간단 논세

Improving Passage Retrieval with Zero-Shot Question Generation

Sachan, Devendra Singh, et al. "Improving passage retrieval with zero-shot question generation." *arXiv preprint arXiv:2204.07496* (2022).

24.12.04
유하영

Introduction

- Text Retrieval은 핵심적인 하위 작업이다.
- Query-passage를 통해 효율적인 검색을 가능하게 하는 연구가 그동안 진행돼 옴.

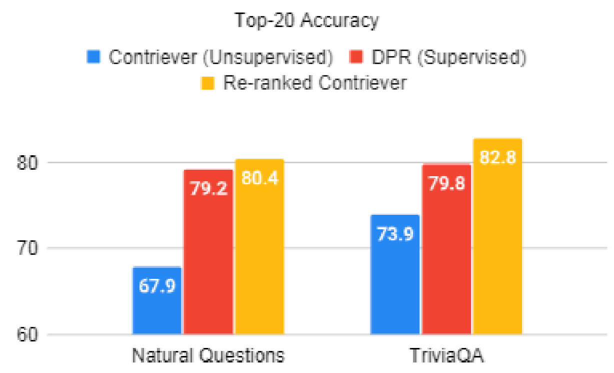


Figure 1: After UPR re-ranking of the Contriever's (unsupervised) (Izacard et al., 2022) top-1000 passages, we outperform strong supervised models like DPR (Karpukhin et al., 2020) on Natural Questions and TriviaQA datasets.

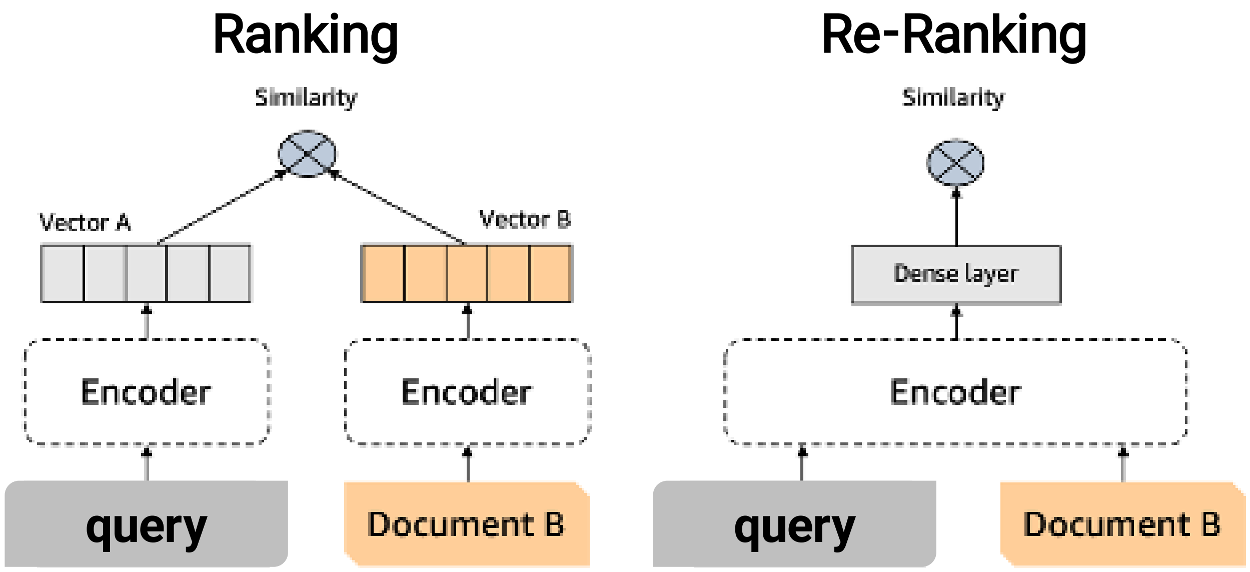
- Re-Ranked시 더 높은 성능을 보임
- re-ranked는 fine-tuning 없이도 여전히 강력한 성능향상을 보인다.

Introduction

Re-Ranking

: 초기에 생성된 문서나 항목의 리스트를 개선하기 위해 다시 정렬하는 과정

초기에 검색된 결과 passages를 받아, 최종적으로 가장 관련성 있는 문서를 최상단에 배치하는 것이 목표



(a) Bi-encoder

(b) Cross-encoder

Introduction

- Text Retrieval은 핵심적인 하위 작업이다.
- Query-passage를 통해 효율적인 검색을 가능하게 하는 연구가 그동안 진행돼 옴.

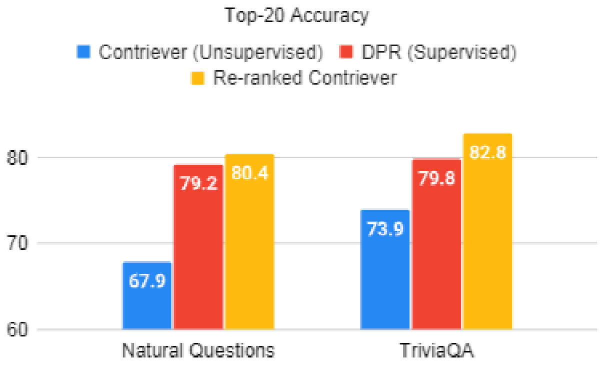


Figure 1: After UPR re-ranking of the Contriever's (unsupervised) (Izacard et al., 2022) top-1000 passages, we outperform strong supervised models like DPR (Karpukhin et al., 2020) on Natural Questions and TriviaQA datasets.

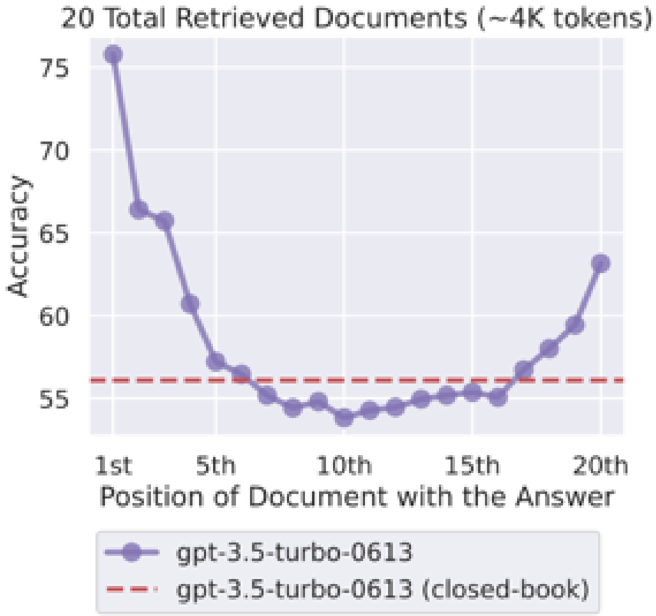
- Re-Ranked시 더 높은 성능을 보임
- re-ranked는 fine-tuning 없이도 여전히 강력한 성능향상을 보인다.

Retriever의 정확도 -> 관련 passage의 '존재 유무'가 아니라 **'순서'** 에 기반함을 확인

모델은 초기 편향 또는 최신 편향에 위치한 관련 정보를 더 잘 사용하며,
입력의 중간에 위치한 정보를 액세스하고 사용하는 경우 성능이 크게 저하.

즉, 관련 정보가 컨텍스트 내 **상위권에 위치하고 있을 때** 좋은 답변을 얻을 수 있음

Liu, Nelson F., et al. "Lost in the middle: How language models use long contexts." *Transactions of the Association for Computational Linguistics* 12 (2024): 157-173.



Introduction

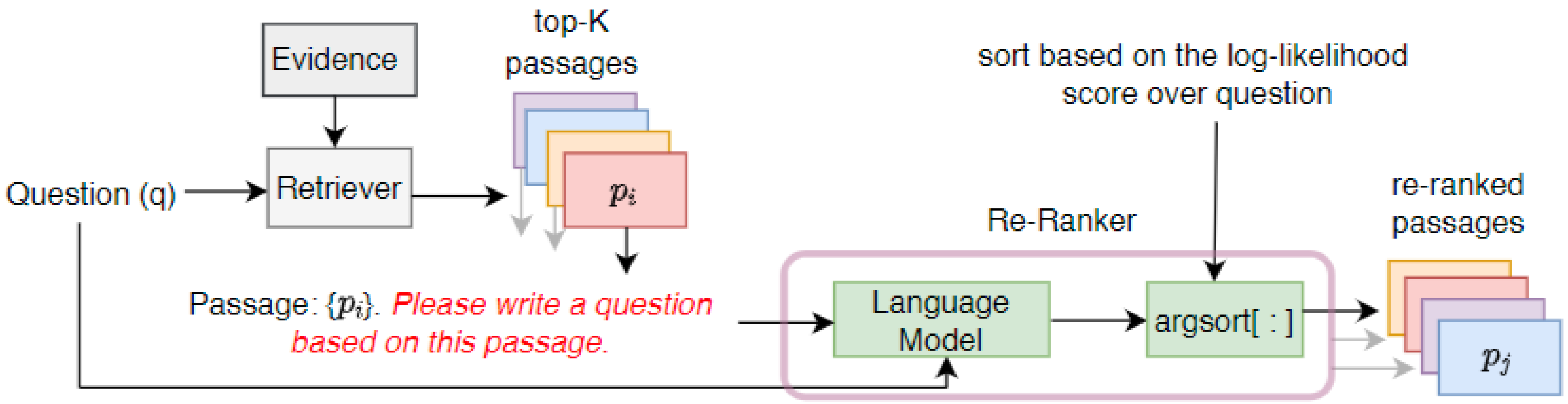
- Text Retrieval은 핵심적인 하위 작업이다.
- Query-passage를 통해 효율적인 검색을 가능하게 하는 연구가 그동안 진행돼 옴.

- 본 논문에서는, task-specific data나 tuning 없이 깊은 토큰 수준 분석을 통해 성능 향상을 이끌어낼 수 있는 방법인 Unsupervised Passage Re-ranker(UPR) 제안.
  - 최근 re-ranking 관련 연구에서는 q-p 쌍에 대해 PLM을 fine-tuning하여 사용
  - 반면, UPR은 간단하게 사용할 수 있는 **zero-shot PLM.**

  - 본 논문에서 제안하는 UPR은 Supervised Dense retriever를 능가함을 증명하는 첫 작업이다. 라고 주장
  - 일반적으로 zero-shot PLM만 사용하므로, 해당 방법이 다양한 검색 문제에 쉽게 적용할 수 있을 것이다.

Method

UPR (Unsupervised Passage Re-ranker)



Evidence
top-K passages
Question (q)
Retriever
p_i
sort based on the log-likelihood score over question
Re-Ranker
re-ranked passages
Passage: $\{p_i\}$. Please write a question based on this passage.
Language Model
argsort[:]
p_j

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

Method

$p(z_i|q)$

- 사용자가 query 제공
- Q와 가장 관련성 높은 passage Retriever
- Passage에서 query의 answer에 해당되는 passage 추출

  -> query-passage 간의 유사도를 바탕으로 ranking
- Answer 생성

** 새로운 방법 제안

-> passage에서 질문을 **generate**하는 방식

- query는 단일질문 -> 고정

$$p(z_i|q)=\frac{p(q|z_i)p(z_i)}{p(q)} \quad \rightarrow \quad \log p(z_i|q) = \ log\ p(q|z_i) + log\ p(z_i) - \ log\ p(q)$$

Bayes' Rule

Method

$p(z_i|q)=\frac{p(q|z_i)p(z_i)}{p(q)}$ -> $\log p(z_i|q) = log\ p(q|z_i) + log\ p(z_i) - log\ p(q)$

$log\ p(z_i|q) \propto log\ p(q|z_i)$

⬇

- Mean Likelihood function 적용

$q = \{q_1, q_2, \dots, q_T\}$: 한 개의 질문

$$p(q|z_i) = \prod_{t=1}^{T} p(q_t|q_{<t}, z_i)$$

$$log\ p(q|z_i) = \sum_{t=1}^{T} log\ p(q_t|q_{<t}, z_i)$$

$$= \frac{1}{|q|}\sum_{t=1}^{T} log\ p(q_t|q_{<t}, z_i\ ;\ \theta)$$

: 한 개의 질문에 대한 문장의 생성 확률

$C = -p(q)$

: 질문 q가 등장할 확률

q는 단일 질문으로 고정
= 상수로 취급

$\log\ p\,(z_i), \forall z_i \in \mathrm{Z}$

: 문서 z_i의 prior 확률

➔ "모든 문서가 동일한 prior 확률을 가진다" 가정
(passage와 query 간의 관계에 집중)

Method

UPR (Unsupervised Passage Re-ranker)

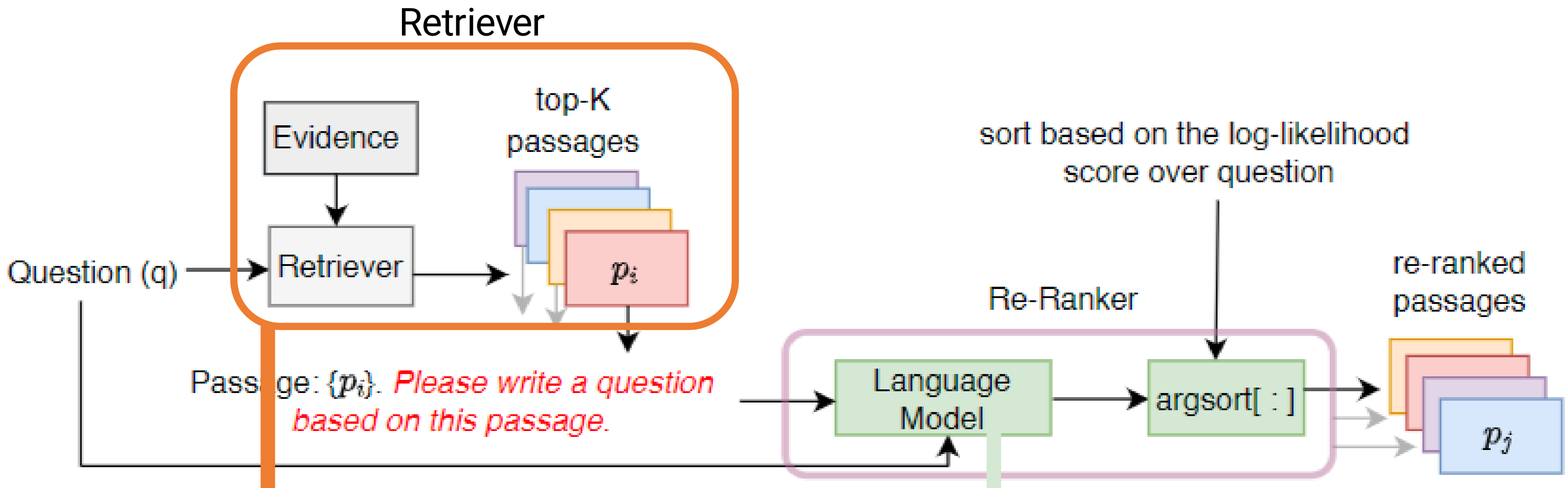

Retriever
Evidence
top-K passages
Question (q)
Retriever
p_i
Passage: $\{p_i\}$. Please write a question based on this passage.
sort based on the log-likelihood score over question
Re-Ranker
re-ranked passages
Language Model
argsort[:]
p_j

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

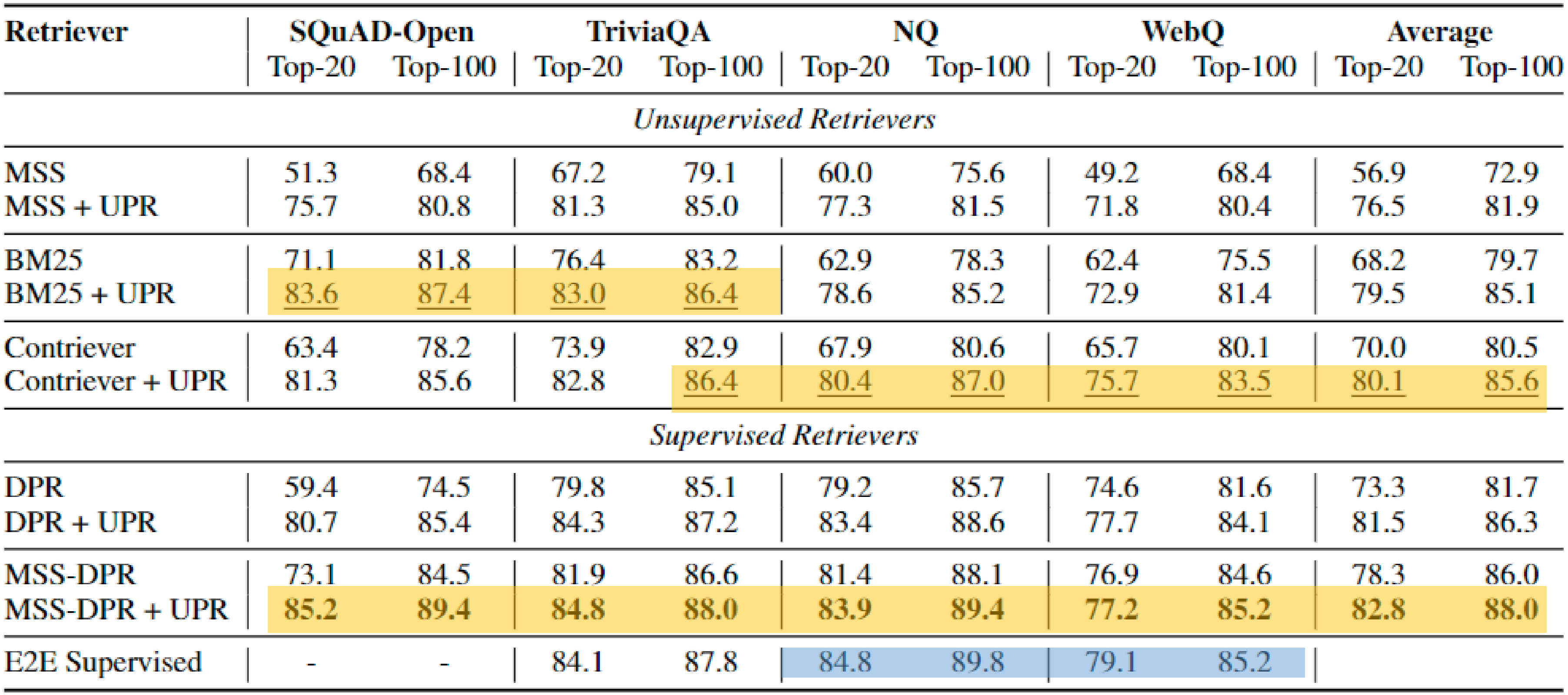
initial ranking
-> 상위 1000개의 passage

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

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

Experiments

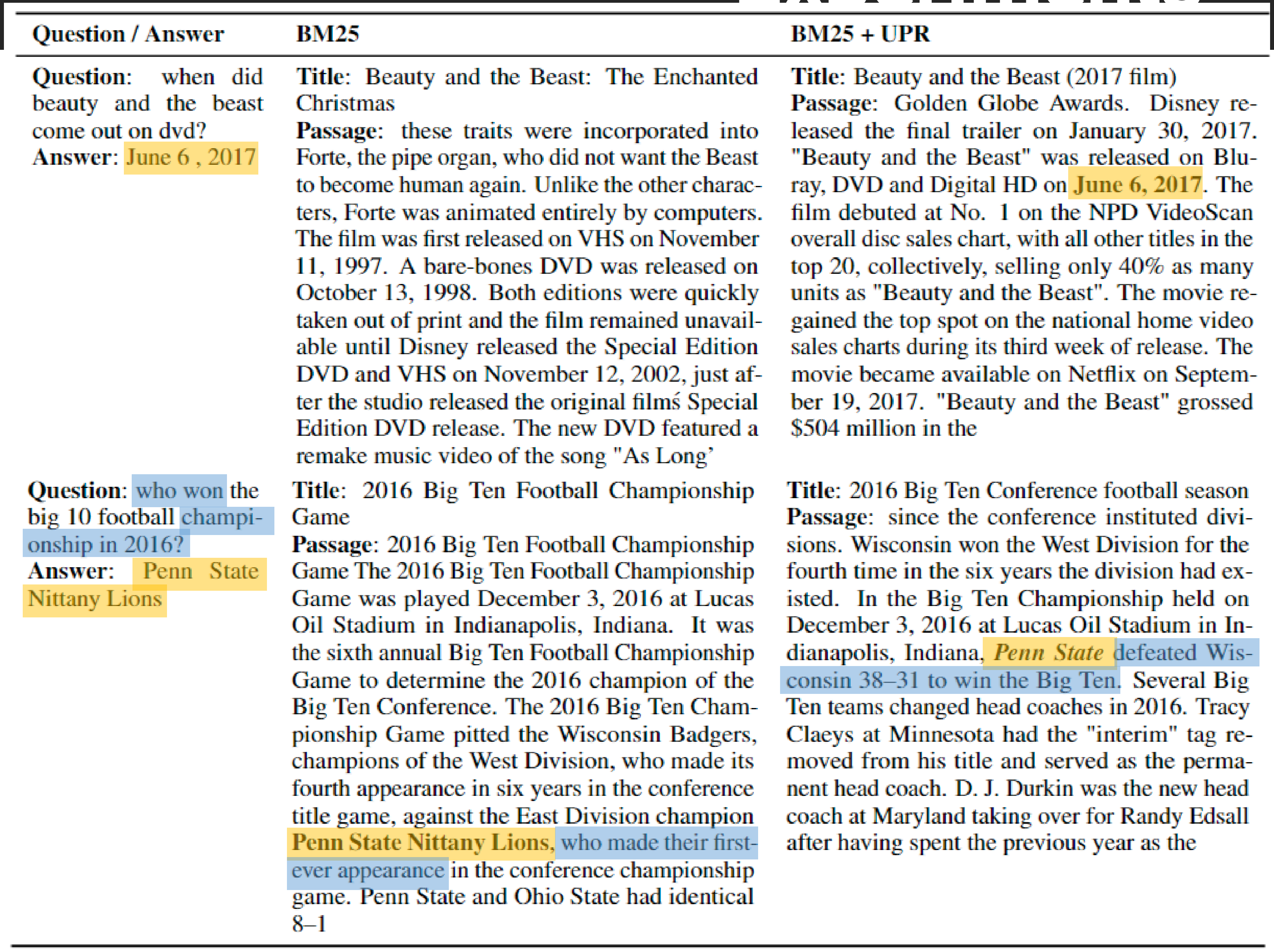
Top-{20, 100} retrieval accuracy on the test set of datasets
before and after UPR re-ranking of the **top-1000** retrieved passages with the T0-3B model.



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	65.7	80.1	70.0	80.5
Contriever + UPR	81.3	85.6	82.8	86.4	80.4	87.0	75.7	83.5	80.1	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	77.7	84.1	81.5	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		

- UPR가 Retriever 성능향상에 도움이 되었다.
- 하지만 여전히 E2E supervised가 가장 성능이 좋다

Experiments



Question / Answer	BM25	BM25 + UPR
Question: when did beauty and the beast come out on dvd? **Answer**: June 6 , 2017	**Title**: Beauty and the Beast: The Enchanted Christmas **Passage**: these traits were incorporated into Forte, the pipe organ, who did not want the Beast to become human again. Unlike the other characters, Forte was animated entirely by computers. The film was first released on VHS on November 11, 1997. A bare-bones DVD was released on October 13, 1998. Both editions were quickly taken out of print and the film remained unavailable until Disney released the Special Edition DVD and VHS on November 12, 2002, just after the studio released the original filmś Special Edition DVD release. The new DVD featured a remake music video of the song "As Long’	**Title**: Beauty and the Beast (2017 film) **Passage**: Golden Globe Awards. Disney released the final trailer on January 30, 2017. "Beauty and the Beast" was released on Blu-ray, DVD and Digital HD on **June 6, 2017**. The film debuted at No. 1 on the NPD VideoScan overall disc sales chart, with all other titles in the top 20, collectively, selling only 40% as many units as "Beauty and the Beast". The movie regained the top spot on the national home video sales charts during its third week of release. The movie became available on Netflix on September 19, 2017. "Beauty and the Beast" grossed $504 million in the
Question: who won the big 10 football championship in 2016? **Answer**: Penn State Nittany Lions	**Title**: 2016 Big Ten Football Championship Game **Passage**: 2016 Big Ten Football Championship Game The 2016 Big Ten Football Championship Game was played December 3, 2016 at Lucas Oil Stadium in Indianapolis, Indiana. It was the sixth annual Big Ten Football Championship Game to determine the 2016 champion of the Big Ten Conference. The 2016 Big Ten Championship Game pitted the Wisconsin Badgers, champions of the West Division, who made its fourth appearance in six years in the conference title game, against the East Division champion **Penn State Nittany Lions**, who made their first-ever appearance in the conference championship game. Penn State and Ohio State had identical 8–1	**Title**: 2016 Big Ten Conference football season **Passage**: since the conference instituted divisions. Wisconsin won the West Division for the fourth time in the six years the division had existed. In the Big Ten Championship held on December 3, 2016 at Lucas Oil Stadium in Indianapolis, Indiana, ***Penn State*** defeated Wisconsin 38–31 to win the Big Ten. Several Big Ten teams changed head coaches in 2016. Tracy Claeys at Minnesota had the "interim" tag removed from his title and served as the permanent head coach. D. J. Durkin was the new head coach at Maryland taking over for Randy Edsall after having spent the previous year as the

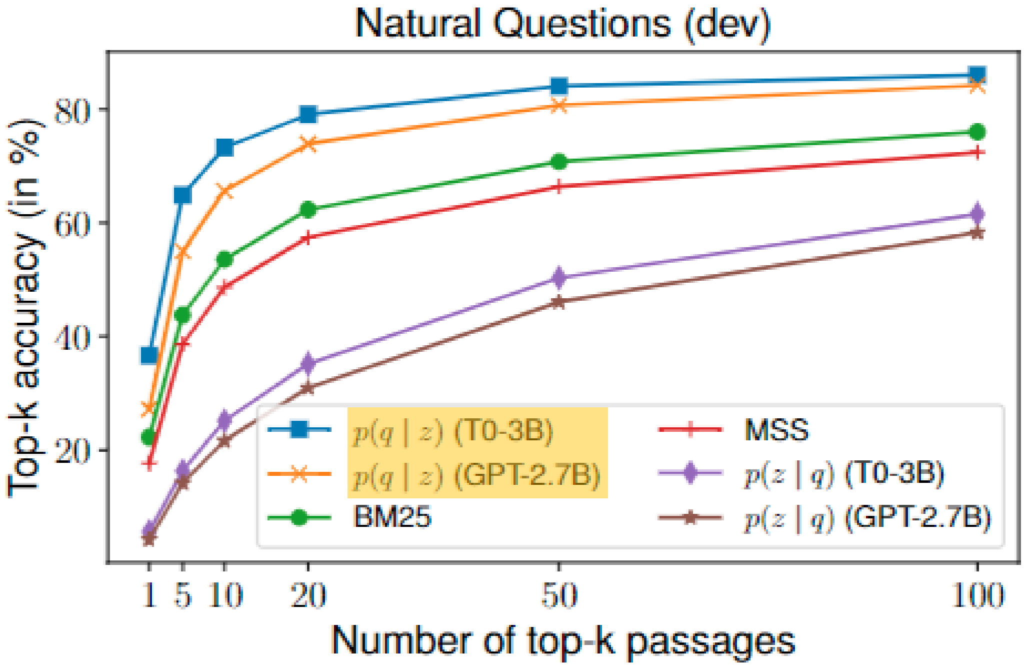
Table 9: Selected examples from the NQ development set of the top-1 retrieved passage from BM25 and the top passage obtained by UPR re-ranking of 1000 passages.

UPR은 cross-attention 메커니즘 덕에 질문과 구문 간의 관계를 더 잘 이해할 수 있다.

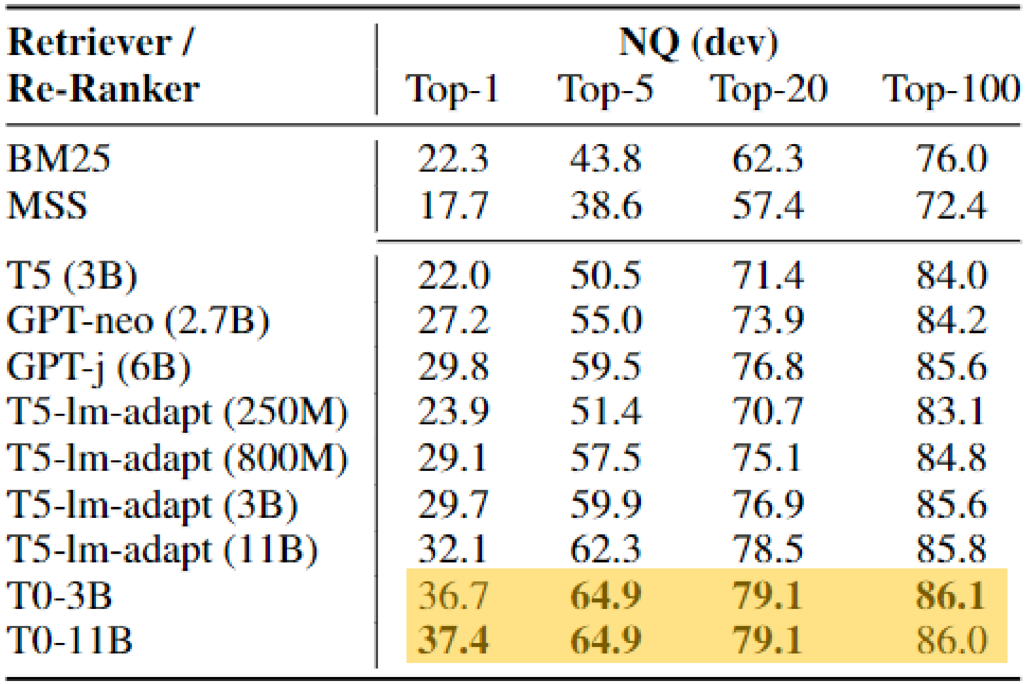
BM25가 검색한 구문이 정답을 그대로 포함하고 있지만, 잘못된 결론 도출

Experiments

4.2.1 Importance of Question Generation



4.2.2 Impact of Pre-trained LanguageModels

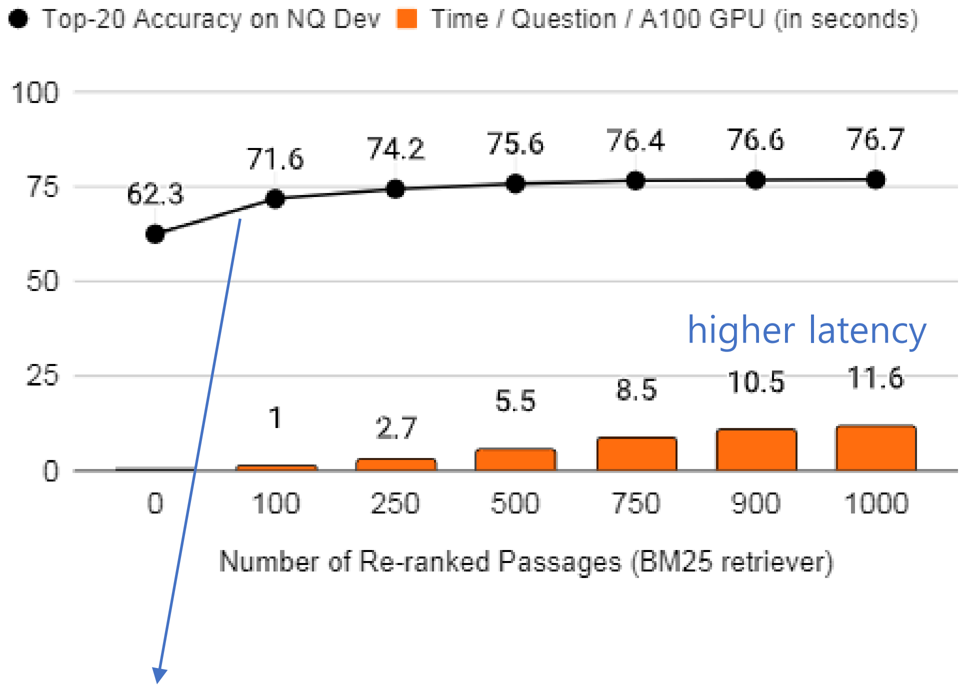


Retriever / Re-Ranker	NQ (dev)			
	Top-1	Top-5	Top-20	Top-100
BM25	22.3	43.8	62.3	76.0
MSS	17.7	38.6	57.4	72.4
T5 (3B)	22.0	50.5	71.4	84.0
GPT-neo (2.7B)	27.2	55.0	73.9	84.2
GPT-j (6B)	29.8	59.5	76.8	85.6
T5-lm-adapt (250M)	23.9	51.4	70.7	83.1
T5-lm-adapt (800M)	29.1	57.5	75.1	84.8
T5-lm-adapt (3B)	29.7	59.9	76.9	85.6
T5-lm-adapt (11B)	32.1	62.3	78.5	85.8
T0-3B	36.7	**64.9**	**79.1**	**86.1**
T0-11B	**37.4**	**64.9**	**79.1**	86.0

모든 PLM이 기준 Retriever보다
상당한 개선을 이뤘음을 확인할 수 있음

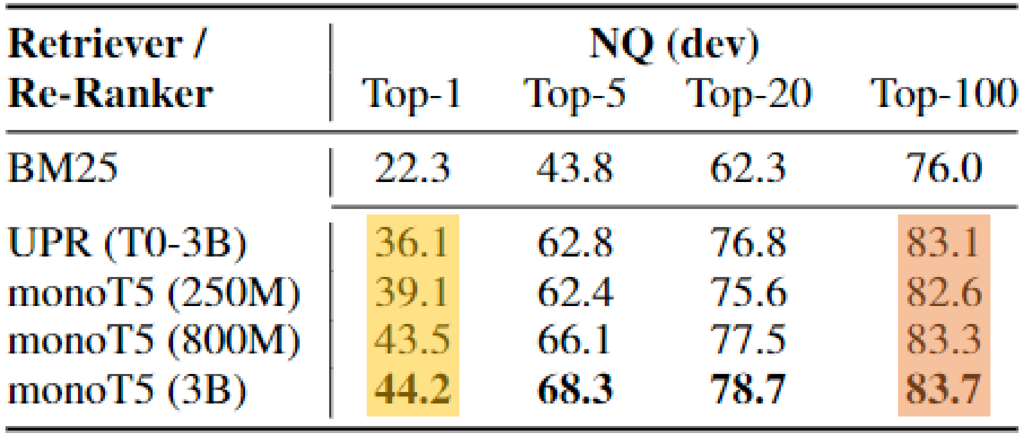
Experiments

4.2.3 Passage Candidate Size vs Latency



improve the performance
But the gains tend to plateau

4.3 Zero-Shot Supervised Transfer



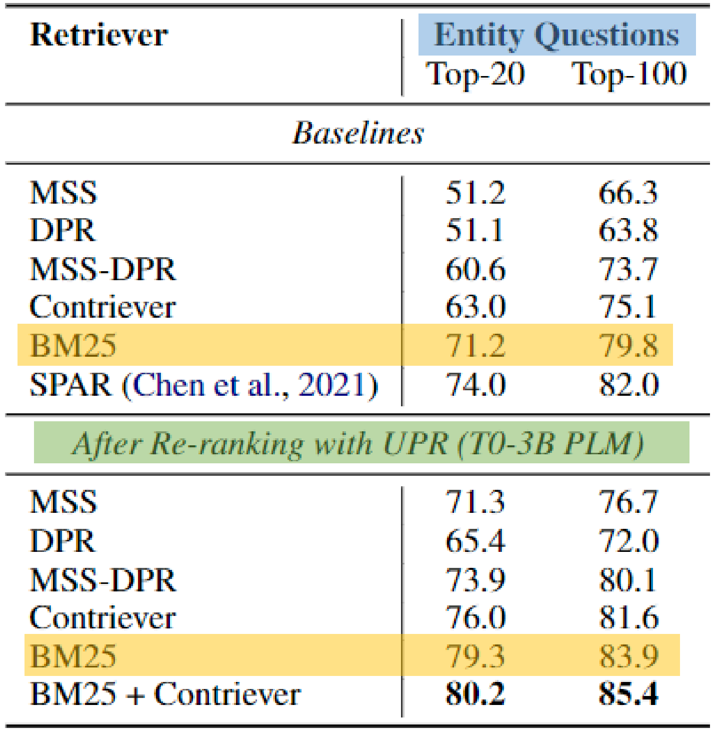
Retriever / Re-Ranker	**NQ (dev)** Top-1	Top-5	Top-20	Top-100
BM25	22.3	43.8	62.3	76.0
UPR (T0-3B)	36.1	62.8	76.8	83.1
monoT5 (250M)	39.1	62.4	75.6	82.6
monoT5 (800M)	43.5	66.1	77.5	83.3
monoT5 (3B)	**44.2**	**68.3**	**78.7**	**83.7**

Table 4: Zero-shot supervised transfer results on the NQ development set.

monoT5 : T5 PLM을 passage ranking dataset으로 fine-tuning된 모델

Experiments

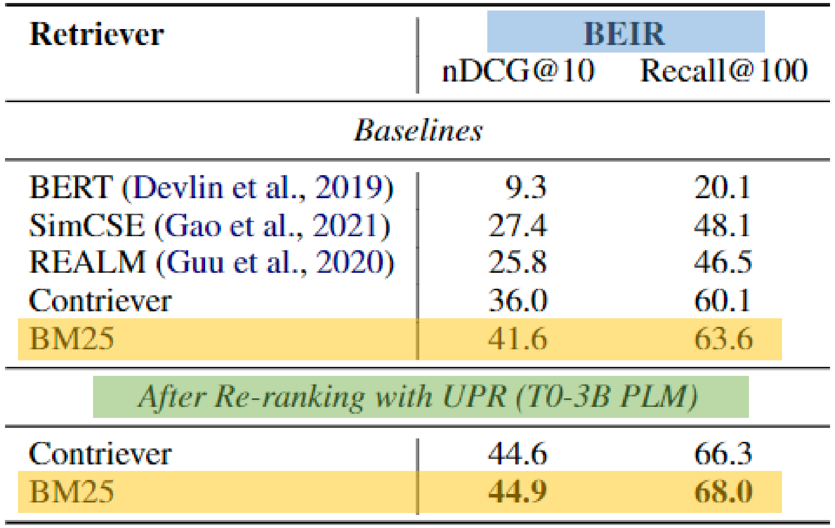
4.4 Evaluation on Keyword-centric Datasets



Retriever	Entity Questions	
	Top-20	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 5: Top-{20, 100} retrieval accuracy on the Entity Questions dataset before and after re-ranking. Following the original paper, we report macro-average scores.

위키백과에서 추출된 **사실 기반(named entity)** 질문
(질문이 짧고, 특정 키워드(예: 장소, 사람 이름 등)에 기반한 질문)



Retriever	BEIR	
	nDCG@10	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. (2022).

다양한 도메인(뉴스, 기술 문서 등)을 포함하는 **복잡한 데이터셋**

Discussion

Conclusions

- open-domain retrieval -> UPR(Unsupervised passage re-ranking) 제안
- **UPR** : retrieved passage에 대해 question generation을 조건으로 관련성 점수를 계산하여 re-ranking

[제안]

- zero-shot 접근, PLM 사용
  - 기존 연구) PLM에 대한 fine-tuning 필수적
  - 본 논문 제안) PLM 자체만 사용, re-ranking 변형만으로도 성능 향상
- 일반화 능력
  - 기존 연구) BM25 : sparse->keyword 강점, DPR : dense-> semantic similarity
    각 domain에 따라 잘 작동하지 않는 경우도.
  - 본 논문 제안) UPR을 제안하여, 성능향상

Limitations

1. re-ranking a large pool of passages can have a **high latency**
   ..involves performing **cross-attention**(4.2.3)
2. **Re-ranking의 성능이 1단계 검색(Initial Retrieval) 결과에 의존한다.**
   리트리버가 초기 검색 단계에서 정답 문서를 반환하지 못했다면, Re-ranking은 그 정답 문서를 처리할 수 없다.
3. 특정 도메인에 맞게 Fine-tuning된 PLM이 일반적인 PLM보다 해당 도메인에서 더 나은 성능을 발휘할 가능성이 높다.