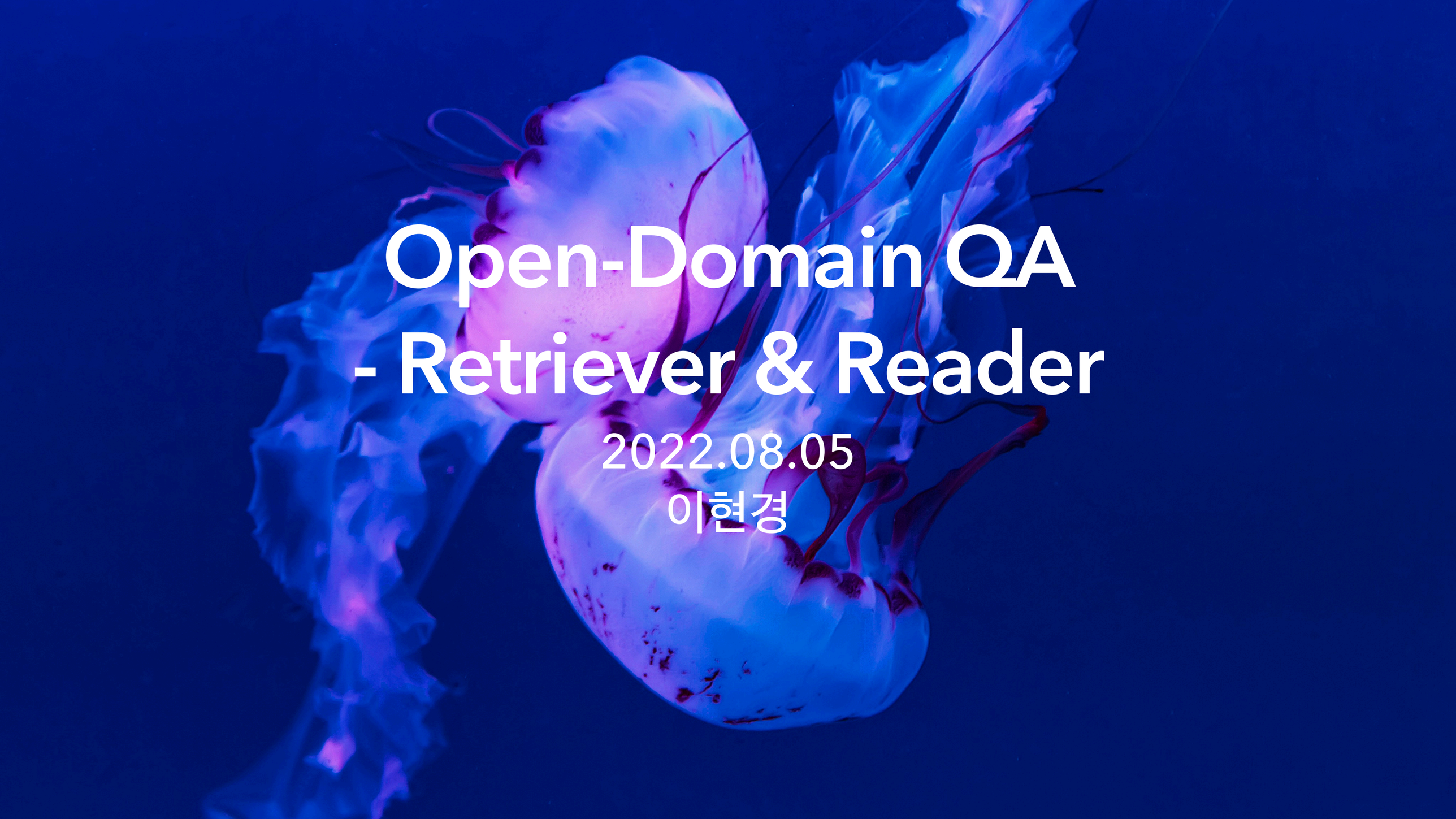


Open-Domain QA
- Retriever & Reader

2022.08.05

이현경

목차

3가지 논문

- Dense Passage Retrieval for Open-Domain Question Answering
  - 새로운 **Retriever** 기법 제안

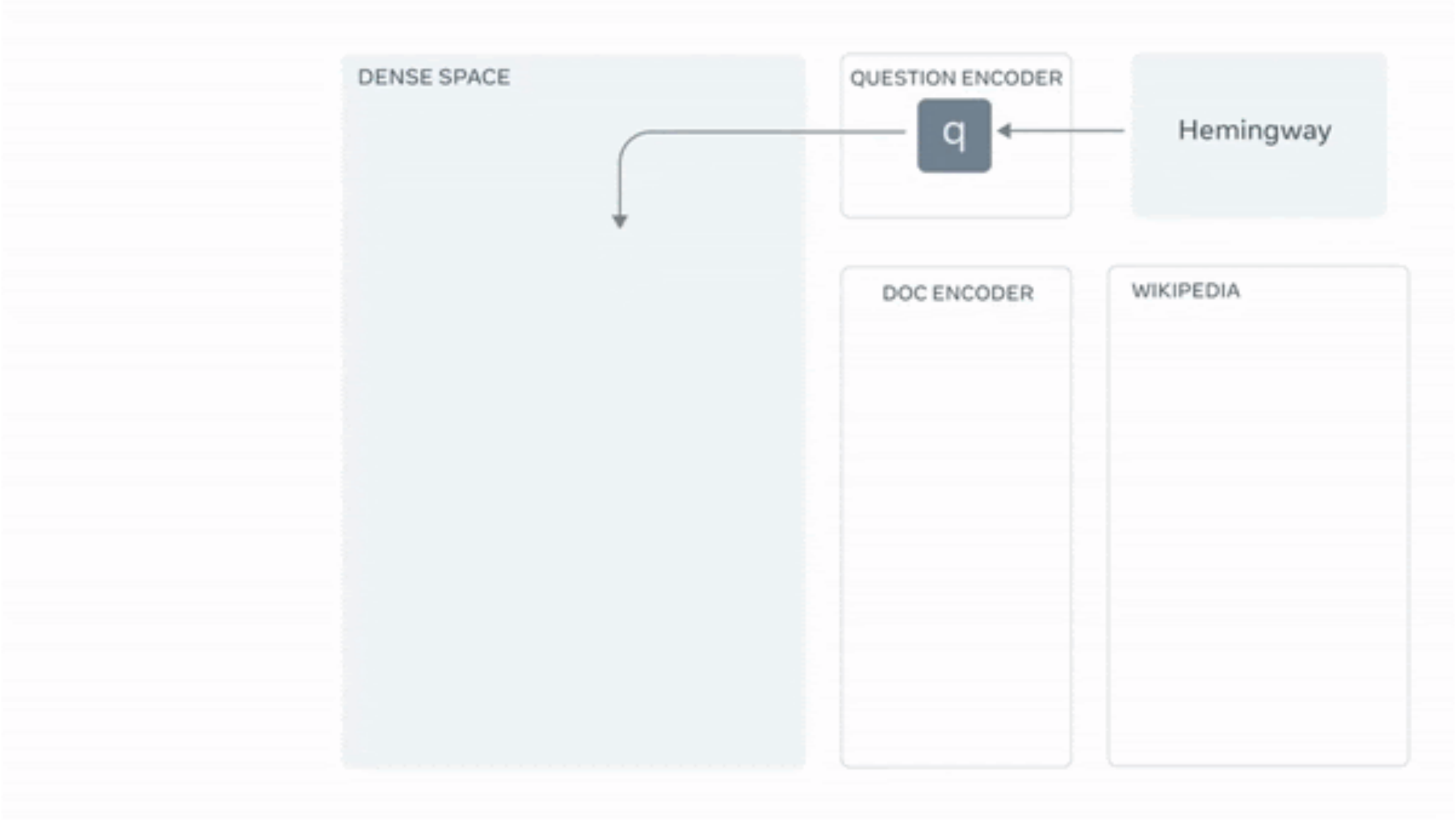
- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks
  - 새로운 **모델 아키텍쳐** 제안 (Retriever ~ Generator)

- Improving Passage Retrieval with Zero-Shot Question Generation
  - Retrieval 개선 과정으로 새로운 **Re-ranker** 기법 제안

Open-Domain QA

- **대규모 document collections**에서 정보를 찾는 **질문**에 대한 **답변**을 생성하는 태스크
- Common approach는 **retriever**와 **reader** 네트워크 컴포넌트로 구성

An overview of RAG



Demo 예시

Retrieval Augmented Generation

Ask a factoid question:

> The model was trained on the Natural Questions dataset of factoid questions, and should be able to answer any question of the same format.

What would you like to ask? ---- select <FREE INPUT> to enter a new query

Which book is Ursula Le Guin best know for?

The model generated the following answers:

> The generator model's beam search allows us to consider several possible outputs for a given query. Here are the top 4 outputs with their generation scores:

- *Model outputs:*
  1. **a wizard of earth** *(-17.53)*
  2. **a wizard of earthsea** *(-17.86)*
  3. **ishi in two worlds** *(-20.06)*
  4. **the left hand of darkness** *(-20.52)*

Demo 예시

Give a target answer:

The model was trained to create Jeopardy!-style questions for a given answer. You can find the training dataset here

What would you like to ask? ---- select <FREE INPUT> to enter a new query

Toussaint Louverture

The model generated the following answers:

> The generator model's beam search allows us to consider several possible outputs for a given query. Here are the top 4 outputs with their generation scores:

- *Model outputs:*
  1. **This coffee liqueur is named for a leader of the Haitian Revolution** *(-17.12)*
  2. **The name of this coffee liqueur is French for "old man"** *(-17.57)*
  3. **The name of this coffee liqueur is French for "saint"** *(-17.59)*
  4. **This coffee liqueur is named for a leader of the French Revolution** *(-17.93)*

Demo 예시

Example-level contribution of Wikipedia passages:

When presented with a query, the model retrieves a set of support documents using a dot product, and weighs each of them using the retrieval score when generating the answer. Scroll down to show the full text of the retrieved passages:

-- Article 1 -- Toussaint Louverture
0.23

-- Article 2 -- Toussaint Coffee Liqueur
0.20

-- Article 3 -- Toussaint Coffee Liqueur
0.15

-- Article 4 -- Impressionism
0.11

-- Article 5 -- Impressionism
0.11

-- Article 6 -- Toussaint Louverture Int...
0.13

-- Article 7 -- Toussaint Louverture
0.05

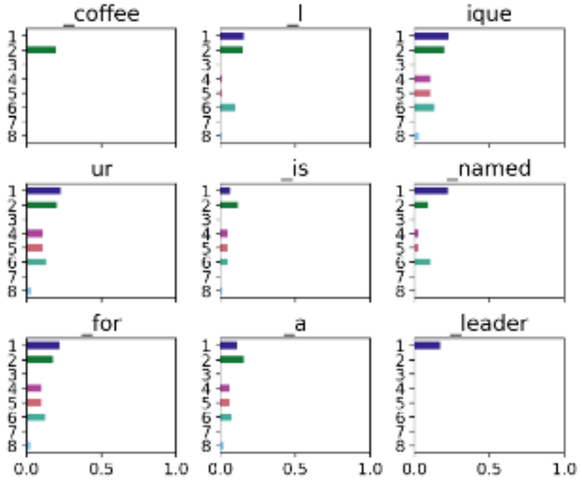
-- Article 8 -- Amadou Toumani Touré
0.03

Word-level contribution of Wikipedia passages:

When generating the output, the contribution of each passage to the next token depends both on its retrieval score and on the tokens that have been generated previously. Select a generated output to show token-level provenance information:

This coffee liqueur is named for a l...

- **This**-1- **coffee**-2- **liqueur**-6- **is**-2- **named**-1- **for**-1- **a**-2- **leader**-1- **of**-1- **the**-1- **Haitian**-1- **Revolution**-1-



Show token-level passage contributions starting from:
1

Dense Passage Retrieval for Open-Domain Question Answering

2020 Apr, 712회 인용

(arXiv)

Facebook AI University of Washington Princeton University

DPR

Dense Passage Retriever

- 후보 context를 고르기 위한 **passage retriever**가 중요하다
- 기존에는 **Sparse Vector Model**을 사용했었음
- 해당 논문에서 dual-encoder 모델 활용한 **Dense Passage Retriever** 제시

Retriever의 변화

Sparse vector model → dense passage retriever

- 효과
  - **표현력**을 높일 수 있음
  - **임베딩**이 고정된 LUT(Look Up Table)가 아닌, **trainable**하다는 장점이 있다.

Retriever 비교

Sparse Models VS Dense Models

- **어휘적 일치(Lexical Overlap)**
  - **Question:** 가장 작은 펭귄 종은 무엇입니까?
  - **Answer:** 쇠푸른펭귄
  - **Gold Passage:** … 가장 작은 펭귄 종은 일반적으로 발견되는 쇠푸른펭귄입니다 …
- **의미적 일치(Semantic Overlap)**
  - **Question:** 반지의 제왕에 나오는 나쁜놈은 누구야?
  - **Answer:** 사우론
  - **Gold Passage:** . … 스토리 상 주요 적수인 다크 로드 사우론은 이전에 …

Dense Passage Retriever
수식

- 서로 다른 인코더(BERT) *E_Q, E_P* 사용하여 question, passage를 *d* 차원에 **임베딩**
- 두 임베딩 간 **유사성 계산**

$$\text{sim}(q, p) = E_Q(q)^\intercal E_P(p)$$

Encoder
특징

- BERT 사용
- [CLS] token output을 임베딩 값으로 사용
- 임베딩 matrix 차원 : (batch_size x embeded_dim)

Inference
특징

- 모든 passage를 학습된 E_P에 임베딩시키고
- **FAISS**를 이용해 index를 달아준다
  - index를 통해 question이 주어졌을 때 가까운 passage vector를 빠르게 Inference 할 수 있음

Training
원리

- “The goal is to create a vector space such that **relevant pairs** of questions and passages will have **smaller distance**”
  - 관련 있는(Positive) question 및 passages는 **가깝게**,
  - 관련 없는(Negative) question 및 passages는 **멀게** 임베딩

Training
수식

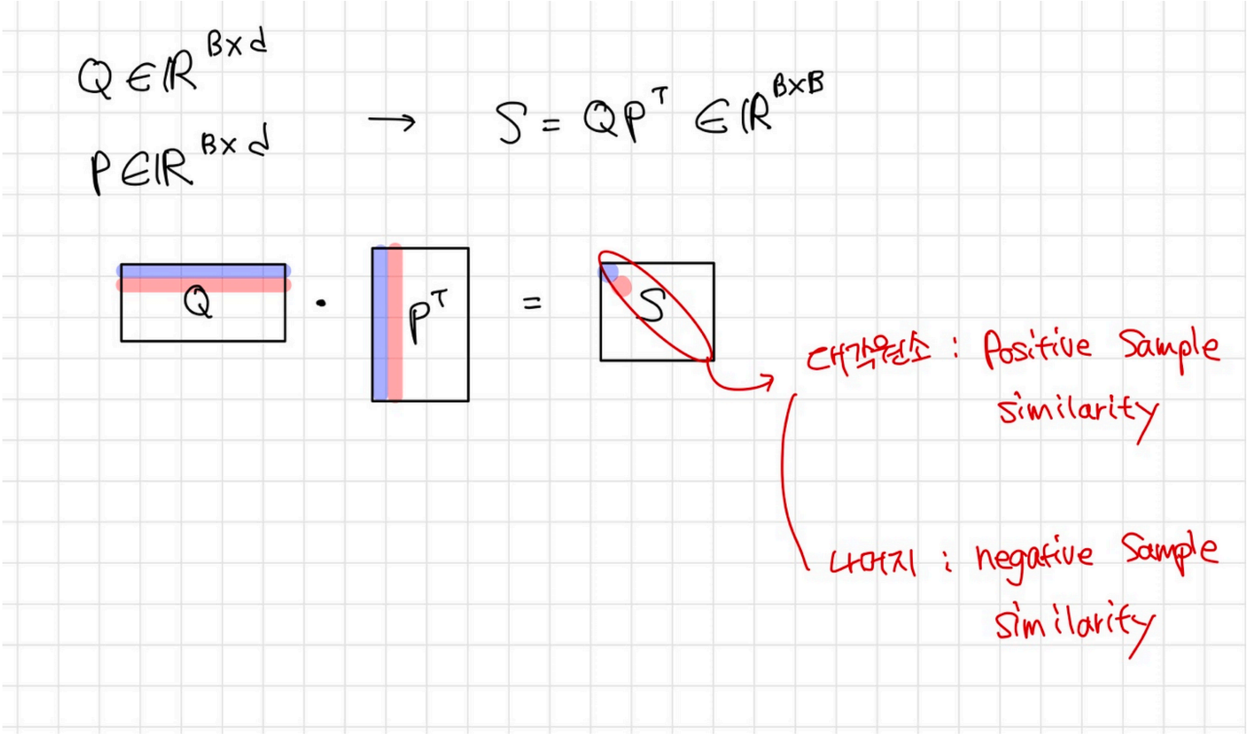
- Negative log likelihood loss
  - 하나의 positive sample과 n개의 negative sample을 하나의 instance로 학습

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

In-batch negative

빠르고 효과적인 Training method

- 따로 positive, negative sample 지정 않고 **하나의 batch** 안에서 계산 가능
- S(유사도, Similarity)에 softmax 취한 뒤, Positive sample을 타겟으로 하여 NLL loss 계산해줌



Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

2020 May, 368회 인용

(arXiv)

Facebook AI Research University College London New York University

RAG

Retrieval-Augmented Generation

- Augmented?
  - **외부 지식**(ex. Wikipedia 문서) 활용하여 지식을 보강하고 지식 집약적 태스크에 대한 SOTA를 달성하는 새로운 아키텍쳐
  - Seq2seq 모델 인풋에 **문맥 정보**를 반영해서 좀더 **특정 domain에 특화**된 성능 결과
- Information **retrieval** component **+** seq2seq **generator** 결합한 **end-to-end**의 미분 가능한 모델 제안

Facebook AI의 DPR + Facebook AI의 BART 사용

모델 아키텍쳐

Retriever + Generator 구조

- **Retriever**
  - BERT 기반 pre-trained neural retriever(DPR)
  - 문맥정보를 반영하는 고정된 임베딩 모델

- **Generator**
  - BART 기반 seq2seq transformer

모델 아키텍쳐

Retriever + Generator 구조

- **Retriever**
  - BERT 기반 pre-trained neural retriever(DPR)
  - 문맥정보를 반영하는 고정된 임베딩 모델
- **Generator**
  - BART 기반 seq2seq transformer

Non-parametric memory
+
Parametric memory

의 결합

모델 아키텍쳐

Retriever + Generator 구조

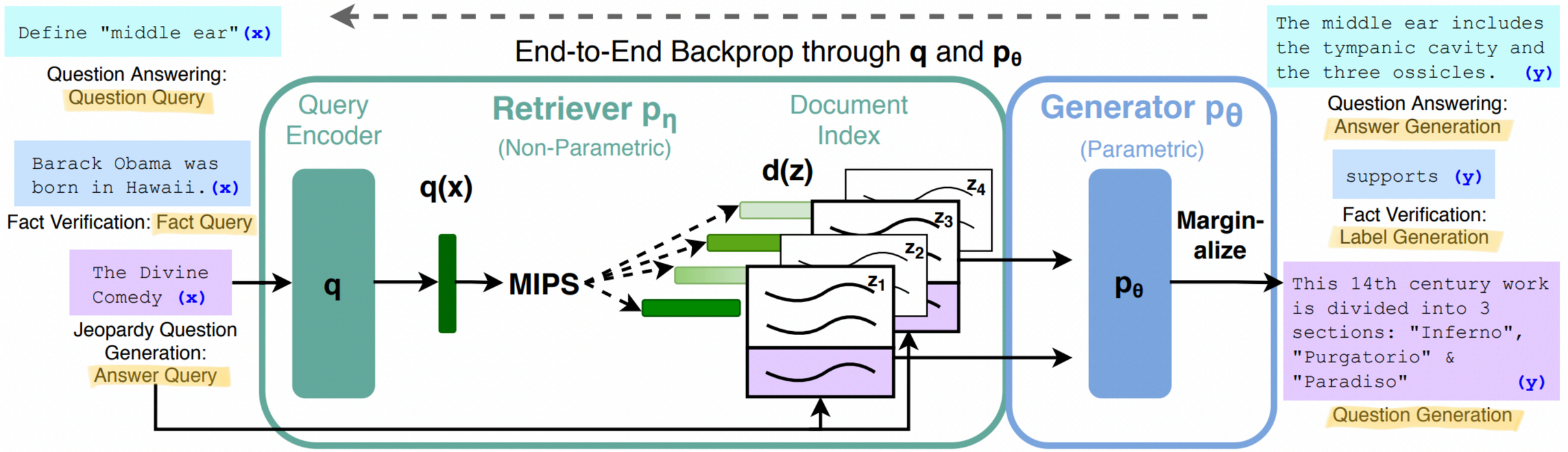
- **Retriever**
  - BERT 기반 pre-trained neural retriever(DPR)
  - 문맥정보를 반영하는 고정된 임베딩 모델
- **Generator**
  - BART 기반 seq2seq transformer

두 컴포넌트를
End-to-end probabilistic model로
결합하여 학습

모델 아키텍쳐

Overview

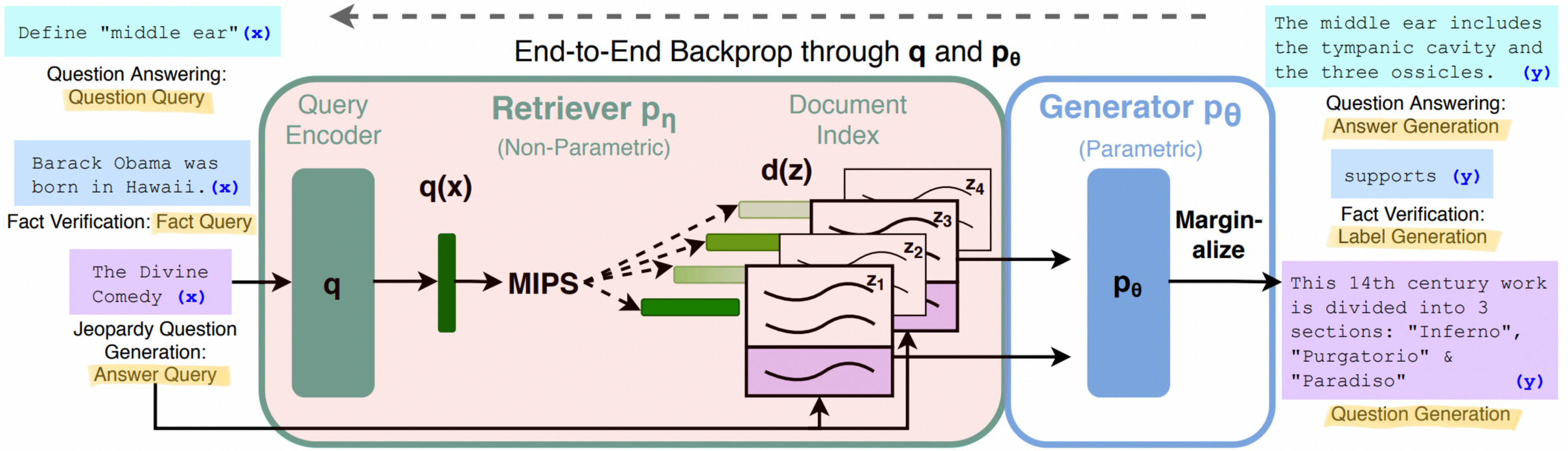
- Query (x) → Answer/Label/Question (y)



모델 아키텍쳐

Overview

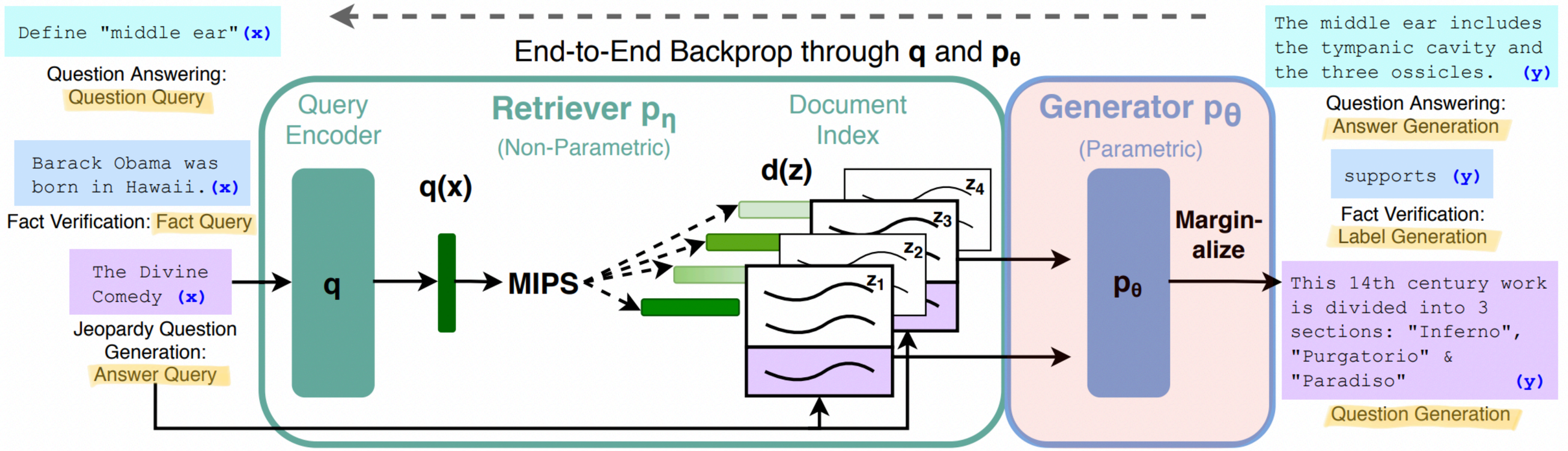
- Pre-trained Retriever: **Query Encoder + Document Index**



* Maximum Inner Product Search(MIPS) : 모든 documents (*z_i*) 중 top-K 가장 관련 있는 documents 찾기 위해 사용한 기법

모델 아키텍쳐
Overview

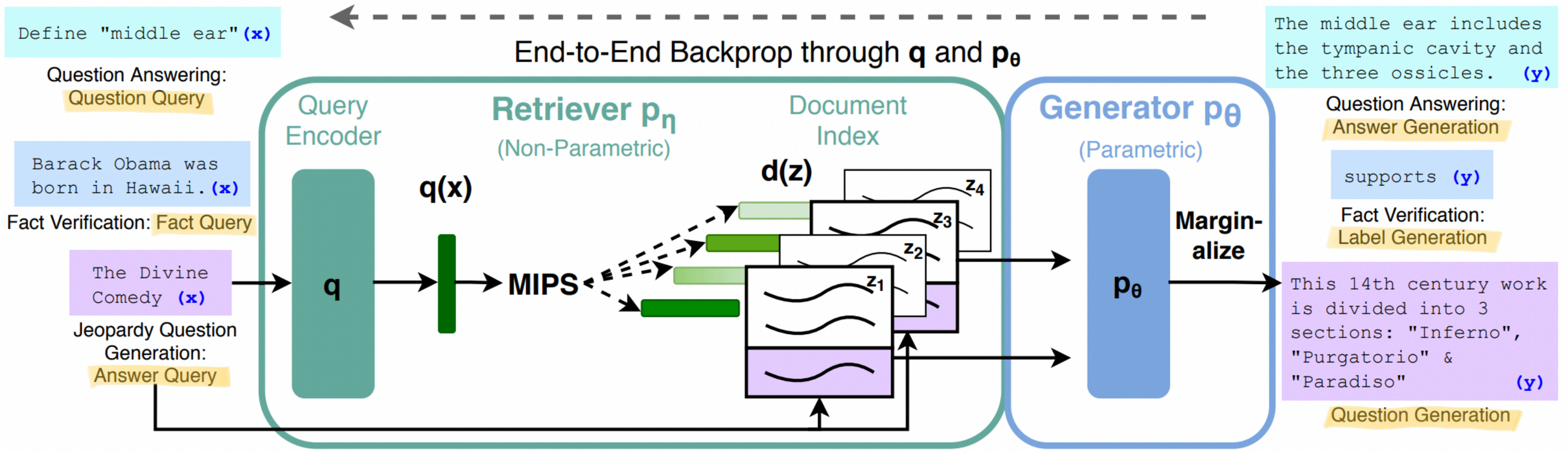
- Pre-trained encoder-decoder: **Generator**



* Marginalize : 검색된 documents (*z*)를 잠재 변수로 취급하여 generated text에 대한 분포를 생성하기 위해 marginalization 함

모델 아키텍쳐
Overview

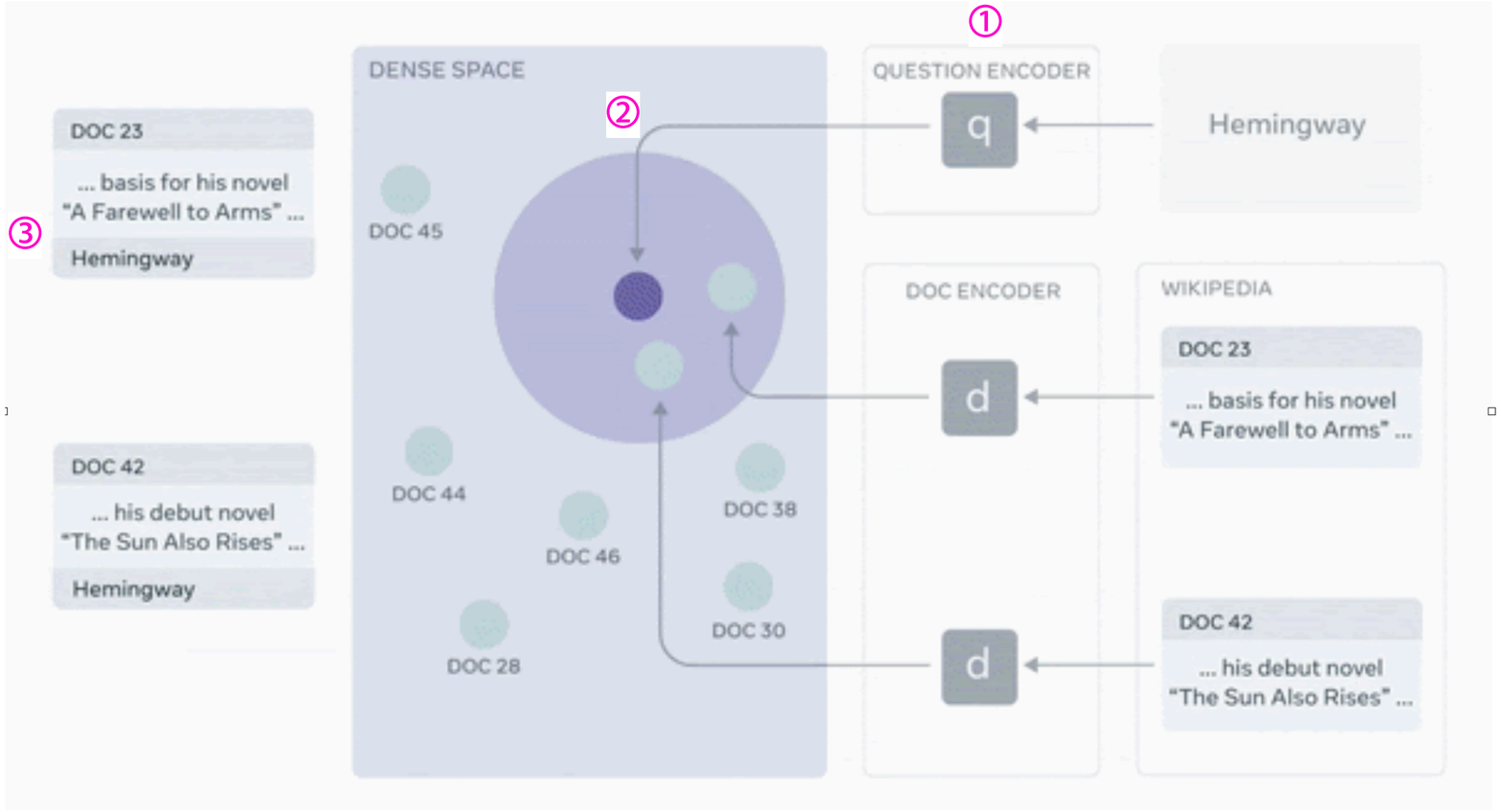
- Training : question-answer pairs(x, y), Adam, NLL



학습 시 Document 인코더는 고정시키고, Query 인코더와 Generator만 학습시킴

Retriever
원리

- BERT 기반의 **DPR 모델** 사용



Retriever
수식

- Input sequence: *x*, Wikipedia document: *z*

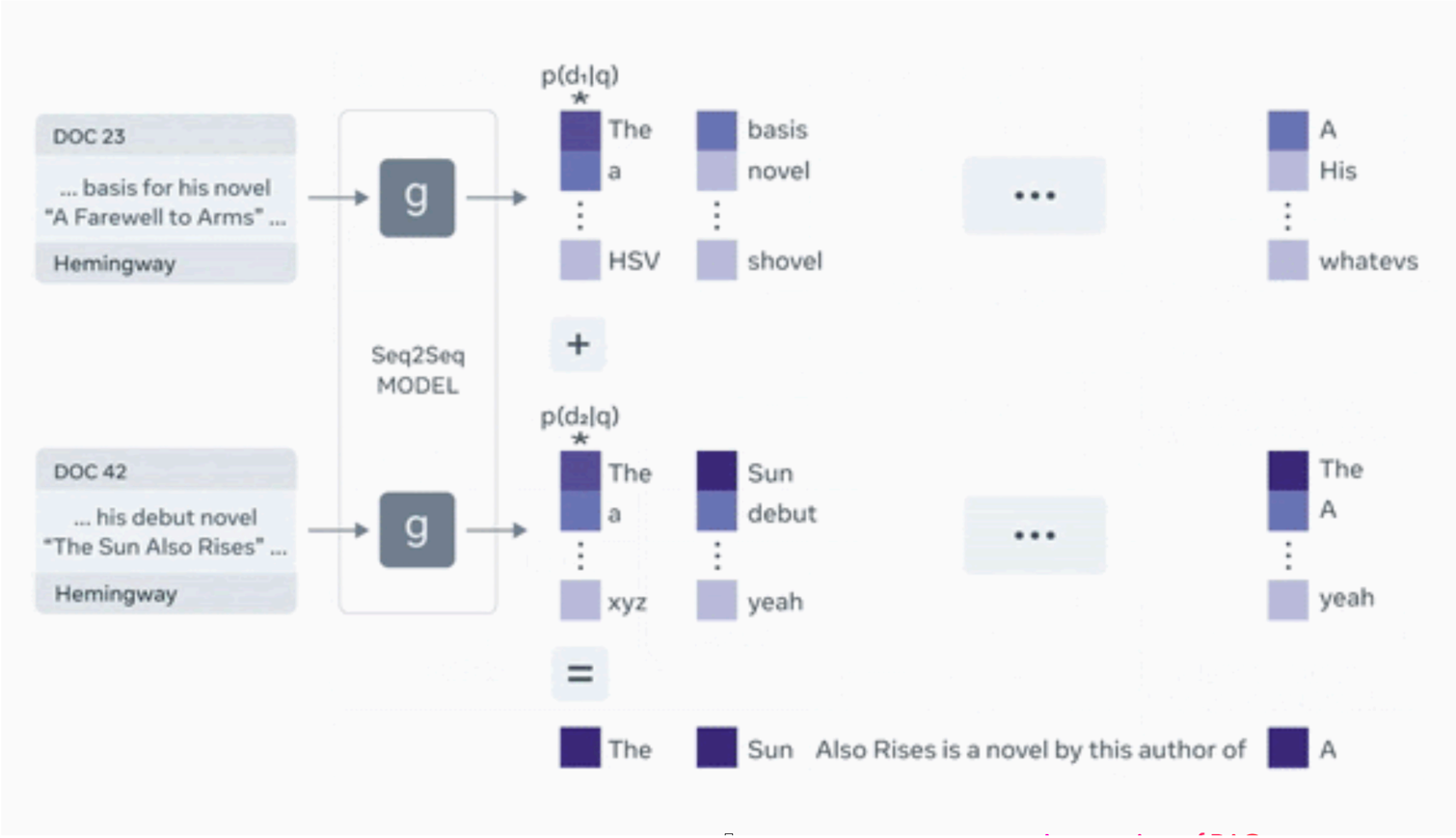
$$retrieval: P_\eta(z|x) = exp(d(z)^T q(x))$$
$$d(z) = bert_d(z),\ q(x) = bert_q(x)$$

- 가장 높은 확률의 *P(z|x)*를 찾기 위해 MIPS(maximum inner product search) 사용

Generator

원리 및 수식

- **BART 모델** 사용



$$generator : p_\theta(y_i|x, z, y_{i-1})$$

Training

End-to-end 모델을 jointly training

- 2가지 모델 제안
  - top-k개의 documents(z)에 대해 marginalize하는 방식에 따라

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

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

하나의 latent z 로 여기고 동일한 문서를 사용함 VS. 서로 다른 문서(latent z)를 사용함

Training

End-to-end 모델을 jointly training

- RAG-Token document posterior

Document 1: his works are considered classics of American literature ... His wartime experiences formed the basis for his novel **"A Farewell to Arms"** (1929) ...

Document 2: ... artists of the 1920s "Lost Generation" expatriate community. His debut novel, **"The Sun Also Rises"**, was published in 1926.

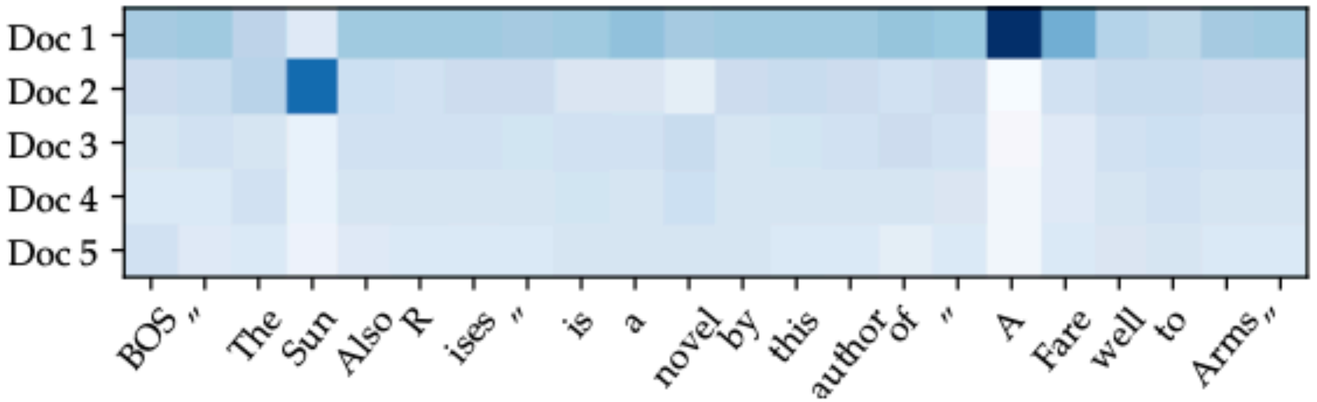


Figure 2: RAG-Token document posterior $p(z_i|x, y_i, y_{-i})$ for each generated token for input "Hemingway" for Jeopardy generation with 5 retrieved documents. The posterior for document 1 is high when generating "A Farewell to Arms" and for document 2 when generating "The Sun Also Rises".

하나의 latent z 로 여기고 동일한 문서를 사용함 VS. 서로 다른 문서(latent z)를 사용함

Improving Passage Retrieval with Zero-Shot Question Generation

2022 Apr, 2회 인용

(arXiv)

McGill University Mila - Quebec AI Institute Meta AI University of Washington

UPR

Unsupervised Passage Re-ranking

- Open-domain QA 분야에서 **passage 검색 개선**을 연구
- **새로운 Re-ranker**로 UPR을 제안
  - UPR 통해 passage를 retrieval하는 과정에서 성능 개선
- A re-ranker based on zero-shot question generation with PLM

Approach

Overview

- New **re-ranker**

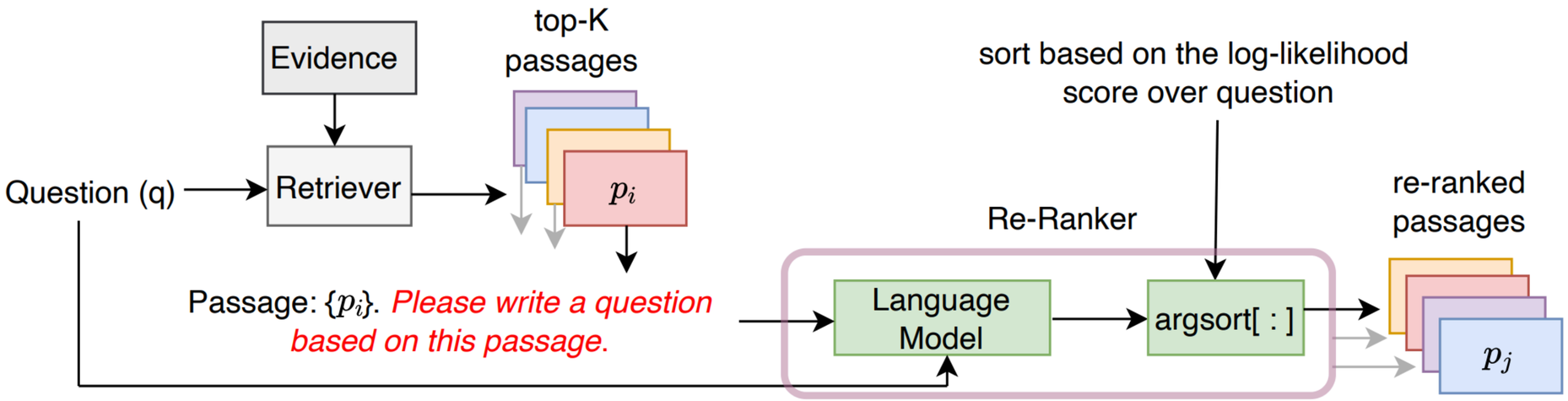


Figure 2: An illustration of the different components in UPR. For more details, please refer to text.

Retrieval Accuracy

Before VS After re-ranking

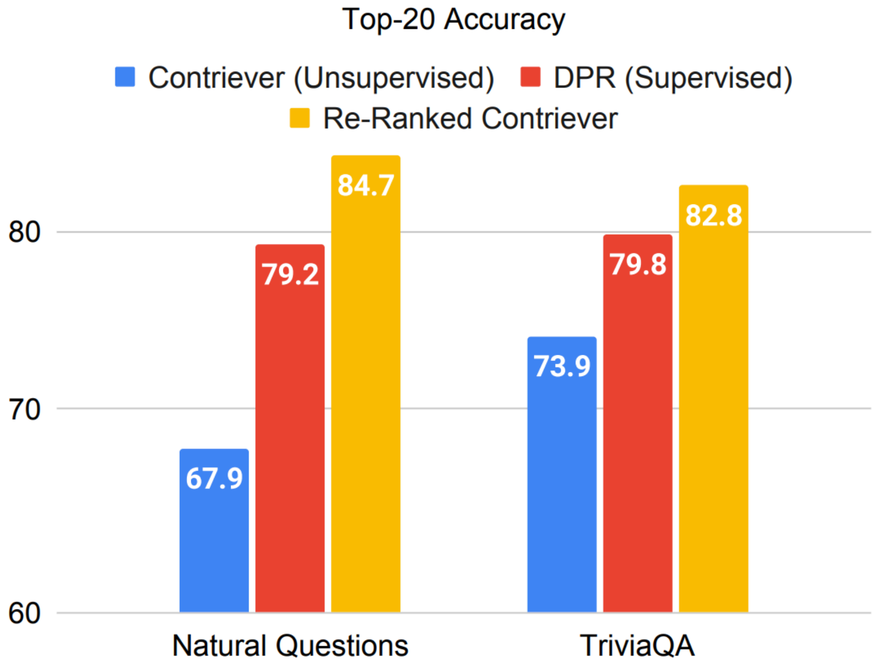


Figure 1: After UPR re-ranking of the Contriever's (unsupervised) (Izacard et al., 2021) top-1000 passages, we outperform strong supervised models like DPR (Karpukhin et al., 2020) on Natural Questions and TriviaQA datasets.

Re-ranker 방식

주요 특징

- **UPR**
  - off-the-shelf PLMs 사용, training data나 fine-tuning 필요 없으나 성능 향상
  - A fully unsupervised pipeline(consisting of a retriever and re-ranker)

- **“Question Generation model”**
  - Zero-shot

“*Please write a question based on this passage*”

UPR

Method

- the **goal** of the re-ranker is
  - To **reorder** top-K retrieved passages such that a passage with the **correct answer** is ranked as **high** as possible.

- **A relevance score**

$$p(\boldsymbol{z}_i \mid \boldsymbol{q}) \text{ for each passage } \boldsymbol{z}_i \in \mathcal{Z}.$$

UPR

Method

- UPR uses a **PLM**
  - to score the probability of generating the **question *q*** given the **passage text *z***

UPR
수식

- Bayes' rule

$$\log p(\boldsymbol{z}_i \mid \boldsymbol{q}) = \log p(\boldsymbol{q} \mid \boldsymbol{z}_i) + \log p(\boldsymbol{z}_i) + c \,,$$

$$\log p(\boldsymbol{z}_i \mid \boldsymbol{q}) \propto \log p(\boldsymbol{q} \mid \boldsymbol{z}_i),\ \forall \boldsymbol{z}_i \in \mathcal{Z} \,.$$

$$\log p(\boldsymbol{q} \mid \boldsymbol{z}_i) = \frac{1}{|\boldsymbol{q}|} \sum_t \log p(q_t \mid \boldsymbol{q}_{<t}, \boldsymbol{z}_i; \Theta) \,.$$

Experiments

Datasets Setup

- 1. Open-Domain QA Datasets
  - SQuAD-Open
  - TriviaQA
  - Natural Questions
  - WebQuestions

- 2. Keyword-centric Datasets
  - Entity Questions
  - BEIR Benchmark

Experiments: Passage Retrieval

Retriever

- 1. Unsupervised Retrievers
  - dense
  - sparse
  - dense
- 2. Supervised Retrievers
  - dense
  - dense
- 3. E2E supervised
  (기존 SOTA)

Retriever	SQuAD-Open		TriviaQA		NQ		WebQ		Average	
	Top-20	Top-100	Top-20	Top-100	Top-20	Top-100	Top-20	Top-100	Top-20	Top-100
Unsupervised Retrievers										
MSS	51.3	68.4	67.2	79.1	60.0	75.6	49.2	68.4	56.9	72.9
MSS + UPR	75.7	80.8	81.3	85.0	77.3	81.5	71.8	80.4	76.5	81.9
BM25	71.1	81.8	76.4	83.2	62.9	78.3	62.4	75.5	68.2	79.7
BM25 + UPR	83.6	87.4	83.0	86.4	78.6	85.2	72.9	81.4	79.5	85.1
Contriever	63.4	78.2	73.9	82.9	67.9	80.6	74.9	80.1	70.0	80.5
Contriever + UPR	81.3	85.6	82.8	86.4	84.7	87.0	80.9	83.5	82.4	85.6
Supervised Retrievers										
DPR	59.4	74.5	79.8	85.1	79.2	85.7	74.6	81.6	73.3	81.7
DPR + UPR	80.7	85.4	84.3	87.2	83.4	88.6	81.6	84.1	82.5	86.3
MSS-DPR	73.1	84.5	81.9	86.6	81.4	88.1	76.9	84.6	78.3	86.0
MSS-DPR + UPR	**85.2**	**89.4**	**84.8**	**88.0**	**83.9**	**89.4**	**77.2**	**85.2**	**82.8**	**88.0**
E2E Supervised	-	-	84.1	87.8	84.8	89.8	79.1	85.2		

Table 2: Top-{20, 100} retrieval accuracy on the test set of datasets before and after UPR re-ranking of the top-1000 retrieved passages with the T0-3B model. Best results of the unsupervised retriever are underlined while those of the supervised retriever are highlighted in bold. For reference, we also include the state-of-the art supervised results in the last row, which is obtained from end-to-end or joint training of the retriever and language model using question-answer pairs (Sachan et al., 2021a,b).

Experiments: Passage Retrieval

Retriever

Retriever	Entity Questions Top-20	Entity Questions Top-100
Baselines		
MSS	51.2	66.3
DPR	51.1	63.8
MSS-DPR	60.6	73.7
Contriever	63.0	75.1
BM25	71.2	79.8
SPAR (Chen et al., 2021)	74.0	82.0
After Re-ranking with UPR (T0-3B PLM)		
MSS	71.3	76.7
DPR	65.4	72.0
MSS-DPR	73.9	80.1
Contriever	76.0	81.6
BM25	79.3	83.9
BM25 + Contriever	**80.2**	**85.4**

Table 4: Top-{20, 100} retrieval accuracy on the Entity Questions dataset before and after re-ranking.

Retriever	BEIR nDCG@10	BEIR Recall@100
Baselines		
BERT (Devlin et al., 2019)	9.3	20.1
SimCSE (Gao et al., 2021)	27.4	48.1
REALM (Guu et al., 2020)	25.8	46.5
Contriever	36.0	60.1
BM25	41.6	63.6
After Re-ranking with UPR (T0-3B PLM)		
Contriever	44.6	66.3
BM25	**44.9**	**68.0**

Table 6: Macro-average nDCG@10 and Recall@100 scores on the BEIR benchmark. Performance numbers of the baseline models are from (Izacard et al., 2021).

Experiments: Passage Retrieval

Retriever

Dataset	#Q	#E	NDCG@10				Recall@100			
			BM25		Contriever		BM25		Contriever	
			original	re-ranked	original	re-ranked	original	re-ranked	original	re-ranked
Scifact	300	5K	66.5	70.3	64.9	69.6	90.8	94.2	92.6	94.3
Scidocs	1000	25K	15.8	17.0	14.9	17.3	35.6	39.0	36.0	39.0
Nfcorpus	323	3.5K	32.5	34.8	31.7	33.3	25.0	28.0	29.0	31.3
FIQA-2018	648	57K	23.6	44.4	24.5	45.0	53.9	67.7	56.2	72.8
Trec-covid	50	0.2M	65.5	68.8	27.4	60.4	49.8	54.8	17.2	36.7
Touche-2020	49	0.4M	36.8	20.6	19.3	21.3	53.8	45.7	22.5	42.4
NQ	3452	2.7M	32.9	45.4	25.4	44.2	76.0	87.7	77.1	88.4
MS-Marco	6980	8.8M	22.8	30.2	20.6	30.7	65.8	76.9	67.2	79.1
HotpotQA	7405	5.2M	60.3	73.3	48.1	72.2	74.0	82.5	70.4	80.8
ArguAna	1406	8.7K	31.5	37.2	37.9	50.3	94.2	98.2	90.1	97.5
CQADupStack	13145	0.5M	29.9	41.6	28.4	41.7	60.6	70.1	61.4	71.3
Quora	10000	0.5M	78.9	83.1	83.5	82.8	97.3	98.8	98.7	98.9
DBpedia	400	4.6M	31.3	35.4	29.2	33.8	39.8	53.3	45.3	47.8
Fever	6666	5.4M	75.3	59.1	68.2	57.3	93.1	84.2	93.6	83.1
Climate-Fever	1535	5.4M	21.3	11.7	15.5	9.5	43.6	39.2	44.1	31.3
Average			41.6	**44.9**	36.0	44.6	63.6	**68.0**	60.1	66.3

Table 7: UPR re-ranking results on the BEIR benchmark (Thakur et al., 2021). #Q and #E denotes the number of queries and evidence documents, respectively. Upon re-ranking the top-1000 documents with the T0-3B language model, on average, the performance of both BM25 and Contriever improve on the NDCG@10 and Recall@100 metrics. We also observe a drop in scores on some datasets which is highlighted in red.

Experiments: Impact of PLM

Re-ranker

- 4가지 PLM 비교

Retriever / Re-Ranker	NQ (dev) Top-1	Top-5	Top-20	Top-100
BM25	22.3	43.8	62.3	76.0
MSS	17.7	38.6	57.4	72.4
GPT-neo (2.7B)	27.2	55.0	73.9	84.2
T5 (3B)	22.0	50.5	71.4	84.0
T5-lm-adapt (3B)	29.7	59.9	76.9	85.6
T0-3B	**36.7**	**64.9**	**79.1**	**86.1**

Table 3: Comparison of different pre-trained language models (PLMs) as re-rankers on NQ dev. We re-rank the union of BM25 + MSS retrieved passages with UPR. Results demonstrate that the T0-3B PLM achieves the best top-K accuracy among the compared PLMs.

Experiments

Passage Candidate Size VS Latency

- **소요 시간** 대비 **향상되는 Retrieval Accuracy** 정체되는 경향

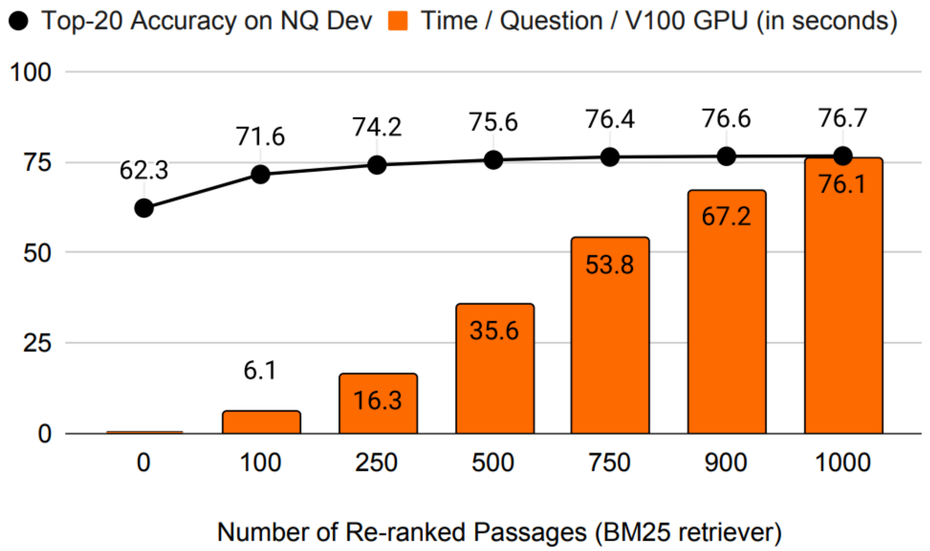


Figure 4: Effect of the number of passage candidates on top-20 accuracy and latency when re-ranked with T0-3B PLM. Evaluation is done on the NQ development set using BM25 retrieved passages.

Experiments: Question Answering

Reader

- reader: **Fusion-in-Decoder(FiD) model** (2021)
  - 구조 - T5 encoder 및 T5 decoder
  - Training - NLL loss, Adam, teacher-forcing, gradient clipping
  - Inference - Greedy decoding

Results

Model	top-K	SQuAD-Open dev	SQuAD-Open test	TriviaQA dev	TriviaQA test	NQ dev	NQ test	*# of params*
Baselines								
BM25 + BERT (Lee et al., 2019)	5	28.1	33.2	47.2	47.1	24.8	26.5	220M
ORQA (Lee et al., 2019)	5	26.5	20.0	45.1	45.0	31.3	33.3	330M
REALM (Guu et al., 2020)	5	-	-	-	-	38.2	40.4	330M
DPR (Karpukhin et al., 2020)	25	-	38.1	-	56.8	-	41.5	330M
RAG-Sequence (Lewis et al., 2020)	50	-	-	55.8	56.8	44.0	44.5	626M
Individual Top-K (large) (Sachan et al., 2021a)	-	-	-	-	59.6	-	48.1	1.2B
Joint Top-K (large) (Sachan et al., 2021a)	50	-	-	-	68.3	-	51.4	1.2B
FiD-base (Izacard and Grave, 2021b)	100	-	53.4	-	65.0	-	48.2	440M
FiD-large (Izacard and Grave, 2021b)	100	-	56.7	-	67.6	-	51.4	950M
FiD-KD-base (Izacard and Grave, 2021a)	100	-	-	68.6	68.8	48.0	49.6	440M
FiD-KD-large (Izacard and Grave, 2021a)	100	-	-	71.9	72.1	51.9	53.7	950M
EMDR^2-base (Sachan et al., 2021b)	50	46.8	51.1	71.1	71.4	50.4	52.5	440M
Our Implementation								
FiD-base (MSS retriever, T5 reader)	100	36.2	39.6	60.9	60.3	43.7	44.5	440M
—- Inference with UPR re-ranked passages		43.7	50.1	68.5	68.9	45.8	47.3	
FiD-base (DPR retriever, T5 reader)	100	48.8	45.8	67.9	68.5	49.4	50.8	440M
—- Inference with UPR re-ranked passages		51.5	54.0	70.1	71.2	49.8	51.3	
FiD-base (MSS-DPR retriever, T5 reader)	100	50.1	52.2	69.9	70.2	49.7	50.8	440M
—- Inference with UPR re-ranked passages		51.9	55.6	71.5	71.8	49.9	51.5	
FiD-large (MSS-DPR retriever, T5 reader)	100	51.9	54.4	71.5	71.6	**51.8**	53.6	950M
—- Inference with UPR re-ranked passages		**53.1**	**58.1**	**72.7**	**73.2**	51.5	**54.5**	

Table 5: Exact match scores for the open-domain QA task. We train one FiD model for each retriever as indicated and then perform inference with its respective re-ranked outputs. Top-K denotes the number of retrieved passages that are used by the reader to produce an answer. The best performing models are highlighted in bold.

감사합니다