간 판단을 포함한다. (즉 시스템에 의한 요약이 인간에 의한 요약보다 품질이 좋다). 이러한 문제로 인해, 현재 사용되는 평가 척도의 성능이 새로운 환경에서 의심스러워졌다. 또 다른 결점으로는 기존 평가 방식이 일관성, 정보성, 의미론적 일치성 및 새로운 정보와 같은 다양한 측면을 고려하지 한다. (참조: 페이지 2)

**기존 문제 해결 방법:**

CNN/DailyMail 데이터셋에서 무작위로 선택된 100개의 기사에 대한 16개의 최신 신경망 요약 모델의 요약 평가를 수집하였다. 이러한 모델 출력은 Kryściński et al. (2019)와 같이 네 가지 차원(일관성, 정보성, 의미론적 일치성 및 새로운 정보)으로 평가하였다. 이러한 자동 평가 척도는 SummEval 툴킷을 사용하여 일관된 방식으로 벤치마크 하였다.

Summeval:

* ROUGE, BLEU, METEOR, BERTScore 등과 같은 다양한 정량적 평가 지표를 제공
* 인간 평가자가 생성된 요약과 참조 요약을 직접 비교하고 평가할 수 있는 기능을 제공

기존 평가 지표:

ROUGE (Recall-Oriented Understudy for Gisting Evaluation):

설명: ROUGE는 주로 텍스트 요약에서 사용되는 지표로, 생성된 요약과 참조 요약 간의 n-그램 일치를 측정한다. ROUGE-N, ROUGE-L, ROUGE-W 등 다양한 변형이 있다.

계산 방법: ROUGE-N은 생성된 요약과 참조 요약 간의 n-그램 일치를 계산한다. 일치하는 n-그램의 수를 참조 요약의 전체 n-그램 수로 나누어 recall을 구하고, 이를 평균하여 최종 점수를 산출한다.

BLEU (Bilingual Evaluation Understudy):

설명: BLEU는 주로 기계 번역에서 사용되는 지표로, 생성된 번역과 참조 번역 간의 n-그램 정확도를 측정한다.

계산 방법: 생성된 번역과 참조 번역 간의 n-그램 일치를 계산하고, 생성된 번역의 전체 n-그램 수로 나누어 정확도를 구한다. 또한, 생성된 번역의 길이가 참조 번역보다 짧을 경우를 고려한 브레버티 페널티를 적용하여 최종 점수를 산출한다.

METEOR (Metric for Evaluation of Translation with Explicit ORdering):

설명: METEOR는 기계 번역에서 사용되는 지표로, 생성된 번역과 참조 번역 간의 단어 정렬과 동시에 일치하는 n-그램을 고려한다. 이 지표는 단어 순서를 고려하여 평가하기 때문에 정확성과 일관성을 함께 고려할 수 있다.

계산 방법: 생성된 번역과 참조 번역 간의 단어 정렬을 구한 다음, 일치하는 n-그램을 고려하여 정확성과 recall을 계산한다. 이러한 값을 고려하여 조화 평균을 구하고, 패널티를 적용하여 최종 점수를 산출한다.