Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

저자	Patrick Lewis, Ethan Perez, et al. (Facebook AI Research)
저널명	Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks
출판연도	2025
주요내용 요약	이 논문은 대규모 사전 학습 언어 모델(Large pre-trained language models)의 지식 접근 및 조작 능력 한계를 극복하고, 지식 집약적인 자연어 처리(NLP) 태스크에서 최첨단 성능을 달성하기 위해 Retrieval-Augmented Generation (RAG) 모델을 제안하였다. 언어 생성 태스크에 대해, 우리는 RAG 모델이 최첨단 모수적 전용 seq2seq 기준 모델보다 더 구체적이고, 다양하며, 사실적인 언어를 생성한다는 것을 발견했다.
PDF/Citation	2005.11401v4.pdf
상태	읽는중
연구 형태	Methodology
키워드	NLP
BibTex	@misc{lewis2021retrievalaugmentedgenerationknowledgeintensivenlp, title={Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks}, author={Patrick Lewis and Ethan Perez and Aleksandra Piktus and Fabio Petroni and Vladimir Karpukhin and Naman Goyal and Heinrich Küttler and Mike Lewis and Wen-tau Yih and Tim Rocktäschel and Sebastian Riedel and Douwe Kiela}, year={2021}, eprint={2005.11401}, archivePrefix={arXiv}, primaryClass={cs.CL}, url={https://arxiv.org/abs/2005.11401}, }

1. **요약**

- 논문의 핵심 기여도와 결과

이 논문은 대규모 사전 학습 언어 모델(Large pre-trained language models)의 지식 접근 및 조작 능력 한계를 극복하고, 지식 집약적인 자연어 처리(NLP) 태스크에서 최첨단 성능을 달성하기 위해 Retrieval-Augmented Generation (RAG) 모델을 제안하였다. 언어 생성 태스크에 대해, 우리는 RAG 모델이 최첨단 모수적 전용 seq2seq 기준 모델보다 더 구체적이고, 다양하며, 사실적인 언어를 생성한다는 것을 발견했다.

2. **배경 및 동기 / 문제정의**

- 기존 연구들의 한계점
- 이 연구가 왜 시작되었는지에 대한 이유

pre-trained language models은 방대한 양의 데이터를 통해 상당한 깊이 있는 지식을 학습하였으며, parametric memory에 지식을 저장한다. 따라서 model이 외부 메모리 접근하지 않고 스스로 지식 기반 역할을 하기 때문에, 여러 task에서 뛰어난 성능을 보였다.

그러나 “Knowledge-Intensive NLP Tasks”에서는 지식 접근 및 조작 능력, 의사 결정의 출처(provenance) 제공, 세계 지식 업데이트 등에서 한계가 있었다.

a. **메모리 확장/수정의 어려움** : 새로운 정보가 생기거나 기존 정보가 업데이트될 때 모델의 지식을 쉽게 확장하거나 수정하기 어렵다.

b. **예측에 대한 통찰력 부족** : 모델이 왜 특정 답변을 생성했는지에 대한 명확한 근거(provenance)를 제공하기 어렵습니다. 이는 "블랙박스" 문제와 관련이 있습니다.

c. **Hallucinations 발생** : 모델이 사실과 다른, 그럴듯하지만 틀린 정보를 생성할 수 있기 때문에 Knowledge-Intensive Tasks에서는 치명적일 수 있다.

이러한 문제점을 해결하기 위해 REALM과 ORQA 등 추출형 질문 답변(Extractive QA)에 집중하여parametric memory와 non-parametric memory를 결합한 하이브리드 모델이 등장했다. 기존 Extractive QA에 집중한 하이브리드 모델은 모델이 외부 지식을 활용하여 답변의 사실성(factuality)을 확보하고, 답변이 검색된 문서에 근거한다는 출처를 제공할 수 있다는 장점을 가진다. 그러나 open-domain extractive question answering task에 관련해서만 주목해왔었다.

이 논문은 이러한 하이브리드 접근 방식을 NLP의 핵심인 seq2seq 모델에 적용하여, RAG라는 범용적인 미세 조정 방식을 제시하였다.

2. **제안 방법론**

- 모델 아키텍쳐
- 핵심 알고리즘

3-1. Model Architecture

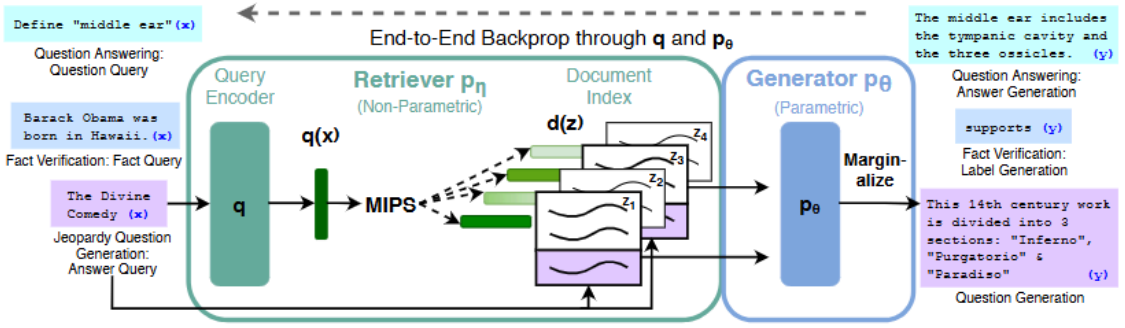

Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder + Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x, we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i. For final prediction y, we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

RAG model은 input sequence x를 받게 되면, Query Encoder가 이를 embedding으로 변환한다. embedding을 사용해서 Retriever는 방대한 Document Index에서 가장 관련성 높은 문서 z들을 찾는다. Generator는 원본 질문과 검색된 문서들을 기반으로 Marginalize을 거쳐 최종 답변 y를 생성한다.

a. **Input (x)**

b. **Query Encoder (q)** : query embedding $q(x)$로 변환한다. (쿼리의 벡터표현)

c. **Retriever (p_η, non-parametric)** : Retriever는 입력 x가 주어졌을 때 문서 z가 검색될 확률 $p_\eta(z|x)$를 정의한다. ⇒ 모델이 어떤 외부 지식을 활용할지 결정하는 부분

  - Document Index : 대규모 non-parametric memory에서 가져온 문서 z들의 밀집 벡터 표현 $d(z)$를 저장한다
  - MIPS (Maximum Inner Product Search): Query Encoder에서 생성된 $q(x)$를 사용하여, Document Index에 있는 모든 문서 임베딩 $d(x)$와 내적(inner product)을 계산하고 가장 유사한(내적 값이 높은) 상위 K개의 문서 z를 검색한다.

d. **Generator (p_θ, parametric memory)** : seq2seq 모델을 의미하며, Retriever가 검색한 상위 K개의 문서 z가 input(x)를 조건으로 하여 target sequence y가 생성될 확률 $p_\theta(y \mid x, z)$를 구한다. Generator는 이러한 정보를 활용하여 parametric memory에 내재된 지식을 바탕으로 y를 구성한다.

e. **Marginalization**

- 단순히 하나의 문서에만 의존하는 것이 아니라, 검색된 여러 문서의 정보를 통합하여 최종 예측을 수행한다.
- 이 논문에서는 RAG-Sequence 모델과 RAG-Token 모델이라는 두 가지 방식으로 주변화를 수행하였다.
  - RAG-Sequence는 전체 시퀀스를 생성하는 데 동일한 문서를 사용

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x,z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_i^N p_\theta(y_i|x,z,y_{1:i-1})$$

  - RAG-Token은 각 토큰을 예측하는 데 다른 문서를 사용

$$p_{\text{RAG-Token}}(y|x) \approx \prod_i^N \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x,z,y_{1:i-1})$$

3-2. RAG Model의 종류

a. RAG-Sequence Model : 전체 시퀀스 y를 생성하는데 동일한 단일 검색 문서 z를 사용한다.

b. RAG-Token Model : 각 target token을 예측할 때마다 Retriever가 검색한 상위 K개의 문서 z를 검색한 후, generator를 거쳐서 output을 도출한다.

만약 RAG-Sequence Model의 시퀀스가 1이면, RAG-Sequence Model = RAG-Token Model 이다.

3-3. Retriever(검색기) : DPR

Retriever는 DPR(Dense Passage Retrieval) 주어진 입력 쿼리 x와 관련성이 높은 문서를 찾아낸다.

$p_\eta(z|x)$ ⇒ x 가 주어졌을 때 텍스트 문서 z의 확률 분포를 계산하여, 어떤 질문이나 문장이 있을 때 그 내용과 가장 관련이 깊은 문서를 찾을 수 있다.

$$p_\eta(z|x) \propto \exp\left(\mathbf{d}(z)^\top \mathbf{q}(x)\right) \qquad \mathbf{d}(z) = \text{BERT}_d(z), \;\; \mathbf{q}(x) = \text{BERT}_q(x)$$

DPR은 bi-encoder 구조로, 문서와 쿼리를 각각 별도의 BERT based encoder를 사용하여 독립적으로 인코딩하는 방식이다. d의 transpose와 q를 내적하면 x 와 z의 유사도를 구할 수 있고, 이것을 exponential 하면 확률 분포로 정규화 할 수 있다. ⇒ **non-parametric memory**

3. **실험 및 결과**

- 어떤 데이터를 사용했는지, 성능은 얼마나 좋아졌는지
- 표와 수치를 사용하여 비교

a. Dataset : single non-parametric momory로 2018년 12월자 Wikipedia dump를 사용함

b. 문서 검색 : 각 쿼리에 대해 상위 k개 문서 k ∈ {5, 10} 로 검색하고 테스트 시에는 개발 데이터를 활용해 최적 k를 설정했다.

c. Tasks

- 오픈 도메인 질의응답 (Open-domain QA) : 질문(x)과 답변(y)을 텍스트 쌍으로 처리하여 답변의 음의 한계 로그-가능도를 최소화하는 방식
  - 데이터셋: Natural Questions (NQ), TriviaQA (TQA), WebQuestions (WQ), CuratedTrec (CT)
  - 평가: Exact Match (EM)
  - 결과 : Closed book 방식의 generation flexibility와 Open book 방식의 retrieval-based performance의 장점을 모두 가지고 있기 때문에 가장 우수한 성능을 보임

Table 1: Open-Domain QA Test Scores. For TQA, left column uses the standard test set for Open-Domain QA, right column uses the TQA-Wiki test set. See Appendix D for further details.

	Model	NQ	TQA	WQ	CT
Closed Book	T5-11B [52]	34.5	- /50.1	37.4	-
	T5-11B+SSM[52]	36.6	- /60.5	44.7	-
Open Book	REALM [20]	40.4	- / -	40.7	46.8
	DPR [26]	41.5	**57.9**/ -	41.1	50.6
	RAG-Token	44.1	55.2/66.1	**45.5**	50.0
	RAG-Seq.	**44.5**	56.8/**68.0**	45.2	**52.2**

- 추상적 질의응답 (Abstractive QA) : 자연어 생성(NLG) 능력을 평가하며, 제공된 참조 문단 없이 질문과 답변만을 사용하여 오픈 도메인 방식으로 수행
  - 데이터셋: MSMARCO NLG v2.1
  - 결과 : RAG 모델은 BART보다 2.6 Bleu 포인트 및 2.6 Rouge-L 포인트 더 나은 성능을 보였다. 정성적으로는 BART보다 hallucination이 적고 사실적으로 정확한 답변을 생성할 수 있음을 입증하였다.

Table 2: Generation and classification Test Scores. MS-MARCO SotA is [4], FEVER-3 is [68] and FEVER-2 is [57] *Uses gold context/evidence. Best model without gold access underlined.

Model	Jeopardy B-1	Jeopardy QB-1	MSMARCO R-L	MSMARCO B-1	FVR3 Label Acc.	FVR2 Label Acc.
SotA	-	-	**49.8***	**49.9***	**76.8**	**92.2***
BART	15.1	19.7	38.2	41.6	64.0	81.1
RAG-Tok.	**17.3**	**22.2**	40.1	41.5	72.5	89.5
RAG-Seq.	14.7	21.4	40.8	44.2		

Table 3: Examples from generation tasks. RAG models generate more specific and factually accurate responses. '?' indicates factually incorrect responses, * indicates partially correct responses.

Task	Input	Model	Generation
MS-MARCO	define middle ear	BART	?The middle ear is the part of the ear between the middle ear and the nose.
		RAG-T	The middle ear is the portion of the ear internal to the eardrum.
		RAG-S	The middle ear includes the tympanic cavity and the three ossicles.
	what currency needed in scotland	BART	The currency needed in Scotland is Pound sterling.
		RAG-T	Pound is the currency needed in Scotland.
		RAG-S	The currency needed in Scotland is the pound sterling.
Jeopardy Question Gener-ation	Washington	BART	?This state has the largest number of counties in the U.S.
		RAG-T	It's the only U.S. state named for a U.S. president
		RAG-S	It's the state where you'll find Mount Rainier National Park
	The Divine Comedy	BART	*This epic poem by Dante is divided into 3 parts: the Inferno, the Purgatorio & the Purgatorio
		RAG-T	Dante's "Inferno" is the first part of this epic poem
		RAG-S	This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso"

- Jeopardy 질문 생성 (Jeopardy Question Generation) : 특정 답변 개체에 대한 Jeopardy 스타일 질문 생성
  - 데이터셋: SearchQA
  - 평가: Q-BLEU-1 및 인간 평가(사실성, 구체성)
    - Table 2 : RAG-Token 모델이 Jeopardy 질문 생성에서 RAG-Sequence 모델보다 더 나은 성능을 보였으며, 두 RAG 모델 모두 BART 기준선보다 Q-BLEU-1 점수가 높음
    - 인간 평가 결과 또한 BART보다 훨씬 더 사실적이고 구체적인 답변을 생성한다고 판단했다.
- 사실 검증 (Fact Verification) : 자연어 주장이 위키피디아에 의해 '지지', '반박', 또는 '정보 부족'인지 분류. 검색된 증거에 대한 직접적인 감독 없이 훈련
  - 데이터셋: FEVER (3방향 및 2방향 분류)
  - 평가: 레이블 정확도
    - 3가지 분류 태스크에서 최신 기술 모델 대비 4.3% 이내의 경쟁력 있는 점수를 달성 ⇒ RAG는 검색 감독 불필요함
    - 2가지 분류 태스크에서 주장만 제공받고 자체적으로 증거를 검색함에도 불구하고, RoBERTa 기반의 최신 모델과 2.7% 이내의 정확도를 보임

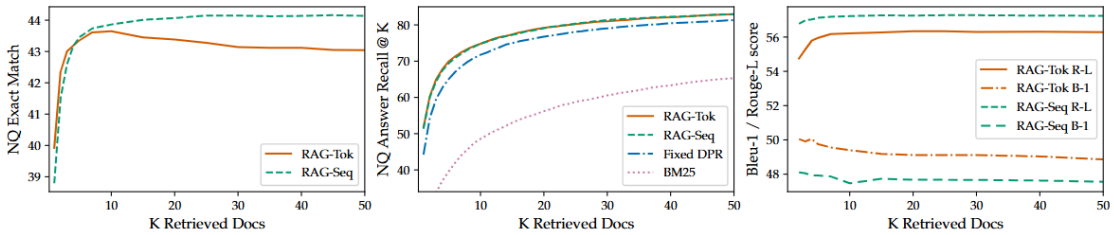

Figure 3: Left: NQ performance as more documents are retrieved. Center: Retrieval recall performance in NQ. Right: MS-MARCO Bleu-1 and Rouge-L as more documents are retrieved.