Corrective Retrieval Augmented Generation

2024.02.14
Lab Seminar
Park Kieun

HUMAN
CENTERED
COMPUTING
LABORATORY

Corrective Retrieval Augmented Generation

Shi-Qi Yan[1*], **Jia-Chen Gu**[2*], **Yun Zhu**[3], **Zhen-Hua Ling**[1]
[1]National Engineering Research Center of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China
[2]Department of Computer Science, University of California, Los Angeles
[3]Google Research
yansiki@mail.ustc.edu.cn, gujc@ucla.edu, yunzhu@google.com, zhling@ustc.edu.cn

간단 소개 (Recap)

- **Corrective Retrieval Augmented Generation** (Yan, Shi-Qi, et al., 24.01.29)
  - 수정 검색 증강 생성(CRAG)
  - 검색기의 자가 수정 구성 요소를 구현하고 증강 생성을 위해 검색된 문서의 활용도를 향상
  - 경량 검색 평가기(lightweight retrieval evaluator)는 쿼리에 대해 **검색된 문서의 전반적인 품질을 평가**하여 다양한 지식 검색 작업을 트리거할 수 있는 **신뢰도를 반환**하도록 설계
    - {Correct, Incorrect, Ambiguous}
  - 검색된 문서가 핵심 정보에 선택적으로 집중하고 **관련 없는 정보를 걸러**낼 수 있도록 분해 후 재구성 알고리즘 설계
  - 짧은 형식과 긴 형식의 생성 작업을 포함하는 4개의 데이터셋에 대한 실험 결과, CRAG가 RAG 기반 접근법의 성능을 크게 향상

CRAG (Corrective Retrieval Augmented Generation)

- Motivation
  - 왜 부정확한 검색 결과를 그대로 쓰지?
  - 왜 검색 결과를 통으로 넣지?
- Method
  - Retrieval Evaluator
  - Knowledge Refinement
  - Knowledge Searching (query rewrite)
- Task
  - PopQA (단답형) – Accuracy
  - Biography (서술형) - FactScore
  - PubHealth (T/F) - Accuracy
  - Arc-Challenge (객관식) - Accuracy

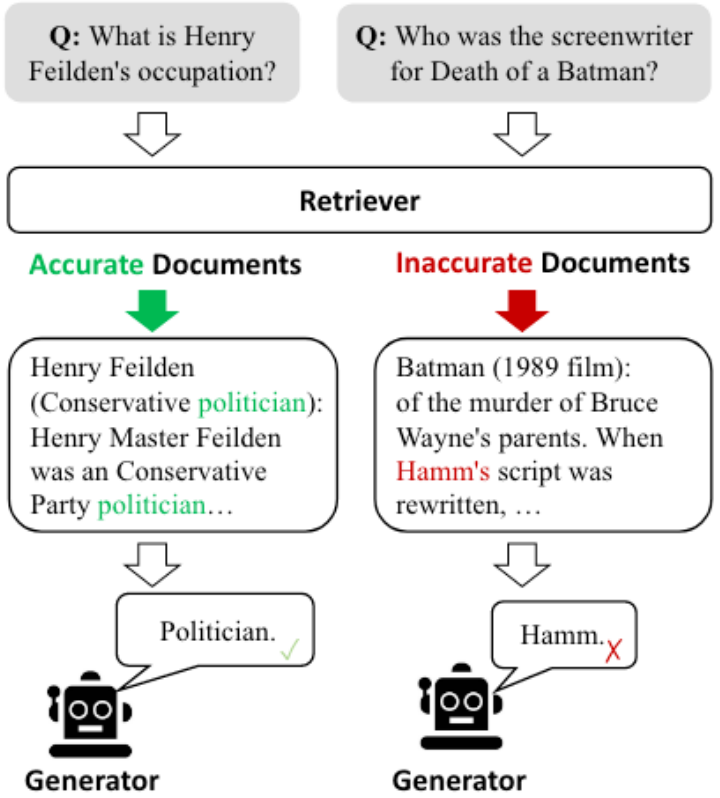


Figure 1: The examples show that a low-quality retriever is prone to introducing a substantial amount of irrelevant information, impeding the generators from acquiring accurate knowledge and potentially misleading them.

CRAG overview

1. Input query 로 DR (any retriever)
2. Retrieval Evaluation
    1. Correct → Retrieved Document 에서 연관된 정보만 뽑기
    2. Incorrect → Retrieved Documents 버리고 web search
    3. Ambiguous → 둘 다
3. Query와 optimized retrieval result 를 generative model 에 넣어 답을 얻기

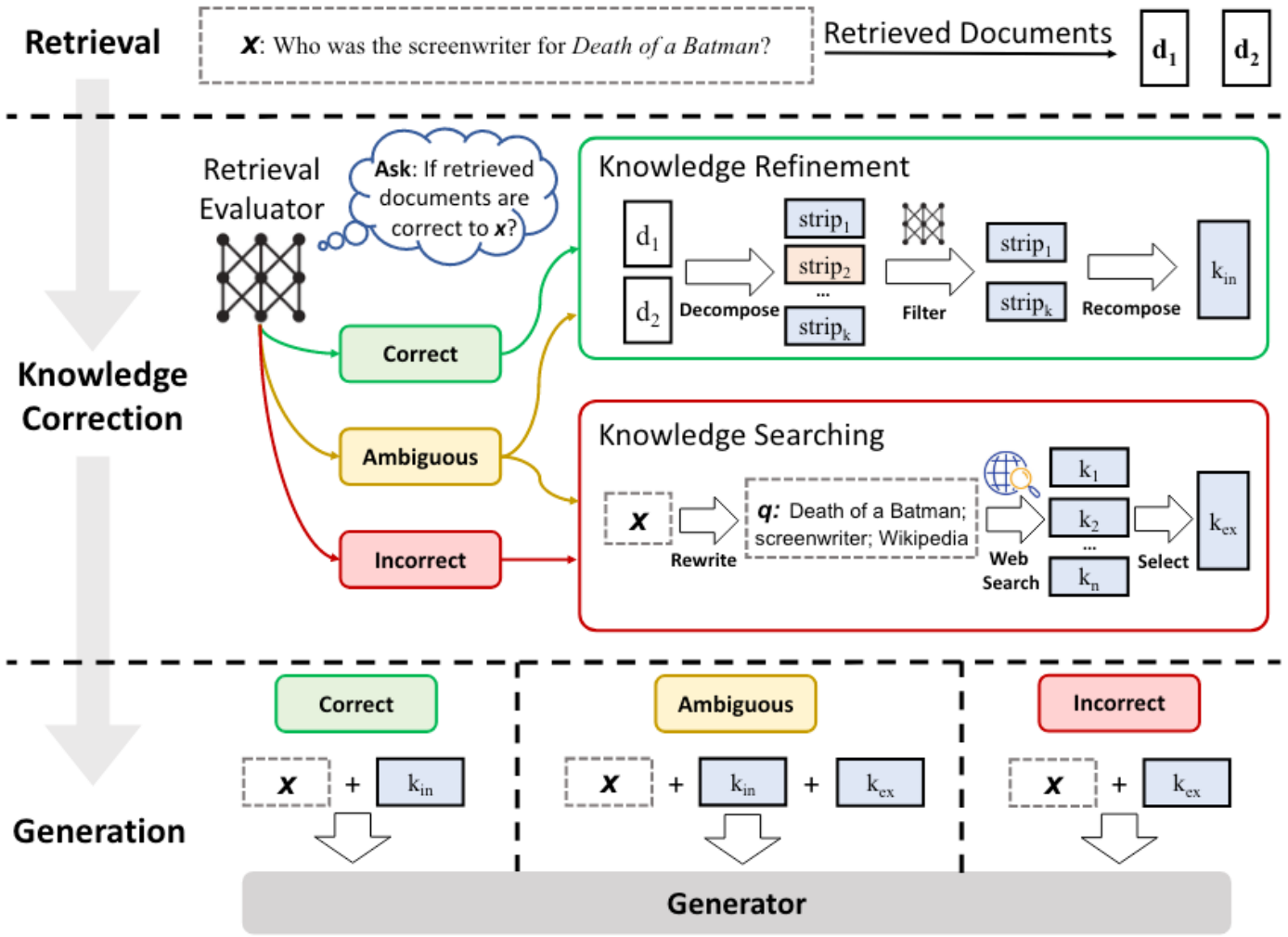


Figure 2: An overview of CRAG at inference. A retrieval evaluator is constructed to evaluate the relevance of the retrieved documents to the input, and estimate a confidence degree based on which different knowledge retrieval actions of {Correct, Incorrect, Ambiguous} can be triggered.

Retrieval Evaluator

- workflow
  - Retrieve 10 document
  - Concat question and document (question-document pair)
  - Predict relevance score [-1~1]
- Fine-tuning
  - Positive sample : Label [1]
  - Negative sample : Label [-1]

Retrieval Evaluator

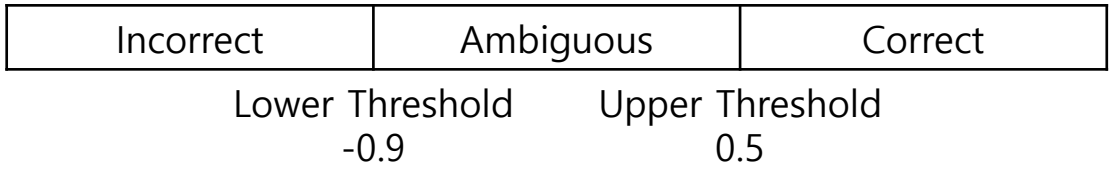
- Lightweight Retrieval Evaluator
  - Fine-tuned T5-large (0.77B)
  - No need Human or LLM annotation (for fine-tunning)
- Critic model (Self-RAG, Asai el al., 2023)
  - Fine-tuned LLaMA-2 (7B)
  - Requires GPT-4 annotated data (for instructi on tuning)
- Chat GPT

	Accuracy
Our Retrieval Evaluator (T5-based)	84.3
ChatGPT	58.0
ChatGPT-CoT	62.4
ChatGPT-few-shot	64.7

Table 4: Evaluation of our retrieval evaluator and ChatGPT for the retrieval results on the PopQA dataset.

Action Trigger

- Relevence score (=confidence score) 를 바탕으로 다음 action 결정



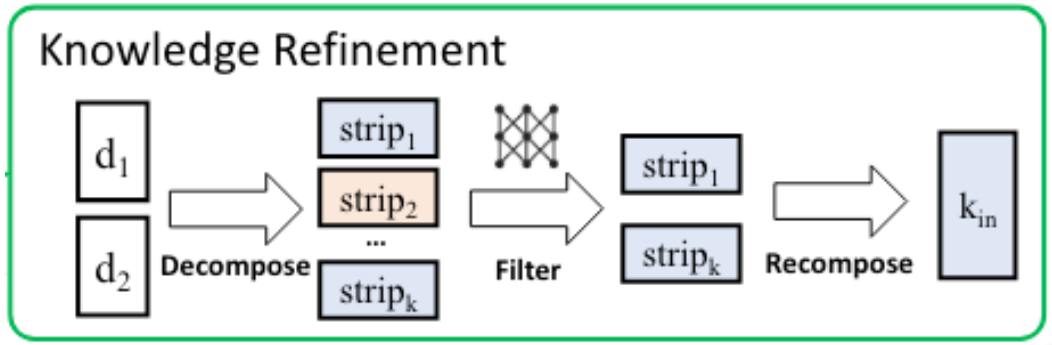
$$S = \max\left(\frac{TP + TN}{n_{total}} - \alpha\frac{FP}{n_{neg}}\right)$$

Threshold set for each benchmark:
maximize the overall accuracy
with a false-positive punishment

if ***Confidence*** == *[CORRECT]* **then**
    Internal_knowledge = Knowledge_Refine(x, D)
    k = Internal_knowledge
else if ***Confidence*** == *[INCORRECT]* **then**
    External_knowledge = Web_Search(W Rewrites x for searching)
    k = External_knowledge
else if ***Confidence*** == *[AMBIGUOUS]* **then**
    Internal_knowledge = Knowledge_Refine(x, D)
    External_knowledge = Web_Search(W Rewrites s for searching)
    k = Internal_knowledge + External_knowledge

Knowledge Refinement

- Extract the most critical knowledge strips (internal knowledge)
  1. Segmentation → fine-grained knowledge strips (heuristic rules)
     - 적당히 몇 문장으로 구성되도록 나누었음
  2. Calculate relevence score (아까 만든 retrieval evaluator 이용)
     - Top-k = 5
     - Filter threshold = -0.5
  3. Relevant ones 만 모아서 원래 순서 유지하며 concat



Web Search

- Questions to query (ChatGPT) →
- Using Google Search API
  - Top-k = 5
- Employ same knowledge refinement method

Extract at most three keywords separated by comma from the following dialogues and questions as queries for the web search, including topic background within dialogues and main intent within questions. question: What is Henry Feilden's occupation? query: Henry Feilden, occupation question: In what city was Billy Carlson born? query: city, Billy Carlson, born question: What is the religion of John Gwynn? query: religion of John Gwynn question: What sport does Kiribati men's national basketball team play? query: sport, Kiribati men's national basketball team play question: [question] query:

Table 5: The few-shot prompt to GPT-3.5 Turbo for generating knowledge keywords as web search queries.

Evaluation

- Tasks
  - PopQA (단답형) – Accuracy
  - Biography (서술형) - FactScore
  - PubHealth (T/F) - Accuracy
  - Arc-Challenge (객관식) - Accuracy

- Methods
  - 여러 모델에 Retrieval 을 넣었을 때와 넣지 않았을 때 비교
  - Generation model 을 바꿔꼈을 때의 비교
  - Standard RAG 와 CRAG 의 비교
  - Self-RAG 와 Self-CRAG(Self-RAG + CRAG)의 비교

Result

Method	PopQA (Accuracy)	Bio (FactScore)	Pub (Accuracy)	ARC (Accuracy)
LMs trained with propriety data				
LLaMA2-c$_{13B}$	20.0	55.9	49.4	38.4
Ret-LLaMA2-c$_{13B}$	51.8	79.9	52.1	37.9
ChatGPT	29.3	71.8	70.1	**75.3**
Ret-ChatGPT	50.8	-	54.7	**75.3**
Perplexity.ai	-	71.2	-	-
Baselines without retrieval				
LLaMA2$_{7B}$	14.7	44.5	34.2	21.8
Alpaca$_{7B}$	23.6	45.8	49.8	45.0
LLaMA2$_{13B}$	14.7	53.4	29.4	29.4
Alpaca$_{13B}$	24.4	50.2	55.5	54.9
CoVE$_{65B}$	-	71.2	-	-

Method	PopQA (Accuracy)	Bio (FactScore)	Pub (Accuracy)	ARC (Accuracy)
Baselines with retrieval				
LLaMA2$_{7B}$	38.2	78.0	30.0	48.0
Alpaca$_{7B}$	46.7	76.6	40.2	48.0
SAIL	-	-	69.2	48.4
LLaMA2$_{13B}$	45.7	77.5	30.2	26.0
Alpaca$_{13B}$	46.1	77.7	51.1	57.6
LLaMA2-hf-7b				
RAG	37.7	44.9	9.1	23.8
CRAG	39.8	47.7	9.1	25.8
Self-RAG*	29.0	32.2	0.7	23.9
Self-CRAG	49.0	69.1	0.6	27.9
SelfRAG-LLaMA2-7b				
RAG	40.3	59.2	39.0	46.7
CRAG	59.3	74.1	**75.6**	54.8
Self-RAG	54.9	81.2	72.4	67.3
Self-CRAG	**61.8**	**86.2**	74.8	67.2

Ablation Study

	LLaMA2-hf-7b	SelfRAG-LLaMA2-7b
CRAG	47.3	59.3
w/o. Correct	44.5	58.1
w/o. Incorrect	46.8	58.6
w/o. Ambiguous	45.7	58.5
Self-CRAG	49.0	61.8
w/o. Correct	43.6	59.6
w/o. Incorrect	47.7	60.8
w/o. Ambiguous	48.1	61.5

Table 2: Ablation study for removing each single action on the PopQA dataset in terms of accuracy. When the action Correct or Incorrect was removed, the proportion that originally triggered Correct or Incorrect would trigger Ambiguous. When Ambiguous was removed, all input queries clearly triggered Correct or Incorrect.

	LLaMA2-hf-7b	SelfRAG-LLaMA2-7b
CRAG	47.3	59.3
w/o. refinement	38.9	47.0
w/o. rewriting	44.8	56.6
w/o. selection	44.0	53.8
Self-CRAG	49.0	61.8
w/o. refinement	35.9	52.2
w/o. rewriting	37.2	58.4
w/o. selection	57.9	57.9

Table 3: Ablation study for removing each knowledge utilization operation on the PopQA dataset in terms of accuracy. Removing document refinement denoted that the original retrieved documents were directly fed to the generator. Removing search query rewriting denoted that questions were not rewritten into queries consisting of keywords during knowledge searching. Removing knowledge selection denoted that all searched content of web pages was all regarded as the external knowledge.

Contributions

- First attempt to design corrective strategies for RAG
- Introduce Plug-and-play method “CRAG”
- Adaptability to RAG-based approaches
- generalizability across short- and long-form generation task

QA

HUMAN
CENTERED
COMPUTING
LABORATORY