AI는 모르는 정보를 어떻게 찾아서 답할까?

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

"지식 집약적 자연어 처리(NLP) 작업을 위한 검색 증강 생성"

30214 진재화

이 논문은 무엇을 연구했을까?



Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis[1,2], Ethan Perez[3],

Aleksandra Piktus[1], Fabio Petroni[1], Vladimir Karpukhin[1], Naman Goyal[1], Heinrich Küttler[1],

Mike Lewis[1], Wen-tau Yih[1], Tim Rocktäschel[1,2], Sebastian Riedel[1,2], Douwe Kiela[1]

[1]Facebook AI Research; [2]University College London; [3]New York University;
plewis@fb.com

Abstract

Large pre-trained language models have been shown to store factual knowledge in their parameters, and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, their performance lags behind task-specific architectures. Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-trained models with a differentiable access mechanism to explicit non-parametric memory have so far been only investigated for extractive downstream tasks. We explore a general-purpose fine-tuning recipe for retrieval-augmented generation (RAG) — models which combine pre-trained parametric and non-parametric memory for language generation. We introduce RAG models where the parametric memory is a pre-trained seq2seq model and the non-parametric memory is a dense vector index of Wikipedia, accessed with a pre-trained neural retriever. We compare two RAG formulations, one which conditions on the same retrieved passages across the whole generated sequence, and another which can use different passages per token. We fine-tune and evaluate our models on a wide range of knowledge-intensive NLP tasks and set the state of the art on three open domain QA tasks, outperforming parametric seq2seq models and task-specific retrieve-and-extract architectures. For language generation tasks, we find that RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline.

1 Introduction

Pre-trained neural language models have been shown to learn a substantial amount of in-depth knowledge from data [47]. They can do so without any access to an external memory, as a parameterized implicit knowledge base [51, 52]. While this development is exciting, such models do have downsides: They cannot easily expand or revise their memory, can't straightforwardly provide insight into their predictions, and may produce "hallucinations" [38]. Hybrid models that combine parametric memory with non-parametric (i.e., retrieval-based) memories [20, 26, 48] can address some of these issues because knowledge can be directly revised and expanded, and accessed knowledge can be inspected and interpreted. REALM [20] and ORQA [31], two recently introduced models that combine masked language models [8] with a differentiable retriever, have shown promising results,

arXiv:2005.11401v4 [cs.CL] 12 Apr 2021

Retrieval-Augmented Generation

Patrick Lewis 외 11명

NeurIPS 2020

외부 문서 검색과 언어 생성 모델을 결합해
지식이 필요한 질문에 더 잘 답하는 방법을 연구했다.

출처 | Lewis et al., NeurIPS 2020 · 원본 논문 1쪽

AI가 모르는 사실을 물어보면?

질문

“우리 학교의 오늘 점심 메뉴는 뭐야 ?”

답변

“돈가스와 미역국입니다.”

그럴듯한 말이지만, 근거 없는 답변일 수도 있다.

왜 새로운 방법이 필요했을까?

01	정확성	학습한 지식을 필요한 순간에 정확하게 활용하기 어려움
02	출처	답변을 뒷받침하는 근거를 쉽게 보여주기 어려움
03	업데이트	새로운 정보를 모델 내부에 바로 반영하기 어려움

중요: RAG 이전에도 외부 검색을 활용한 AI 연구는 존재했다. (예: REALM, ORQA)

찾아보고, 읽고, 대답한다.

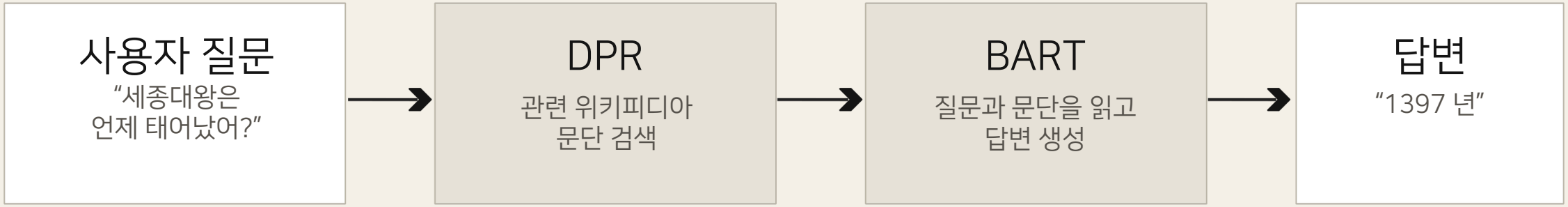
RAG

Retrieval-Augmented Generation

검색으로 얻은 외부 자료를 언어 모델에 제공해
답변을 생성하는 방식

이 논문은 위키피디아 문서를 검색 대상으로 사용했다.

RAG는 어떻게 동작할까?



검색 결과는 답변을 그대로 복사하기 위한 것이 아니라,
답변을 만들 때 참고하는 추가 정보가 된다.

논문 Methods: DPR 검색기 + BART 생성기 · 설명을 위한 예시

찾은 문서를 어떻게 활용할까?

RAG-Sequence

답변 전체를 만들 때
한 문서에 의존하는 생성 경로를
기준으로 결과들을 결합

RAG-Token

다음 토큰을 만들 때마다
여러 검색 문서의 정보를
다르게 반영할 수 있음

두 방식 모두 여러 문서를 검색한다.

논문 2.1 Models · 설명을 위한 단순화. 둘 다 검색기/생성기를 사용한다.

정말 효과가 있었을까?

연구진은RAG를기존모델과 비교했다.

질문 답변
위키피디아 지식이 필요한 질문에 정확하게 답하는가?

문장 생성
더 구체적이고 사실적인 문장을 만드는가?

사실 검증
주어진 주장에 근거가 있는지 판별하는가?

실험은 발표자가 새로 수행한 것이 아니라 논문 저자들이 수행한 것이다.

실험에서는 어떤 차이가 났을까?

Natural Questions · Exact Match



RAG는 같은 평가에서 기존 검색·추출 방식보다
더 높은 점수를 기록했다.

+3.0 점

원본 논문 Table 1 · NQ 테스트 · EM은 정답 문자열이 일치하는 비율을 나타내는 지표

그렇다면 RAG는 항상 정확할까?

아니다.

검색오류	잘못된 문서를 찾으면 답변도 틀릴 수 있다.
자료의 범위	검색 대상에 없는 정보는 찾기 어렵다.
생성 오류	올바른 문서를 찾아도 답변을 잘못 만들 수 있다.

논문의 목적은 문제를 완전히 없애는 것이 아니라, 검색 결합으로 성능을 개선하는 것이다.

AI가 모든 것을
기억할 필요는 없다.

학습한 지식 + 필요한 정보의 검색 + 답변 생성

RAG는 외부 지식을 활용해 더 정확하고 구체적인
언어 생성을 가능하게 하는 방향을 보여줬다.

Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” NeurIPS 2020

감사합니다.

원문 Lewis et al., NeurIPS 2020

한국어 해설 Better Jeong, 「RAG 논문 번역」

arxiv.org/abs/2005.11401