EVAL +

날짜	@2026/01/27

▼ G-EVAL

▼ Background

- 인공지능이 생성한 텍스트의 품질을 자동으로 정량 평가하는 것은 본질적으로 어려운 문제
- BLEU, ROUGE 등 전통적인 레퍼런스 기반 메트릭은 새로운 작업에 적용하기 어렵고, 인간 판단 및 평가와 낮은 상관관계를 보이는 것으로 보고
- 최근에는 정답 없이도 LLM을 활용한 NLG 평가 방식이 제안되었으나, 여전히 신뢰성 관련 한계가 제기

→ 이러한 한계를 극복하고, 평가 과정에서 발생할 수 있는 편향을 완화하며, 인간 평가와의 높은 일치도를 제공하기 위한 새로운 LLM 기반 자동 평가 방법이 제안

▼ Method

- 구성요소
  - *Prompt for NLG Evaluation*
    - 평가 작업과 평가 기준을 정의하는 자연어 지침

```
# TASK: 요약
You will be given one summary written for a news article.
Your task is to rate the summary on one metric.
Please make sure you read and understand these instructions carefully.
Please keep this document open while reviewing, and refer to it as need
ed.

# 기준: 일관성
Evaluation Criteria:
Coherence (1-5) - the collective quality of all sentences. We align thi
s dimension with the DUC quality question of structure and coherence wh
ereby ”the summary should be well-structured and well-organized. The su
mmary should not just be a heap of related information, but should buil
d from sentence to sentence to a coherent body of information about a t
opic.”
```

  - *Auto Chain-of-Thoughts for NLG Evaluation*
    - LLM이 텍스트를 평가할 때 더 많은 컨텍스트와 지침을 제공할 수 있고, 평가 과정과 결과를 설명하는 데에도 도움이 됨

```
1. Read the news article carefully and identify the main topic and key
points.
2. Read the summary and compare it to the news article. Check if the su
mmary covers the main topic and key points of the news article, and if
it presents them in a clear and logical order.
3. Assign a score for coherence on a scale of 1 to 5, where 1 is the lo
west and 5 is the highest based on the Evaluation Criteria.
```

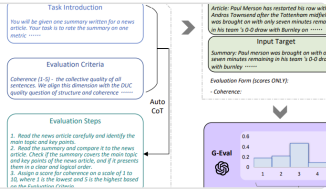
  - *Scoring Function*

        - *'Prompt for NLG Evaluation'*, *'Auto Chain-of-Thoughts'*, *'input context'*, *'target text'* 를 입력으로 LLM 호출
        - *'target text'* 을 생성할 조건부 확률을 메트릭으로 사용하는 GPTScore 와는 달리, G-EVAL은 *form-filling* 방식으로 평가 작업 수행
    - Limitations
        - 일부 평가 작업에서 하나의 숫자가 점수 분포를 지배 → 점수의 분산 저하 및 인간의 상관관계가 낮아지는 결과 초래
        - 프롬프트 내 소수 값 요청을 명시하더라도, 보통 정수 점수만 출력 → 동점 상황 발생
    - Solutions
        - LLM의 출력 토큰 확률을 사용하여 점수를 정규화하고, 가중합을 취하는 방법 제안

$$score = \sum_{i=1}^{n} p(s_i) \cdot s_i$$

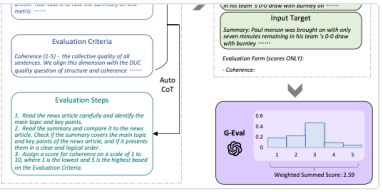
▼ References

[논문리뷰] G-Eval: LLM을 사용해 인간의 견해와 보다 일치하는 NLG 평가 시스템 구축하기
지난 포스트에서는 LLM 기반의 시스템을 평가하는 방법에 대해 알아보았다 LLM Evaluation | LLM 기반의 시스템을 어떻게 평가할 수 있을까 지난 포스팅에서 다루었던 것처럼 LLM의 문맥 이해 및 자연어 생성 능력 능력이 향상되었고, fine-tuning API, Plug-in 지원 등이 이루어지면서 다양한 애플리케이션 개
https://littlefoxdiary.tistory.com/123



G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment
The quality of texts generated by natural language generation (NLG) systems is hard to measure automatically. Conventional reference-based metrics, such as BLEU and ROUGE, have been shown to have...
https://arxiv.org/abs/2303.16634

G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment
https://arxiv.org/abs/2303.16634 G-Eval: NLG Evaluation using GPT-4 with Better Human AlignmentThe quality of texts generated by natural language generation (NLG) systems is hard to measure automatically. Conventional reference-based
https://dytis.tistory.com/86



▼ RAGAs

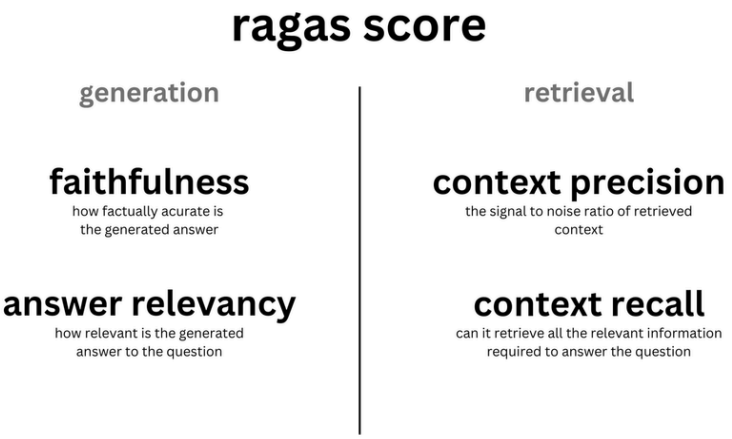
▼ Background

- 일단 RAG에는 평가에 영향을 미치는 요소가 너무 많아 정확한 평가가 어려움
    - *e.g. retrieving, generator, prompt*
- 그래서 이전에는 일반 언어 모델의 평가와 동일한 방식으로 이루어짐
    - e.g. *(Q, A)* 데이터쌍을 만들어, *retrieval system*이 *ground truth answer*를 검색했는가의 여부를 기준으로 정확도(*accuracy, recall, precision*) 기반 성능을 평가

→ 이를 해결하기 위해 RAGAs는 자동화된 평가 프레임워크를 제안

▼ Method

RAGAs는 주요 성능 매트릭을 두 측면(Retrieval, Generation)에서 정의하고 있음



▼ Generation

- faithfulness
  - *a_s(q)*가 얼마나 *c(q)*에 기반하여 hallucination 없이 생성되었는지를 평가하는 지표

$$Faithfulness = \frac{\text{Number of Correct Facts in the Response}}{\text{Total Number of Facts in the Response}}$$

- answer relevancy
  - *a_s(q)*가 유저 질문(*q*)와 얼마나 관련이 있는지를 평가하는 지표

$$Answer\ relevancy = \frac{\text{Number of Relevant concepts in the Response}}{\text{Total Number of concepts in the Response}}$$

▼ Retrieval

- context recall
  - 검색된 *c(q)*에 포함된 내용이 생성된 답변 *a_s(q)* 누락없이 잘 반영되었는지를 평가하는 지표

$$CR_{recall} = \frac{N_{\text{relevant facts in } c(q)}}{N_{\text{relevant facts needed for } q}}$$

- context precision
  - *c(q)*에 대해 질문에 대한 답변과 관련있는 문서가 얼마나 상위에 랭크되어있는지 측정하는 지표

$$CR_{precision} = \frac{N_{\text{relevant facts in } c(q)}}{N_{\text{total facts in } c(q)}}$$

▼ References

[논문리뷰] RAGAs: Automated Evaluation of Retrieval Augmented Generation

🥲 →0. Abstract RAGAs(RAG Assessment), 이 프레임워크는 사람이 만드는 ground truth 데이터 없이도 RAG를 평가할 수 있게 해주는 'reference-free' evaluation framework이다. RAG는 knowledge-intensive task에서 LLM이 사용가능한 지식을 제공해주고, hallucination 문제

🔻 https://song9ski-program.tistory.com/entry/%EB%85%BC%EB%AC%B8%EB%A6%AC%EB%B7%B0-RAGAs-Automated-Evaluation-of-Retrieval-Augmented-Generation

RAGAS for RAG in LLMs: A Comprehensive Guide to Evaluation Metrics.

Introduction

https://dkaarthick.medium.com/ragas-for-rag-in-llms-a-comprehensive-guide-to-evaluation-metrics-3aca142d6e38