PDFTriage: Question Answering over Long, Structured Documents

Announcement Date	@2025년 2월 27일
Conference Name	EMNLP 2024 Industry Track
Keywords	LLM PDF

Abstract

- LLM은 문서의 컨텍스트 길이 제한으로 인해 문서 기반 QA에서 어려움을 겪으며, 기존 연구는 문서를 일반 텍스트로 변환하여 검색하는 방식에 집중해 왔다.
- 그러나 PDF, 웹페이지, 프레젠테이션과 같은 문서는 고유한 구조를 가지며, 이를 단순 텍스트로 변환하면 사용자의 인식과 맞지 않아 QA 시스템이 오류를 일으킬 수 있다.
- 이를 해결하기 위해 우리는 **PDFTriage**를 제안하며, 이는 문서의 구조나 내용을 기반으로 컨텍스트를 검색할 수 있도록 한다.
- 실험 결과, PDFTriage를 적용한 모델이 기존 검색 기반 LLM이 실패하는 여러 유형의 질문에서 우수한 성능을 보였으며, 연구 확장을 위해 관련 데이터셋을 공개할 예정이다.

Introduction

- LLM은 문서의 컨텍스트 길이 제한으로 인해 QA 성능이 저하되며, 기존 방식은 문서를 일반 텍스트로 변환해 검색하는 방법을 사용하지만, 이는 구조적 문서와 맞지 않음.
- 예를 들어, “페이지 5-7의 핵심 내용을 요약해줄 수 있나요?”나 “표 3에서 가장 높은 수익이 발생한 연도는 언제인가요?” 같은 질문은 문서의 구조적 정보를 필요로 하지만, 기존 방식으로는 이를 제대로 처리하지 못함.
- 이를 해결하기 위해 PDFTriage를 제안하며, 문서 구조 정보를 활용해 모델이 특정 페이지나 표를 검색할 수 있도록 함.
- PDFTriage는 문서 구조 메타데이터를 포함한 프롬프트 보강과 `fetch_pages(pages: list[int])` 같은 함수 제공을 통해 구조적 정보 검색을 가능하게 함.

- 실험 결과, PDFTriage를 적용한 모델이 기존 검색 기반 LLM이 실패한 질문 유형에서도 신뢰할 수 있는 답변을 제공함을 확인함.
- 연구 확장을 위해 90개의 문서와 900개의 인간이 작성한 질문을 포함한 데이터셋을 구축하고, 코드 및 프롬프트와 함께 공개할 예정임.

Related Work

2.1 Tool and Retrieval Augmented LLMs

- Tool-augmented LLM은 외부 도구를 활용해 LLM의 성능을 향상시키는 방식으로, ReAct, Self-ask, Toolformer, TALM, Taskmatrix, Gorilla 등의 기법이 개발되었으며, 각 방법은 API 활용, 자체 학습, 반복적인 성능 향상 등 다양한 접근을 사용함.
- API-Bank와 ToolQA 같은 벤치마크가 도구 활용 능력을 평가하기 위해 개발되었으며, LLM이 API를 올바르게 계획하고 실행하는 능력이나 외부 도구를 이용한 QA 수행 능력을 측정함.
- Retrieval-augmented LLM은 검색 엔진이나 외부 데이터베이스를 활용해 필요한 정보를 검색하여 LLM의 추론 능력을 보강하는 방식으로, HyDE는 가상의 문서를 생성해 관련 정보를 검색하고, InteR는 검색 질의를 반복적으로 개선하여 검색 성능을 높이는 방법을 제안함.
- 전반적으로, 도구 및 검색 기반 보강 기법들은 LLM이 보다 정확하고 효과적으로 질문에 답할 수 있도록 하며, 최근 연구들은 이러한 기법을 대규모로 확장하는 방향으로 발전하고 있음.

2.2 Question Answering

- 기존 QA 연구는 주로 GLUE와 같은 추출 기반 QA에 집중하며, SQuAD나 NaturalQuestions 같은 데이터셋은 표나 그림이 포함되지 않은 텍스트 중심 문서를 다룸.
- 문서 기반 QA를 평가하기 위해 다양한 데이터셋이 개발되었으며, DocVQA는 문서 스캔을 활용한 시각적 질문 응답을, DUDE는 스캔 및 디지털 PDF를 활용한 문서 이해 및 평가를 다룸.
- DUDE와 DocVQA는 평균적으로 짧은 답변을 요구하는 반면, QASPER는 연구 논문을 기반으로 정보 탐색 질문과 그에 대한 답변을 포함하며, LaTeX 소스에서 문서를 파싱하여 활용함.
- PDFTriage 데이터셋은 기존 데이터셋의 한계를 확장하여 문서 구조와 내용을 참조하는 질문, 추출적·생성적 답변, 장문 응답 및 재작성까지 포함하는 평가 기준을 제시함.

PDFTriage : Structed Retrieval from Document Metadata

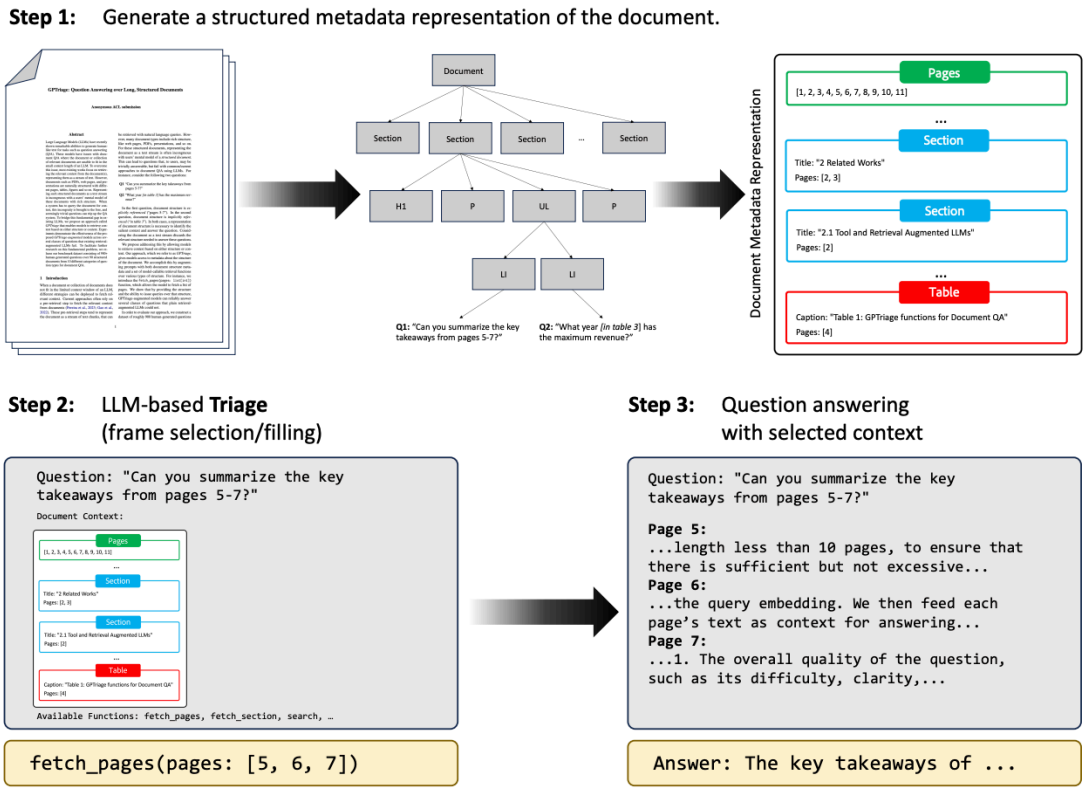

Figure 1: **Overview of the PDFTriage technique**: PDFTriage leverages a PDF's structured metadata to implement a more precise and accurate document question-answering approach. It starts by generating a structured metadata representation of the document, extracting information surrounding section text, figure captions, headers, and tables. Next, given a query, a LLM-based Triage selects the document frame needed for answering the query and retrieves it directly from the selected page, section, figure, or table. Finally, the selected context and inputted query are processed by the LLM before the generated answer is outputted.

PDFTriage 접근 방식은 사용자의 질문에 답변하기 위해 다음 세 단계로 구성됨(그림 1 참조):

1. **문서 메타데이터 생성 (섹션 3.1)**: 문서의 구조적 요소를 추출하고 이를 읽을 수 있는 메타데이터로 변환함.
2. **LLM 기반 분류 (섹션 3.2)**: LLM을 활용하여 문서에서 정확한 내용(페이지, 섹션, 검색된 컨텍스트)을 선택함.
3. **검색된 내용을 이용한 답변 생성 (섹션 3.3)**: 질문과 검색된 컨텍스트를 기반으로 답변을 생성함.

3.1 Document Representation

- 우리는 사용자가 상호작용할 구조화된 문서로 디지털 PDF를 고려하며, Adobe Extract API를 사용하여 PDF를 HTML과 유사한 트리 구조로 변환함.
- 이 API는 섹션 제목, 페이지 정보, 표, 그림 등을 포함한 계층적 트리를 생성하며, 각 요소는 페이지와 위치 같은 메타데이터를 포함함.
- 변환된 구조적 정보는 JSON 형식으로 매핑되어 LLM의 초기 프롬프트로 사용되며, 콘텐츠는 마크다운 형식으로 변환됨.

3.2 LLM Querying of Document

- PDFTriage는 `fetch_pages`, `fetch_sections`, `fetch_table`, `fetch_figure`, `retrieve`의 다섯 가지 기능을 사용하여 PDF 문서의 구조적 데이터를 정확하게 검색함.
- 각 기능은 헤더, 하위 헤더, 그림, 표, 섹션 단락 등의 정보를 가져오며, 질문별로 별도의 쿼리를 수행하여 최종 답변을 생성함.
- OpenAI의 함수 호출 API를 활용하여 각 기능을 개별 채팅 턴에서 호출하며, ReAct 또는 Toolformer 방식으로 프롬프트를 구성하는 것도 가능함.

Function	Description
`fetch_pages`	Get the text contained in the pages listed.
`fetch_sections`	Get the text contained in the section listed.
`fetch_figure`	Get the text contained in the figure caption listed.
`fetch_table`	Get the text contained in the table caption listed.
`retrieve`	Issue a natural language query over the document, and fetch relevant chunks.

Table 2: PDFTriage Functions for Document QA.

3.3 Question Answering

- PDFTriage는 GPT-3.5의 시스템 프롬프트를 활용하여 문서의 메타데이터를 포함한 질문 응답을 수행함.
- 사용자의 질문을 입력하면, PDFTriage는 미리 정의된 기능을 이용해 문서에서 필요한 정보를 검색함.
- 각 턴에서 하나의 기능을 사용해 정보를 수집한 후 최종적으로 모델이 답변을 생성하며, 실험에는 gpt-3.5-turbo-0613 모델이 사용됨.

Dataset Construction

- PDFTriage의 성능을 평가하기 위해 문서 내 텍스트, 표, 그림을 활용한 다양한 질문-응답 과제를 구성함.
- Mechanical Turk를 활용하여 실제 업무 환경에서 발생할 수 있는 문서 기반 질문을 수집함.
- Common Crawl에서 1000개의 문서를 샘플링하고, 읽기 난이도를 고려해 100개의 전문 문서를 최종 선정함.
- 다양한 질문 유형을 반영하기 위해 10개 카테고리의 질문 유형을 정의하고, 각 유형에서 균형 잡힌 샘플을 수집함.
- 이 데이터셋은 다단계 추론을 포함한 다양한 문서 기반 QA 접근 방식을 평가할 수 있도록 설계됨.

1. **그림 관련 질문 (6.5%)**: 문서 내 그림에 대한 질문을 묻는 경우.
2. **텍스트 관련 질문 (26.2%)**: 문서의 내용과 관련된 질문을 묻는 경우.
3. **표 추론 (7.4%)**: 문서 내 표에 대한 질문을 묻는 경우.
4. **구조 관련 질문 (3.7%)**: 문서의 구조에 대한 질문을 묻는 경우.
5. **요약 요청 (16.4%)**: 문서의 일부 또는 전체 내용을 요약해달라는 요청.
6. **추출 요청 (21.2%)**: 문서에서 특정 콘텐츠를 추출해달라는 요청.
7. **재작성 요청 (5.2%)**: 문서 내 일부 텍스트를 다시 작성해달라는 요청.
8. **외부 정보 질문 (8.6%)**: 문서만으로는 답할 수 없는 질문을 묻는 경우.
9. **다중 페이지 작업 (1.1%)**: 문서의 여러 부분을 참조해야 답할 수 있는 질문.
10. **분류 요청 (3.7%)**: 문서의 유형을 묻는 질문.

- 이 데이터셋은 82개의 문서에서 총 908개의 질문을 포함하며, 문서당 평균 4,257개의 토큰으로 구성되어 있으며, 문서의 단어 수 분포 및 질문 유형 설명은 부록에 제공됨.

Experiments

5.1 PDFTriage

- PDFTriage를 사용하여 선택된 PDF 문서 데이터셋에서 다양한 질문에 답변하는 실험을 진행함.

- 이 접근 방식은 PDF 구조와 GPT-3.5의 기능을 활용하여 기존보다 더 정밀하고 정확한 답변을 제공함.

5.2 Retrieval Baselines

- 페이지 검색 방식은 문서의 각 페이지를 `text-embedding-ada-002` 임베딩으로 색인화하고, 코사인 유사도를 이용해 질문과 가장 유사한 페이지를 검색하여 답변을 생성함.
- 청크 검색 방식은 문서를 100단어 단위로 나눈 후, 유사한 청크를 검색하여 문맥으로 제공하고 질문에 답변함.
- 두 가지 검색 방식 모두 GPT-3.5에 동일한 프롬프트를 사용하여, 검색된 페이지 또는 청크를 기반으로 질문에 대한 답변을 생성함.

5.3 Human Evaluation

- PDFTriage와 검색 기반 모델의 성능 차이를 평가하기 위해 Upwork에서 영어 능숙자 12명을 고용하여 인간 평가 연구를 진행함.
- 평가자는 질문의 난이도와 명확성, 질문 유형 분류, 생성된 답변의 순위 및 품질(정확성, 정보성, 가독성, 명확성) 등을 분석함.
- 전체 평가 질문과 평가자의 인구통계학적 정보는 부록 A에 제공되며, 개요에서는 샘플 질문을 사용하여 설명함.

Results and Analysis

다양한 난이도와 유형의 질문을 포함한 데이터셋을 구축하여, PDFTriage가 문서 기반 QA에서 여러 추론 방식을 얼마나 효과적으로 수행하는지 평가할 수 있도록 함.

6.1 PDFTriage yields better answers than retrieval-based approaches. (PDFTriage는 검색 기반 접근 방식보다 더 나은 답변을 제공한다.)

- 평가 연구 결과, PDFTriage가 페이지 검색(Page Retrieval)과 청크 검색(Chunk Retrieval)보다 높은 평가를 받았으며, 전체 질문 중 50.7%에서 가장 선호되는 답변을 제공하고, 모든 질문 유형에서 기존 방식보다 우수한 성능을 보였음.

6.2 PDFTriage improves answer quality, accuracy, readability, and informativeness (PDFTriage는 답변의 품질, 정확성, 가독성, 그리고 정보성을 향상시킨다.)

	PDFTriage	*Page Retrieval*	*Chunk Retrieval*
Readability	**4.2**	4.1	4.1
Informativeness	**3.9**	3.7	3.4
Clarity	2.0	2.1	**2.3**
Accuracy	**3.8**	3.6	3.4
Overall Quality	**3.9**	3.8	3.6

Table 3: Answer Quality Scoring

- PDFTriage는 페이지 검색과 청크 검색보다 명확성을 제외한 모든 답변 품질 기준에서 높은 점수를 받았으며, 특히 전체 품질과 답변 정확성에서 가장 우수한 평가를 받음.
- 평가 결과, PDFTriage는 요약, 표 추론, 정보 추출, 그림 관련 질문처럼 다단계 추론이 필요한 질문에서 강한 성능을 보였으며, 일반적인 텍스트 질문과 문서 분류 작업에서는 기존 방법과 유사한 성능을 나타냄.

6.3 PDFTriage requires fewer retrieved tokens to produce better answers (PDFTriage는 더 나은 답변을 생성하기 위해 적은 수의 검색된 토큰을 필요로 한다.)

- PDFTriage는 평균 1,568개의 토큰을 활용하며, 기존 페이지 검색(3,611 토큰)과 청크 검색(3,934 토큰)보다 더 많은 정보를 비연속적인 문서 섹션에서 검색함.
- 페이지 검색과 청크 검색은 제한된 섹션만을 사용해 답변을 생성하기 때문에 "전체 품질"과 "정확성" 평가에서 낮은 점수를 기록함.
- PDFTriage는 다단계 검색을 통해 필요한 정보를 동적으로 추가하여 문서 QA 작업에서 보다 효과적인 답변을 제공함.

6.4 PDFTriage performs consistently across document lengths (PDFTriage는 문서 길이에 관계없이 일관된 성능을 발휘한다.)

- PDFTriage의 성능과 문서 길이 사이의 피어슨 상관 계수는 -0.015로, 문서 길이가 성능에 거의 영향을 미치지 않음을 확인함.
- PDFTriage는 특정 섹션만 검색하여 과도한 컨텍스트 처리 없이 표 추론, 다중 페이지 검색, 그림 및 구조적 질문을 효과적으로 수행함.

- 이를 통해 GPT-3 및 LLM이 더 적은 연산 자원으로 문서 QA 작업을 수행할 수 있어 비용 효율성이 향상됨.

Future Work & Conclusions

이번 연구에서는 문서 중심 작업에 특화된 새로운 질문 응답(QA) 기술인 **PDFTriage**를 제안한다. 우리는 페이지 검색(Page Retrieval)과 청크 검색(Chunk Retrieval) 등 기존 질문 응답 기법과 비교하여 PDFTriage의 강점을 분석하였다. 실험 결과, PDFTriage는 기존 접근 방식보다 우수한 성능을 보였으며, 다양한 문서 길이와 검색 컨텍스트에서도 효과적으로 작동함을 확인하였다.

우리는 향후 연구 방향으로 다음과 같은 두 가지를 고려하고 있다:

1. **멀티모달 접근 방식 개발** – 표와 그림 정보를 통합하여 GPT-4 기반 문서 질문 응답 시스템을 확장하는 연구.
2. **질문 유형을 반영한 최적화** – 질문 유형을 PDFTriage 접근 방식에 적용하여 시스템의 효율성과 성능을 더욱 향상하는 연구.