Zero-shot 기반 다중 문서 그라운딩된 대화 시스템

박준범°[1], 홍범석[2], 최원석[2], 한영섭[2], 전병기[2], 나승훈[1]
[1]전북대학교, [2]LG 유플러스
pjb2516@naver.com, {bshong, wseokchoi, yshan042, bkjeon}@lguplus.co.kr, nash@jbnu.ac.kr

Zero-shot Dialogue System Grounded in Multiple Documents

Jun-Bum Park°[1], Beomseok Hong[2], Wonseok Choi[2], Youngsub Han[2], Byoung-Ki Jeon[2], Seung-Hoon Na[1]
[1]Jeonbuk National University, [2]LG Uplus

요 약

본 논문에서는 다중 문서 기반의 대화 시스템을 통한 효율적인 정보 검색과 응답 생성에 중점을 둡니다. 대규모 데이터 집합에서 정확한 문서를 선택하는 데 필요한 검색의 중요성을 강조하며, 현재 검색 방법의 한계와 문제점을 지적합니다. 또한 더 자연스러운 답변을 생성하기 위해 대규모 언어 모델을 사용하게 되면서 fine-tuning 시에 발생하는 제약과 낭비를 모델의 제로샷 생성 능력을 활용하여 개선하려는 방안을 제안하며, 모델의 크기와 자원의 효율성에 대한 고려사항을 논의합니다. 우리의 접근 방식은 대규모 언어 모델을 프롬프트와 함께 다중 문서로 학습 없이 정보를 검색하고 응답을 생성하는 방향으로 접근하여 대화 시스템의 효율성과 유용성을 향상시킬 수 있음을 제시합니다.

주제어: Document Grounded Dialogue System, Large Language Model, LLM
Zero-shot, Retrieval Augmented Language Model

1. 서론

다중 문서에 기반한 대화 시스템은 사용자의 질문에 대해 사실에 기반한 응답을 생성하는 목적을 갖는다. 이를 위해서 사용자의 질문이 입력 되었을 때 대화 기록과 질문을 바탕으로 코퍼스에 포함된 여러 개의 문서 중 질문과 가장 관련 있는 문서를 찾는 검색 과정이 대화 시스템에 포함되어 있다. 외부 코퍼스에 대한 대규모 검색의 경우 수십만에서 수십억 개의 항목을 가진 거대한 컬렉션에서 주어진 질문에 대한 관련 문서를 가져오게 되는데 이 과정에서 잘못된 문서가 검색된 경우, 말은 그럴듯하지만 부정확한 오류에 기반한 응답을 생성하거나 개인 데이터가 코퍼스에 포함된 경우 사생활 침해의 영역으로 문제가 발생할 수 있다. 이러한 이유로 정밀한 검색의 중요성은 증가하고 있다.

대규모 검색 과정은 크게 두 단계로 나뉘어 1단계 검색은 수백만에서 수십억 개의 항목을 가진 거대한 컬렉션에서 주어진 텍스트 질문에 대한 관련 문서를 가져오는 것이다. 이 때 질문과 문서 사이의 상관 관계는 sparse vector 방식을 사용하는지 dense embedding을 사용하는지에 따라 BM25[1], TF-IDF[2] 혹은 DPR[3]로 대표하여 1단계 검색기로 꼽을 수 있다. 그리고 Reranking으로 대표되는 2단계 검색에서 재순위화 되는 과정을 거치기도 하지만, 수 많은 데이터와 작업에서 검색의 결과가 완벽했던 경우는 없다. 따라서 본 논문에서는 검색 모듈의 불완전함을 개선하기 위해 대규모 언어 모델을 활용한 검색 단계를 한번 더 거쳐 답변을 생성하는 방식을 제안한다.

대규모 언어 모델을 활용한 작업에 대해 Billion 단위로 나타내는 큰 모델들은 많은 수의 파라미터 탓에 업데이트에 제약이 발생하기도 한다. 이 문제를 해결하기 위한 방법 중 하나로, 대규모 언어 모델의 제로샷 생성 능력을 활용하는 것이 제안되고 있다. 제로샷 학습은 사전 학습된 모델이 특정 작업에 대한 명시적인 fine-tuning 없이 그 작업을 수행하는 능력을 말한다.

OpenAI에서 개발한 GPT-3와 같은 대규모 언어 모델의 성능을 세부적으로 관찰한 결과[4], GPT-3의 다양한 크기의 모델들이 제로샷, 퓨샷, 그리고 많은 샷을 가리지 않고 높은 작업 수행 능력을 보였으며, 특히 큰 모델에서는 특별한 튜닝 없이도 다양한 태스크에서 높은 성능을 보임과 동시에 예제를 몇 개 제공받는 퓨샷 학습을 통해 해당 태스크를 수행하는 데 필요한 지식을 빠르게 학습할 수 있음을 확인했다. 하지만 대규모 언어 모델 또한 때때로 사실에 기반하지 않은 잘못된 정보나 엉뚱한 답변을 제공할 수 있고 논리적 또는 수학적 태스크에서의 일관성은 부족함을 한계점으로 제시하였다.

만약 대규모 언어 모델이 다중 문서를 활용하여 제로샷 생성 능력을 갖추게 된다면, 이는 문서 그라운딩 기반의 대화 시스템의 효율성과 유용성을 크게 향상시킬 수 있을 것이다.

본 논문에서는 대규모 언어 모델의 제로샷 생성 능력과 다중 문서 활용의 가능성에 대해 탐구한다. 또한, 다중 문서 그라운딩된 대화 시스템의 성능을 개선하기 위한 방법론과 이를 통해 얻을 수 있는 혜택에 대해 논의한다.

- DPR-Reranking으로 대표되는 2단계의 검색 과정은 대화 기록과 질문으로부터 답변에 적절한 문서를 찾는 작업 성능에 대해 여전히 한계를 보이고 있다.
- 대규모 언어 모델을 활용하여 작업을 진행할 경우 모델의

규모에 따라서 학습에 제약이 일어나거나 낭비되는 자원이 발생할 수 있다.

- 위에서 제시한 한계점을 개선하고자 본 논문에서는 여러개의 문서를 프롬프트로 대규모 언어 모델에 입력하여 모델이 다시 한번 문맥과 어울리는 문서를 선택할 수 있도록 하며 학습 없이 답변을 생성하면서 대규모 언어 모델의 강점이 되는 제로샷 성능을 효율적으로 활용하는 방법을 제시한다.

2. 관련 연구

다중 문서 그라운딩 된 대화형 질의 응답 작업

seq2seq 모델을 기반으로 사용자와의 대화를 목적으로 개발된 대화 시스템은 단순한 질의 응답 위주로 진행되는 자유 주제 대화 시스템(Open domain Dialogue System)부터 실제 상황을 제시하고 해당 상황에 대한 사용자의 목적을 해결해주기 위해 진행되는 목적 지향 대화 시스템(Task-oriented Dialogue System) 두 가지로 크게 구분이 된다. 이 중 목적 지향 대화 시스템의 경우 사용자가 처한 상황이 복합적일수록 시스템이 정보를 얻어야 하는 문서의 종류도 늘어나기 다중 문서에 기반한 새로운 작업 및 데이터 셋에 대한 연구는 활발히 진행중이다.

다중 문서를 활용한 생성 방식 중 Fusion-in-decoder [5]의 경우는 검색된 상위 문서를 여러개 활용하여 모델의 답변 생성 능력을 개선시키는 방법을 제시한다. 다양한 문서의 활용은 도메인에 대한 호환성이 높아지고 그만큼 다양한 주제의 데이터를 활용하여 응답을 생성할 수 있음을 나타낸다.

Prompt를 활용한 검색 증강형 언어 모델

대규모 언어 모델에 검색 작업을 추가할 때 prompt를 활용하여 모델이 해야할 작업을 구체적으로 지시해 준다면 정보 검색과 해당 정보를 바탕으로 더 정확하거나 풍부한 응답을 생성하는 데 도움이 될 수 있다.

In-context 검색 증강형 언어 모델[6]과 같은 작업은 최근 도입된 모델 아키텍처 변경을 통해 검색 증강형 언어 모델을 강화하는 기존의 시스템 대신 프롬프트와 문장의 맥락을 고려하고 외부 데이터에서 정보를 검색하여 언어 모델을 강화하는 방법을 제안했다. 문맥 중심의 검색은 검색 결과가 사용자의 질문이나 요구와 더 밀접하게 연결될 수 있어 생성하는 답변의 품질이 향상될 수 있다. 그리고 명확한 프롬프트는 모델에게 원하는 작업을 명확하게 알려주는 방법 중 하나이며 파인튜닝이 아닌 프롬프트를 통한 제로샷 혹은 원샷의 작업은 계산 비용과 시간 절약에도 큰 도움을 준다.

Zero shot Reranker로써의 LLM

LLM을 활용한 zero-shot 재순위화 작업[7]에 대해 제로샷 검색기는 종종 BM25 검색보다도 품질이 떨어지기 때문에 질의의 잠재적인 답변을 LLM에서 추출함으로써 제로샷 검색을 달성하는 방식을 제안한다. 또한 dense embedding 대신 Non-parametric Lexicon-based Retriever 방식을 사용하면서 문서 기반의 질의의 증강과 parametirc 기반의 model에서 발생하는 검색기의 성능 병목 현상이 없어지게 됨을 확인하였다. 따라서 대규모 언어 모델이 도메인 특정 주석이 달린 질의-문서 쌍 없이도 강력한 검색기로 단독으로 사용될 수 있음을 나타냈다.

3. 제안 방법

Retrieval

본 논문에서 제시한 검색 과정은 Dense Passage Retrieval 이용한 1차 검색과 Reranking을 통한 2차 검색을 거친다.

Dense Passage Retrieval(DPR)은 이중 인코더 프레임워크를 이용해서 질문과 문서의 dense representation을 어떻게 효율적으로 만들 것인지 다룬다. 텍스트의 semantic한 정보를 인코딩하여 벡터 공간 위로 표현하는 Dense Retrieval의 특징은 질문과 관련있는 문서의 형태가 맞지 않더라도 의미적으로 유사성을 갖는다면 질문과 문서를 매칭시킬 수 있음을 의미한다.

DPR의 기본적인 구조는 사전학습된 BERT[8] 기반의 두개의 인코더를 바탕으로 먼저 질문을 벡터 차원으로 매핑해주는 인코더 $\mathbf{E}_Q(\cdot)$ 와 문서를 벡터 차원으로 매핑해주는 인코더 $\mathbf{E}_P(\cdot)$로 나누어 각각 질문과 문서들의 임베딩을 얻어낸다.

입력된 질문과 문서를 각각 해당하는 인코더에 질문 q 와 문서p를 입력하여 인코딩 후 $\mathbf{E}_Q(\mathbf{q}) \in \mathbb{R}^{d_{model}}$ 와 $\mathbf{E}_P(\mathbf{p}) \in \mathbb{R}^{d_{model}}$ 얻어 식 (1)과 같이 벡터값들을 내적하여 질문-문서간 유사도를 측정한다.

$$\mathrm{sim}(\mathbf{q}, \mathbf{p}) = \mathbf{E}_Q(\mathbf{q})^T \mathbf{E}_P(\mathbf{p}) \quad (1)$$

각 질문에 대한 코퍼스의 유사도를 바탕으로 Top-k 상위 문서들을 1차 검색 후 해당 문서들을 바탕으로 2차 Reranking을 진행하게 되는데 Reranking은 사전학습된 BERT[8] 기반의 인코더를 사용하여 질문과 문서를 함께 인코딩하여 질문-문서간의 상관관계 점수를 distance score로 계산하여 앞서 DPR을 통해 구한 Top-k의 상위 문서들 사이에서 유사도를 2차로 계산한다.

Generation from LLM

앞서 진행된 검색 작업의 결과로 문서 코퍼스로부터 질문과 관련된 최상위 문서들을 찾아낸 후 다중 문서 대화 작업에 대한 대규모 언어 모델의 제로샷 응답 생성 성능을 확인한다.

디코더 기반의 대규모 언어 모델을 위하여 프롬프트를 작성하여 모델이 진행할 작업에 대해 지시하는 과정을 진행한다.

데이터가 영어인만큼 지시문은 영어로 작성 되었고 주어진 문서들과 대화 기록이 주어졌을때 질문에 대한 답변을 생성하라는 주제의 프롬프트를 작성했다.

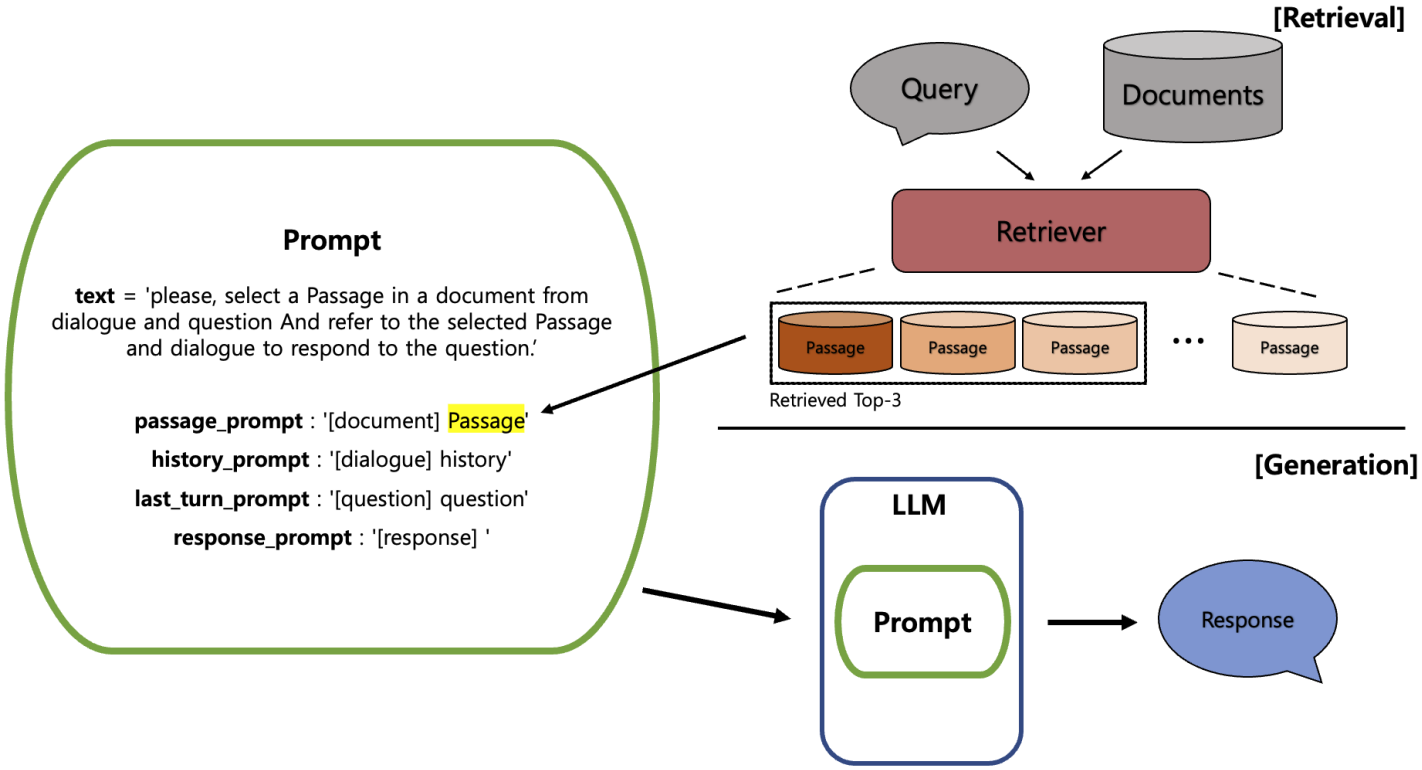

그림 1. 대규모 언어 모델을 이용한 다중 문서에 기반한 제로샷 답변 생성을 위한 구조 그림. [Retrieval] : 시스템에 입력된 질문과 문서 코퍼스를 이용하여 앞에서 진행된 대화 기록을 포함한 문맥과 가장 어울리는 문서 검색, 검색 결과로 선택된 상위 3개의 문서를 포함한 프롬프트 작성. [Generation] : 작성한 프롬프트를 대규모 언어 모델에 입력하여 상위 3개의 문서 중 다시 한번 모델이 문서를 선택할 수 있도록 하며 선택된 문서를 바탕으로 질문에 적합한 답변 생성.

4. 실험

4.1 실험 데이터

실험 및 평가를 위해 MultiDoc2Dial[9] 데이터 셋을 사용하였다. MultiDoc2Dial 데이터 셋은 4개의 도메인, 488개의 문서들을 바탕으로 평균 약 14번의 발화가 오가는 4,796개의 대화데이터로 구성되어 있다. 대화데이터의 질문-답변 쌍을 하나의 인스턴스로 취급하여 Train-Valid-Test 셋으로 나눌 경우 표 1과 같다.

표 1. Multi-doc2dial 데이터 셋

	Train	Valid	Test
Instance	21,451	4,201	4,094

4.2 Retrieval

본 논문에서는 언어 모델 RoBERTa-base[10]를 인코더**E**로 사용하여 DPR 모듈을 구축한다. 인코더에는 질문을 포함한 대화 기록 q = {[*BOS*],질문,[*SEP*],[*SEP*],기록,[*SEP*]}와 정답 문서 p = {[*BOS*],제목,[*SEP*],[*SEP*],내용,[*SEP*]}가 각각 입력되어 인코딩 된다. 그리고 식 (1)과 같이 벡터값들을 내적하여 얻을 수 있는 질문-문서 간 유사도를 측정하고 손실을 계산하여 역전파 하는 방식으로 학습된다.

질문-문서 쌍 총 N개가 모여 배치를 구성한다면 질문과 문서들은 각각 $(N \times d_{model})$ 크기의 행렬을 만들게 된다. 이 상태에서 식 (1) 과 같이 유사도를 곱하는 내적을 하게 되면 $(N \times N)$ 크기의 유사도 행렬을 구할 수 있다.

그렇다면 행렬의 i번째 행은 질문 q_i와 정답이 되는 문서 $p^+_{i(positive)}$와 오답이 되는 문서 $p^-_{i(negative)}$사이의 유사도로 구성된다. 이렇게 구해진 i번째 질문 q_i에 대한 유사도 집합의 손실함수는 식 (2)를 따라 소프트맥스 함수의 모양을 만들게 된다.

$$\mathcal{L}(q_i, p_i^+, p_{i,1}^-, \ldots, p_{i,n}^-) = -\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\mathrm{sim}(q_i, p_{i,j}^-)}} \quad (2)$$

각 배치의 행 마다 계산되는 소프트 맥스로부터 cross entropy loss 계산하여 효율적으로 역전파한다. 또한 배치의 크기가 커질수록 각 질문마다 만나게 되는 오답 문서들이 늘어나기 때문에 더 효과적인 학습이 가능하다.

baseline에 사용된 하이퍼 파라미터는 배치 사이즈-128, 학

습률-2e-05, Optimizer-Adam, Epoch-30의 조건으로 학습하였다.

이후 Reranking은 RoBERTa-large[10]를 인코더$\mathbf{E_{Re}}$로 사용하여 모듈을 구축한다. 인코더에는 대화 기록q과 정답 문서 p를 $q, p = \{[BOS]$,질문,$[SEP]$,$[SEP]$,기록,$[SEP]$,$[SEP]$,제목,$[SEP]$,$[SEP]$,내용,$[SEP]\}$로 입력되어 인코딩 된다. 인코딩의 결과로 모델이 갖는 cls 토큰의 $\mathbf{d_{model}}$ 차원의 벡터 공간 위로 $\mathbf{E_{Re}}$(q,p) $\in \mathbb{R}^{d_{model}}$로 표현되고 $(d_{model} \times 1)$ 크기의 MLP를 거쳐 크기 1의 dist$(\mathbf{q}, \mathbf{p})$ 스코어로 나타낼 수 있다.

각 질문마다 DPR을 통해 선정한 Top-k의 상위 문서들은 각각 인코딩의 결과로 dist$(\mathbf{q}, \mathbf{p_1})$, dist$(\mathbf{q}, \mathbf{p_2})$, ..., dist$(\mathbf{q}, \mathbf{p_k})$ 스코어 셋으로 표현되어 i번째 데이터 인스턴스에 대한 질문 q_i와 정답이 되는 문서 $p^+_{i(positive)}$와 오답이 되는 문서 $p^-_{i(negative)}$ 사이의 스코어로 구성된다.

이렇게 구해진 i번째 질문 q_i에 대한 스코어 집합의 손실함수는 식 (3)를 따라 소프트맥스 함수의 모양을 만들게 된다.

$$\mathcal{L}(q_i, p_i^+, p_{i,1}^-, \ldots, p_{i,k}^-) = -\log \frac{e^{\mathrm{dist}(q_i, p_i^+)}}{e^{\mathrm{dist}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\mathrm{dist}(q_i, p_{i,j}^-)}} \tag{3}$$

각 인스턴스 마다 계산되는 소프트 맥스로부터 cross entropy loss 계산하여 효율적으로 역전파한다.

baseline에 사용된 하이퍼 파라미터는 배치 사이즈-1, 학습률-1e-05, Optimizer-Adam, Epoch-5의 조건으로 학습하였다.

4.3 Zero-shot Generation from LLM

실험을 위해 대규모 언어모델로는 7B 규모의 Llama 모델('meta-llama/Llama-2-7b-hf')[11]을 사용한다. 디코더 기반의 언어모델로부터 다중 문서 기반의 질문에 대한 답변을 얻기 위해 프롬프트를 작성하여 모델이 할 일을 지시한다.

입력 프롬프트 :

- text = 'please, select a Passage in a document from dialogue and question And refer to the selected Passage and dialogue to respond to the question.'
- passage prompt : '[document] **Passage**'
- history prompt : '[dialogue] **history**'
- last turn prompt : '[question] **question**'
- response prompt : '[response] '

Input = text + passage prompt + history prompt + last turn prompt + response prompt

다중 문서에 대한 언어 모델의 제로샷 답변 생성 성능을 확인하기 위해 Passage prompt : '[document] **Passage**'에 입력될 Passage를 변수로 두고 성능을 관찰한다.

표 2. 다중 문서를 활용한 LLM의 제로샷 대화 답변 생성 성능

	F1(%)	Meteor(%)	Rouge(%)
Gold Passage	19.63	18.83	16.17
No Passage	15.55	14.41	12.49
Top-3 Passages	**15.82**	**14.62**	**12.62**
Top-1 Passage	14.99	13.90	12.16

Passage는 1) 정답 문서, 2)No Passage, 3) 검색 결과로 뽑은 Top-k의 문서들 (k=1,3) 총 세가지 경우의 변수로 입력하여 각각의 성능을 확인하였다.

성능의 평가는 F1-score, Meteor-score, Rouge-score을 지표로 이루어진다.

5. 결론

본 논문에서는 대규모 언어 모델을 이용한 다중 문서 기반 질의 응답 작업에 있어 대규모 언어 모델의 크기 탓에 파인튜닝시에 발생하는 제한과 검색 모듈의 한계를 개선하고자 학습 없이 제로샷으로 상위 다중 문서를 입력하여 모델이 문서들 중 적절한 문서를 선택하고 답변을 생성하도록 하는 방식을 제안하였다.

Llama-2를 이용한 다중 문서 기반의 질의 응답 성능을 확인한 결과 문서가 주어진 경우가 주어지 않은 경우는 성능면에서 확연한 차이를 보였고 검색 결과 Top-3의 문서를 프롬프트에 모두 입력하고 모델 스스로 답변의 근거가 될 문서를 선택하고 답변을 생성해본 결과 문서를 주지않은 경우보다 F1-score : 0.27% Meteor-score : 0.21% Rouge-score : 0.13%의 성능이 개선됨을 확인할 수 있었다. 그러나 Top-1의 문서를 활용하였을 때는 문서 없이 생성한 결과보다 오히려 낮은 수치를 보였는데 이 점은 프롬프트의 수정 혹은 문서 입력 개수에 따라 증가하는 입력 길이가 모델의 문맥 이해력에 영향을 주는 것은 아닌지 검증을 통해 확인을 해 봐야 할 부분이다.

6. Acknowledgement

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-02068, 인공지능 혁신 허브 연구 개발)

참고문헌

[1] S. Robertson and HugoZaragoza, "Theprobabilisticrelevanceframework: Bm25andbeyond," 2009.

[2] R. A. Shahzad Qaiser, "Text mining: Use of tf-idf to examine the relevance of words to documents," *International Journal of Computer Applications*, Vol. 181, No. 1, 2018.

[3] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. tau Yih, "Dense passage retrieval for open-domain question answering," *Empirical Methods in Natural Language Processing(EMNLP)*, Vol. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), p. 6769–6781, 2020.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[5] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," 2021.

[6] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham, "In-context retrieval-augmented language models," 2023.

[7] T. Shen, G. Long, X. Geng, C. Tao, T. Zhou, and D. Jiang, "Large language models are strong zero-shot retriever," 2023.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Association for Computational Linguistics*, Vol. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), p. 4171–4186, 2019.

[9] S. Feng, S. S. Patel, H. Wan, and S. Joshi, "Multidoc2dial: Modeling dialogues grounded in multiple documents," *Empirical Methods in Natural Language Processing(EMNLP)*, Vol. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, p. 6162–6176, 2021.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *The International Conference on Learning Representations (ICLR)*, 2020.

[11] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.