Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering

Gautier Izacard, Edouard Grave

Facebook AI Research, ENS, PSL University, Inria

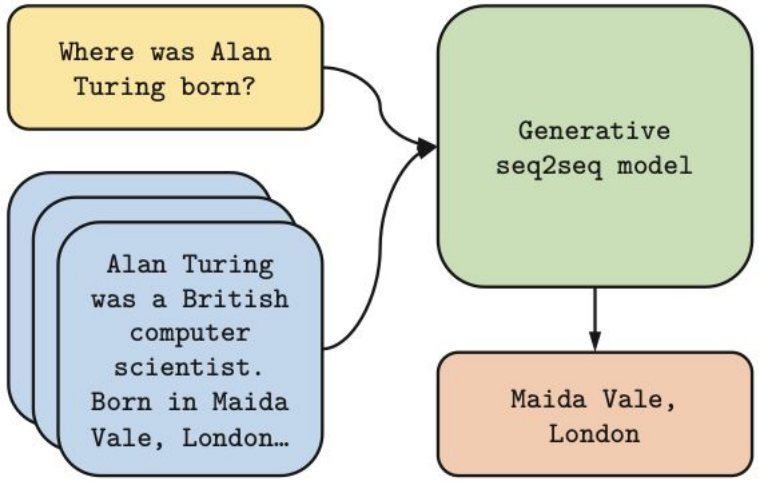
EACL 2021

발표자: 송선영

2024/06/03

Introduction

- 많은 연구들에서 대규모 언어 모델이 factual information을 추출할 수 있음을 발견
- 하지만, 이를 위해서는 수많은 parameter를 가지고 있어야 하기 때문에 훈련하는 데 드는 비용이 큼
- 이를 위해 외부 지식 소스(ex. Wikipedia)에서 관련 문서를 검색해 답변을 생성하는 Retrieval-augmented Generation(RAG)방식이 제안되었음

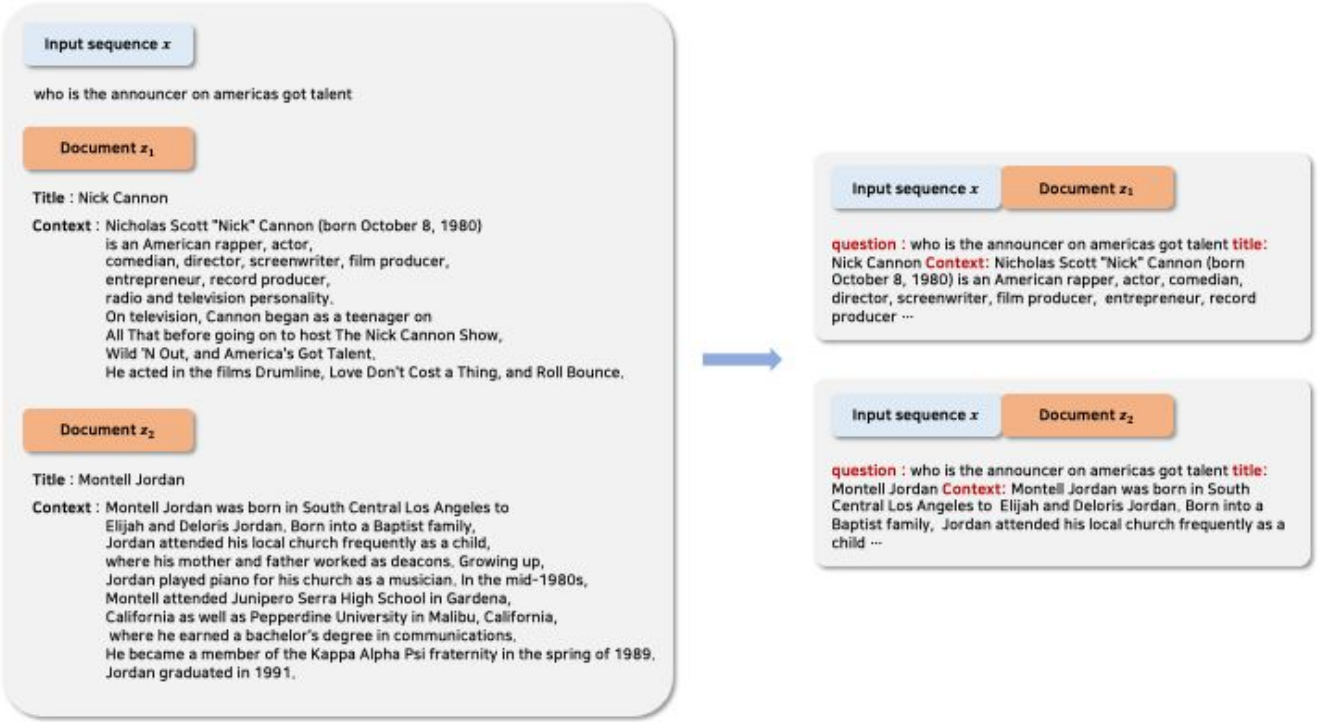


-> 이 논문에서는 또다른 구조의 retrieval augmented language model **FiD(Fusion-in-Decoder)**를 제안

FiD
(Fusion-in-Decoder)

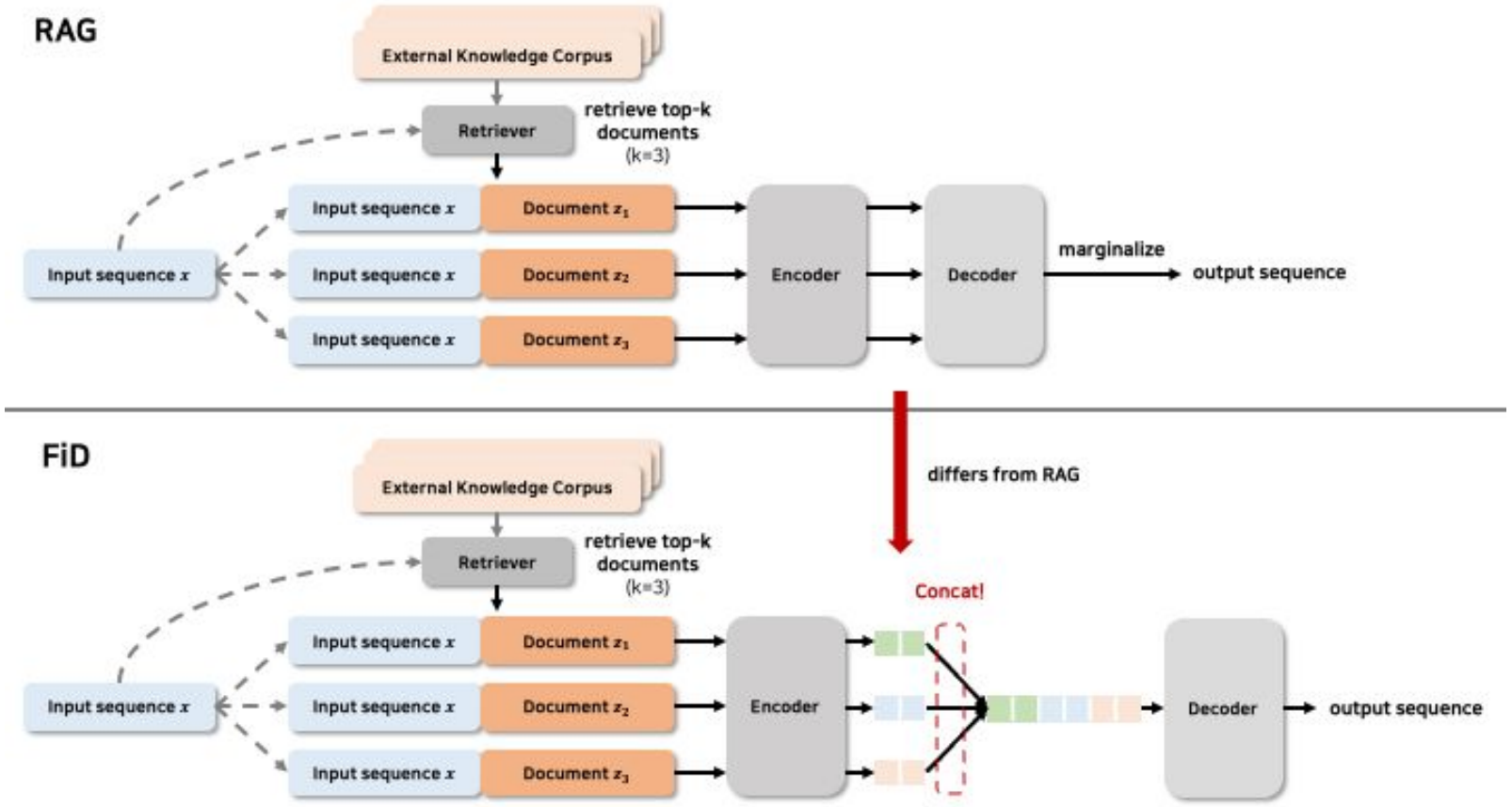
- Retriever-Reader 로 구성
  - Retriever: BM25, DPR
  - Reader: T5

- 각 텍스트 앞에 특수 토큰 “question:”, “title:”, ”context:” 를 추가



FiD
(Fusion-in-Decoder)

- RAG와 FiD의 차이
  - FiD는 RAG와 달리 **여러 document에 대한 encoder의 output을 concat해 decoder에서 한번에 처리**



Datasets

- **Dataset**
  - Open-Domain QA dataset에 대해 실험
    - NaturalQuestions (DPR)
    - TriviaQA (DPR)
    - SQuAD v1.1 (BM25)

- **Metric**
  - Exact Match (EM)

- **Backbone**
  - T5-base(220M), T5-large(770M)

- **Settings**
  - Learning rate: 1e-4
  - dropout: rate 10%
  - Batch size: 64
  - Optimizer: Adam
  - Steps: 10,000 (500마다 검증)
  - 학습과 테스트 시, 100개의 passage 검색

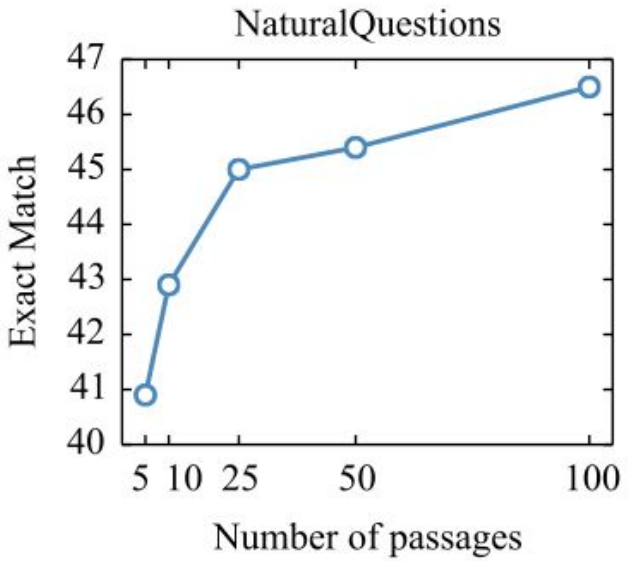
Experiments

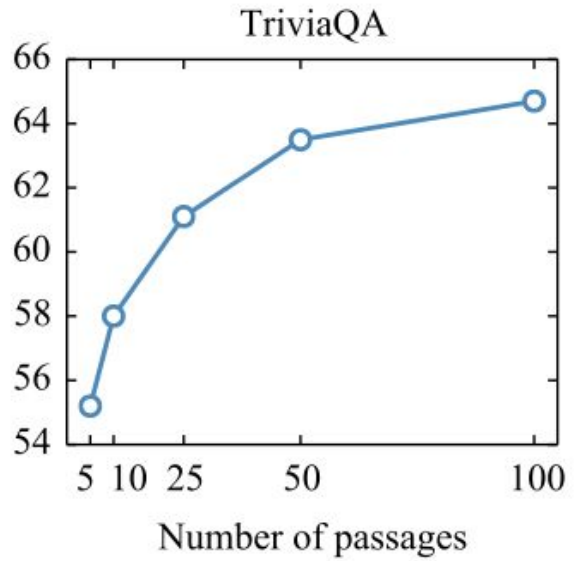
- Generative model이 Extractive model 보다 좋은 성능을 보임
- Retrieval을 사용하면 성능이 향상됨을 확인
- FiD 방법들이 NQ, TriviaQA dataset에서 SoTA 달성

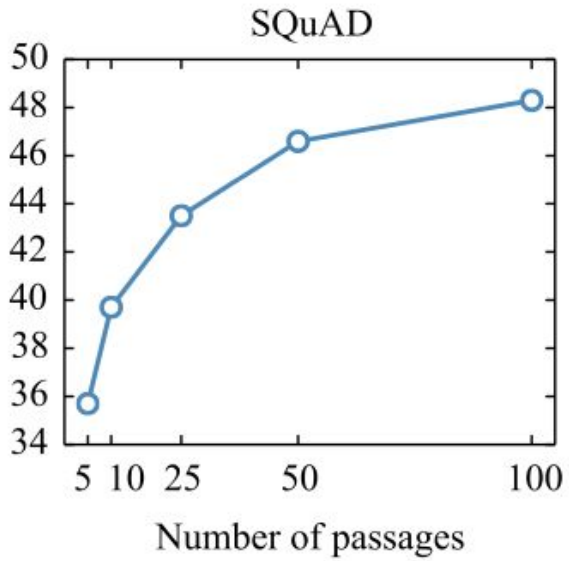
Model	NQ	TriviaQA		SQuAD Open	
	EM	EM	EM	EM	F1
DrQA (Chen et al., 2017)	-	-	-	29.8	-
Multi-Passage BERT (Wang et al., 2019)	-	-	-	53.0	60.9
Path Retriever (Asai et al., 2020)	31.7	-	-	**56.5**	**63.8**
Graph Retriever (Min et al., 2019b)	34.7	55.8	-	-	-
Hard EM (Min et al., 2019a)	28.8	50.9	-	-	-
ORQA (Lee et al., 2019)	31.3	45.1	-	20.2	-
REALM (Guu et al., 2020)	40.4	-	-	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-	36.7	-
SpanSeqGen (Min et al., 2020)	42.5	-	-	-	-
RAG (Lewis et al., 2020b)	44.5	56.1	68.0	-	-
T5 (Roberts et al., 2020)	36.6	-	60.5	-	-
GPT-3 few shot (Brown et al., 2020)	29.9	-	71.2	-	-
Fusion-in-Decoder (base)	48.2	65.0	77.1	53.4	60.6
Fusion-in-Decoder (large)	**51.4**	**67.6**	**80.1**	**56.7**	63.2

Experiments

- Document의 수를 늘릴수록 성능이 좋아짐을 확인







Experiments

- 계산 비용 문제를 완화하기 위해, Training 시 적은 document의 수로 학습하고 평가
  - Testing 시에는 100개의 document 검색

- 적은 document로 학습 시 성능이 하락함을 발견
- 적은 document로 학습 후, 100개의 document로 1000 step동안 fine-tuning 하면 성능 격차 감소→ **효율적인 학습 방법**

	NaturalQuestions		TriviaQA	
Training Passages	w/o finetuning	w/ finetuning	w/o finetuning	w/ finetuning
5	37.8	45.0	58.1	64.2
10	42.3	45.3	61.1	63.6
25	45.3	46.0	63.2	64.2
50	45.7	46.0	64.2	64.3
100	46.5	-	64.7	-

Conclusion

- 이 연구는 ODQA task에 대한 간단한 접근 방식인 FiD를 제안
- FiD는 support passage를 검색한 후, 이를 합쳐 생성 모델로 입력으로 주고 답변을 생성
- 제안 방법은 ODQA task에서 SoTA를 달성했으며, 검색된 passage의 수 확장에 따라 성능이 향상됨을 보여줌
- 향후 연구로 검색 모델 학습을 통합한 end-to-end 학습을 목표로 함

Open Questions

- FiD에서는 passage의 수를 늘릴수록 성능이 향상되는 것에 대해, 생성 기반의 Seq2Seq 모델이 여러 passage의 정보를 잘 활용하기 때문이라고 했는데, RAG에서는 왜 그렇지 않은지?
  - FiD는 T5, RAG는 BART

Thank You

감사합니
다.