Tree of Clarifications: Answering Ambiguous Questions with Retrieval-Augmented Large Language Models

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, Jaewoo Kang

Korea University, NAVER Cloud, NAVER AI Lab, KAIST AI, University of Richmond

EMNLP 2023

발표자: 송선영

2024/05/13

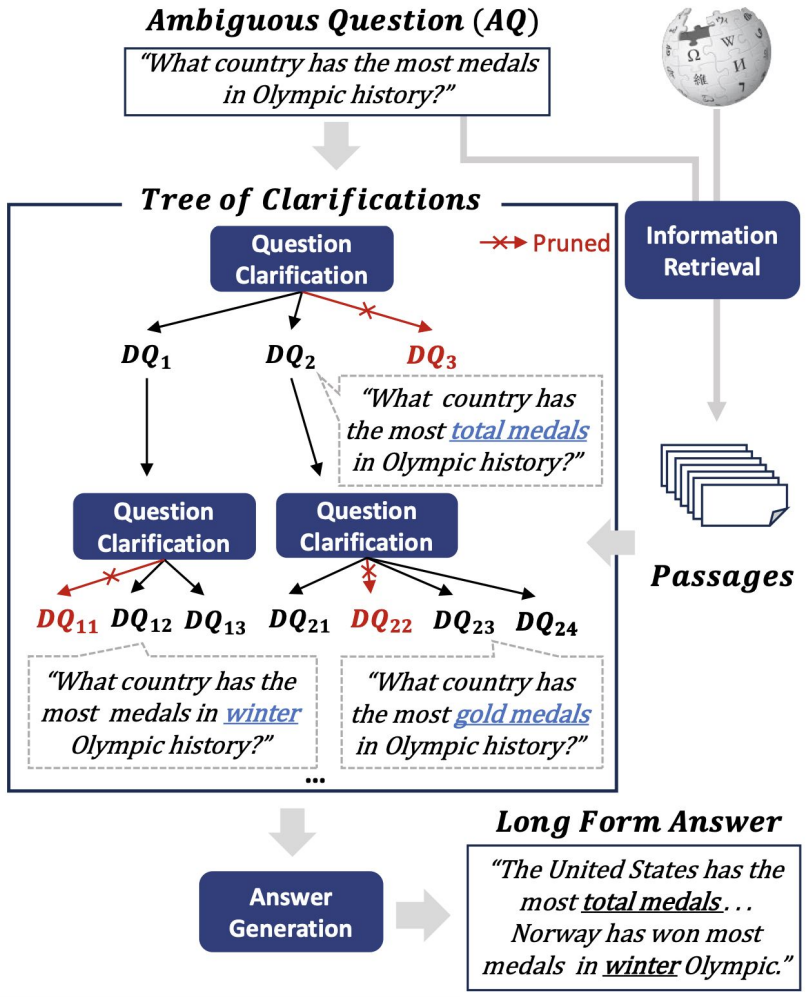
Introduction

- ODQA task에는 여러 방식으로 해석할 수 있는 모호한 질문(Ambiguous Question: AQ)을 하는 경우가 많음
- 기존에 AQ를 해결하기 위해서는, 주어진 AQ에 대해 명확한 질문(Disambiguated Questions: DQ)를 구하고, 이에 대한 long-form answer를 생성해야 함
- 기존 방식의 문제점
  1. Multiple dimensions of ambiguity를 고려해 AQ를 명확히 해야 함
  2. DQ와 그에 대한 답(DA)를 식별하기 위해서는 상당한 지식이 필요

  (예시) AQ: “올림픽 역사상 가장 많은 메달을 획득한 나라는 어디인가?”

-> 이 논문에서는 이 문제를 해결하고 AQ에 대해 long-form answer를 제공할 수 있는 ToC 프레임워크를 제안

Tree of Clarifications (ToC)

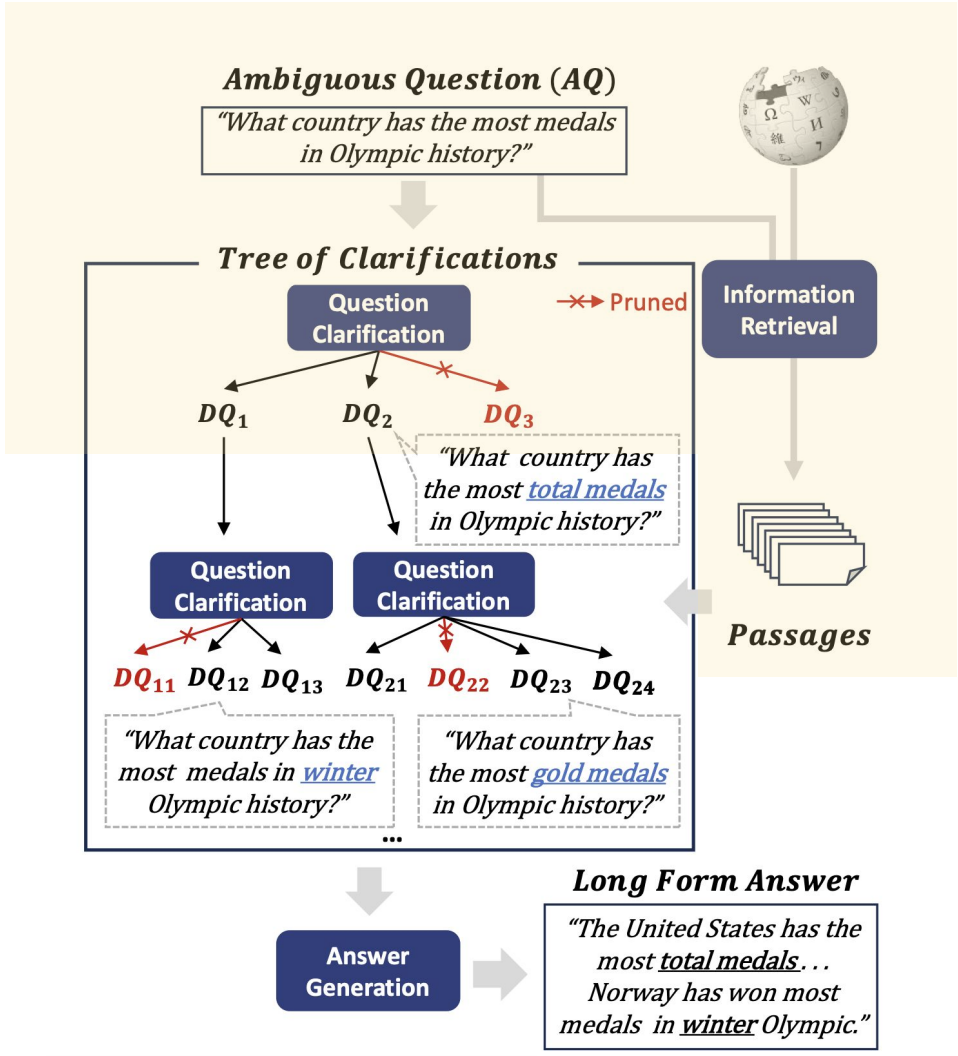


Contributions of ToC

1. LLM이 트리 구조에서 AQ를 설명할 수 있는 다양한 경로를 탐색하게 하고, 도움이 되지 않는 DQ를 잘라낼 수 있음

2. 최초로 LLM과 retrieval system을 결합해 AQ에 대한 long-form answer를 생성하는 방법을 제안

Tree of Clarifications (ToC)



1) Retrieval-Augmented Clarifiacation (RAC)

- Retrieval system을 사용해 AQ에 관련된 Wikipedia document를 검색
  - ColBERT, Bing Search Engine 사용
  - 총 200개 이상의 passage를 구축

- passage set을 수집 후, SentenceBERT를 활용해 rerank하고 top-k개의 passage를 선택해 prompt에 추가
- LLM이 위 prompt로, 가능한 모든 DQ와 그에 해당하는 답(DA)을 생성

Tree of Clarifications (ToC)

I will provide ambiguous questions that can have multiple answers based on their different possible interpretations. Clarify the given question into several disambiguated questions and provide short factoid answers to each question. Subsequently, summarize them into a detailed long-form answer of at least three sentences. Here are some examples.

Context:
[1] Fred and George Weasley | Fred and George Weasley are fictional characters in the "Harry Potter" book series written by J. K. Rowling. The characters are the identical twin brothers ... The twins were played by identical twin brothers James and Oliver Phelps in the film adaptations
[2] James and Oliver Phelps | James Andrew Eric Phelps and Oliver Martyn John Phelps (born 25 February 1986) are English actors and identical twin brothers. They are known for playing Fred and George Weasley in the "Harry Potter" film series from 2001 to 2011 ...
...
[5] Chris Rankin | plays of "Bugsy Malone" and "The Lion, The Witch and The Wardrobe". His professional acting career began when he won the role of Percy Weasley in September 2000 ... after his character's absence from "Harry Potter and the Goblet"

Question: Who played the weasley brothers in harry potter?

Disambiguations:
DQ 1: Who played the fictional characters Fred and George Weasley in the "Harry Potter" book series?
DA 1: James and Oliver Phelps
DQ 2: Who are the English actors and identical twin brothers known for playing Fred and George Weasley in the "Harry Potter" film series?
DA 2: James and Oliver Phelps
DQ 3: Who is the actor that plays Percy Weasley in the Harry Potter series?
DA 3: Chris Rankin

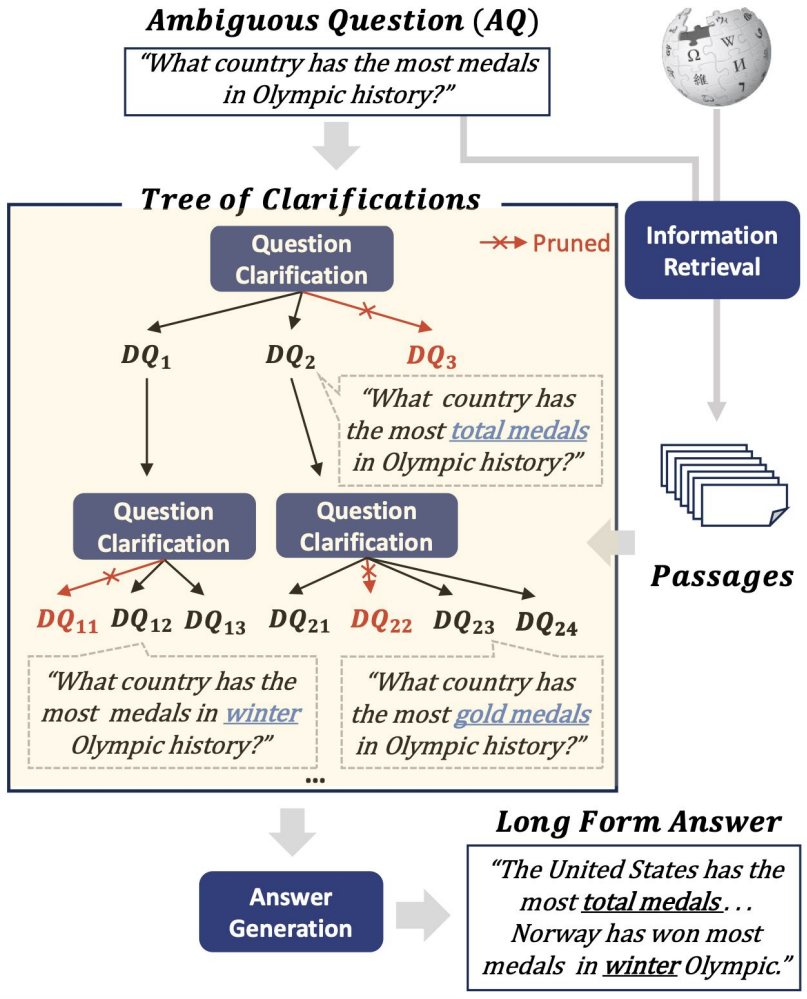
Answer: The Weasley brothers in the Harry Potter series were played by identical twin brothers James and other Oliver Phelps. The Phelps brothers are English actors who have appeared in films and TV shows together as a duo. Chris Rankin is the actor who played Percy Weasley in the Harry Potter series. Rankin is a British actor who has appeared in theatre, film, and TV.

1) Retrieval-Augmented Clarifiacation (RAC)

- Retrieval system을 사용해 AQ에 관련된 Wikipedia document를 검색
  - ColBERT, Bing Search Engine을 모두 사용
  - 각 retrieval system에서 검색된 passage를 결합해 총 200개 이상의 passage를 구축

- AQ를 위한 passage set을 수집 후, SentenceBERT를 활용해 rerank하고 top-k개의 passage를 선택해 prompt에 추가
- LLM이 위 prompt로, 가능한 모든 DQ와 그에 해당하는 답(AQ)를 생성

Tree of Clarifications (ToC)



2) Tree Structure (TS)

- 모호성의 다양한 차원을 탐색하기 위해, 트리 구조를 도입

- AQ가 있는 root node에서 시작해, RAC를 재귀적으로 실행하며 child node를 생성
- 각 expansion step에서 현재 query에 대해 passage가 reranking 됨
- valid node의 최대 수 또는 최대 깊이에 도달하면 트리 탐색 종료

Tree of Clarifications (ToC)

Correct Case 1
DQ: Who was selected to host the 2018 FIFA World Cup?

I will provide a question, relevant context, and proposed answer to it. Identify whether the proposed answer could be correct answers or not with only 'True' or 'False'

Context:
2018 and 2022 FIFA World Cup bids | FIFA's headquarters in Zurich. Russia was chosen to host the 2018 World Cup, and Qatar was chosen to host the 2022 World Cup. This made Russia the first Eastern European country to host the World Cup, while Qatar would be the first Middle Eastern country to host the World Cup. Blatter noted that the committee had decided to "go to new lands" and reflected a desire to "develop football" by bringing it to more countries. In each round a majority of twelve votes was needed. If no bid received 12 votes in a round, the bid with the fewest votes

Question: Who is hosting the next world cup 2022?

Proposed Answer: Russia

`False`

Correct Case 2
DQ: Which player has won the most World Series in baseball?

I will provide a question, relevant context, and proposed answer to it. Identify whether the proposed answer could be correct answers or not with only 'True' or 'False'

Context:
World Series ring | on World Series rings. The New York Yankees Museum, located in Yankee Stadium, has an exhibit with replicas of all Yankees' World Series rings, including the pocket watch given after the 1923 World Series. Yogi Berra won the most World Series rings with 10, as a player. Frankie Crosetti won 17 as a player and as a coach. Yogi Berra Museum and Learning Center. World Series ring A World Series ring is an award given to Major League Baseball players who win the World Series. Since only one Commissioner's Trophy is awarded to the team, a World Series ring is

Question: Who's won the most world series in baseball?

Proposed Answer: Yogi Berra

`True`

2-1) Pruning with Self-Verification

- 도움이 되지 않는 node를 제거하기 위해, self-verification을 수행

- LLM에게 prompt를 주어, Target node의 답변과 root node의 AQ 가 일치하는지, 사실적 일관성을 확인
- LLM은 node pruning을 결정하기 위해, 'True' or 'False' 답변을 생성
  - 만약, 원본과 다르거나 관련 없는 사실을 묻는 DQ가 생성된 경우라면, 이 DQ는 제거하게 됨

Tree of Clarifications (ToC)

I will provide ambiguous questions that can have multiple answers based on their different possible interpretations. Clarify the given question into several disambiguated questions and provide short factoid answers to each question. Subsequently, summarize them into a detailed long-form answer of at least three sentences. Here are some examples.

Context:
[1] Game of Thrones | Game of Thrones Game of Thrones is an American fantasy drama television series created by David Benioff and D. B. Weiss. ... and its seventh season ended on August 27, 2017. The series will conclude with its eighth season
[2] Game of Thrones | Game of Thrones is an American fantasy drama television series created by David Benioff and for HBO. It is an adaptation of "A Song of Ice and Fire", ... Set on the fictional continents of Westeros and Essos, "Game of Thrones" has a large ensemble cast
. . .
[5] A Game of Thrones (comics) | A Game of Thrones (comics) A Game of Thrones is the comic book adaptation of George R. R. Martin's fantasy novel "A Game of Thrones", . . . It is intended to follow the story and atmosphere of the novel closely, at a rate of about a page of art for each page of text, and

Question: What kind of series is game of thrones?

Disambiguations:
DQ 1: What is the genre of the American television series Game of Thrones?
DA 1: fantasy drama
DQ 2: What is the genre of the comic book series A Game of Thrones?
DA 2: fantasy
. . .
DQ 10: What is the genre of the board game A Game of Thrones?
DA 10: strategy

Answer: `There are multiple works that share the title Game of Thrones. The first is a television series that is a fantasy drama, the second is a comic book series that is fantasy, the third is a book series that is fantasy, and the fourth is a board game that is a strategy game.`

2-2) Answer Generation

- ToC는 모든 valid node를 aggregate 하고, AQ에 대한 포괄적인 long-form answer를 생성
- 만약 node의 수가 부족하다면, pruning step에서 제거된 것들 중 root node에 가장 가까운 순서대로 복구

- 최대 10개의 DQ-DA pair와 5개의 관련 passage를 입력으로 사용

Experimental Setup

- Datasets
  - ASQA benchmark dataset 사용
    - AmbigNQ에서 확인된 6,316개의 모호한 질문(AQ)를 기반으로 구축된 long-form QA dataset
    - Train, development, test를 각각 4,353개, 948개, 1,015개로 나누어 사용

- Evaluation metrics
  1. Disambig-F1 (D-F1)
     - Prediction의 사실적 정확도를 측정
       - SQuAD_v2로 학습된 RoBERTa를 사용해, 생성된 long-form answer에서 DQ에 대한 짧은 답변을 찾음.
       - 탐지된 답변의 F1을 계산해 long-form answer에 정확한 정보가 포함되어 있는지 확인

  2. ROUGE-L
     - Reference의 long-form answer와 prediction 간의 어휘 중복도를 측정

  3. DR (Disambiguation-ROUGE)
     - Disambig-F1과 ROUGE-L 의 기하 평균으로 전반적인 성능을 평가

  - DQ에 대한 답변의 최대 F1 정확도를 측정하는 Answer-F1을 추가로 평가

Experiments

1. ToC가 fully-supervised baseline과 few-shot prompting baseline의 성능을 능가

Model	D-F1	R-L	DR
Fully-supervised			
T5-Large Closed-Book	7.4	33.5	15.7
T5-Large w/ JPR	26.4	43.0	33.7
PaLM w/ Soft Prompt Tuning*	27.8	37.4	32.1
Few-shot Prompting (5-shot)			
PaLM*	25.3	34.5	29.6
GPT-3*	25.0	31.8	28.2
Tree of Clarifications (ToC; Ours)			
GPT-3 + RAC	31.1	39.6	35.1
GPT-3 + RAC + TS	32.4	**40.0**	36.0
GPT-3 + RAC + TS w/ Pruning	**33.7**	39.7	**36.6**

* from Amplayo et al. (2023)

- Baseline 중에서는 fully-supervised 가 few-shot prompting 보다 높은 성능을 달성
- RAC가 있는 LLM은 D-F1, DR 점수에서 모든 baseline을 능가
- 트리 구조(TS)을 사용할 경우, D-F1, DR 점수 향상
- Pruning 을 추가하면 모든 척도에서 가장 좋은 성능 달성

Experiments

2. Retrieval system을 통합하는 것이 정확하고 다양한 disambiguations에 크게 기여함

Model	D-F1	R-L	DR
GPT-3 (Baseline)	24.2	36.0	29.5
GPT-3 w/ RAC	**31.1**	**39.6**	**35.1**
– Disambiguations	30.5	37.3	33.7
– Bing Search Engine	28.5	37.4	32.7
– Retrieval Systems	25.6	35.1	30.0

- RAC에서 제안된 각 구성 요소의 기여도를 측정하기 위한 ablation study
- Disambiguation을 제거할 경우, ROUGE-L 점수가 크게 감소
  - 완전한 답을 제공하기 위한 중간 단계의 중요성을 보여줌
- Retrieval system을 통합하면 모델의 성능이 크게 향상됨
  - 외부 지식을 활용하는 것이 설명의 사실 정확도를 높이는 데 중요하다는 것을 나타냄

Experiments

3. Pruning method가 트리에서 유용한 disambiguations을 파악하는데 크게 도움됨

Filtration	#(DQs)	Answer-F1
w/o Pruning (None)	12,838	40.9
w Pruning		
+ Deduplication	10,598	40.1
+ Self-Verification	4,239	59.3

- Target DQ에 대한 답변의 F1 정확도를 측정하는 Answer-F1 점수로 평가
- Self-verification을 사용해 남은 유효한 node에 더 정확한 disambiguations이 포함되어 있어, baseline에 비해 훨씬 더 높은 Answer-F1 점수를 달성
- 중복 제거만으로는 정확도가 향상되지 않음
  - 이 논문에서 제안한 self-verification 방법의 효율성을 보여줌

Conclusion

- 모호한 질문에 대한 다양한 해석을 탐색하기 위해, **LLM과 retrieval system을 결합한 Tree of Clarification (ToC) 방법을 제안**
- 제안 방법은 외부 지식과 few-shot prompt를 통해 AQ에 대한 트리를 재귀적으로 구축해 long-form answer를 생성함
- 실험 결과, ToC가 트리 구조 내에서 주어진 AQ에 대한 다양한 설명 경로를 탐색하고, 포괄적인 답변을 생성하도록 LLM을 안내한다는 것을 보여줌

- 한계점
  - 실험이 ASQA benchmark dataset에 대해서만 수행되었음
  - ToC를 사용하면 LLM이 반복적으로 prompt해 다양한 경로를 탐색할 수 있지만, 여러번 prompt 하는데 드는 비용이 큼
  - CoT 방식을 시도했지만 성능을 향상시킬 수 없었음
    - 이는 모호성 제거 process에서 외부 지식이 필요하다는 것을 의미

Thank You

감사합니
다.