REALM을 이용한 한국어 오픈도메인 질의 응답

Dong-Chan Kang[O1], Seung-Hoon Na[1], Yun-Su Choi[2], Hye-Woo Lee[2], Du-Seong Chang[2]

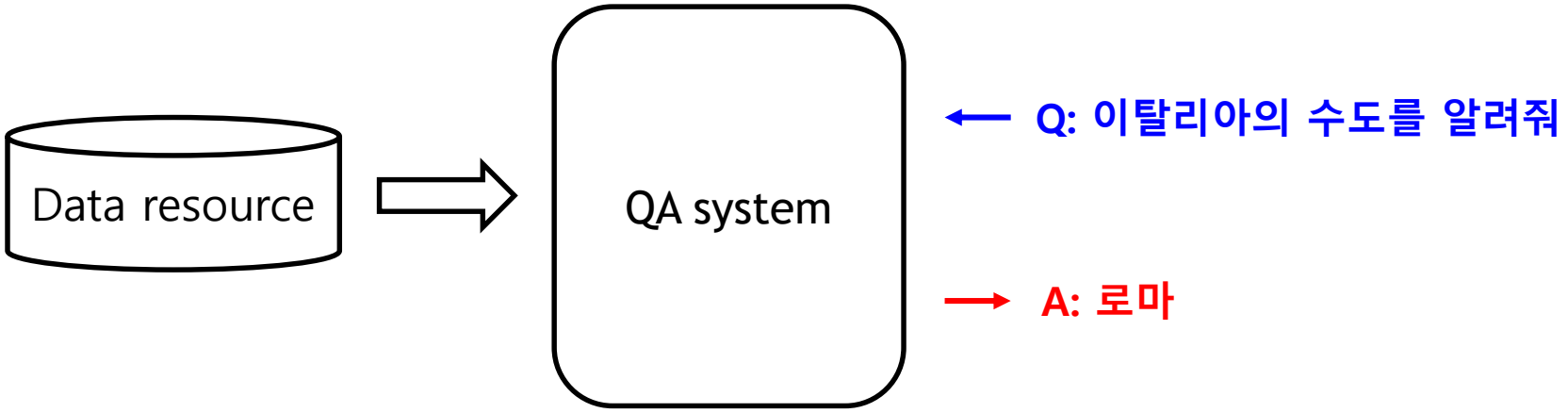
[1]Jeonbuk National University, [2]KT

목차

1. Background
2. Related works
3. Experiment setting and result
4. Conclusion

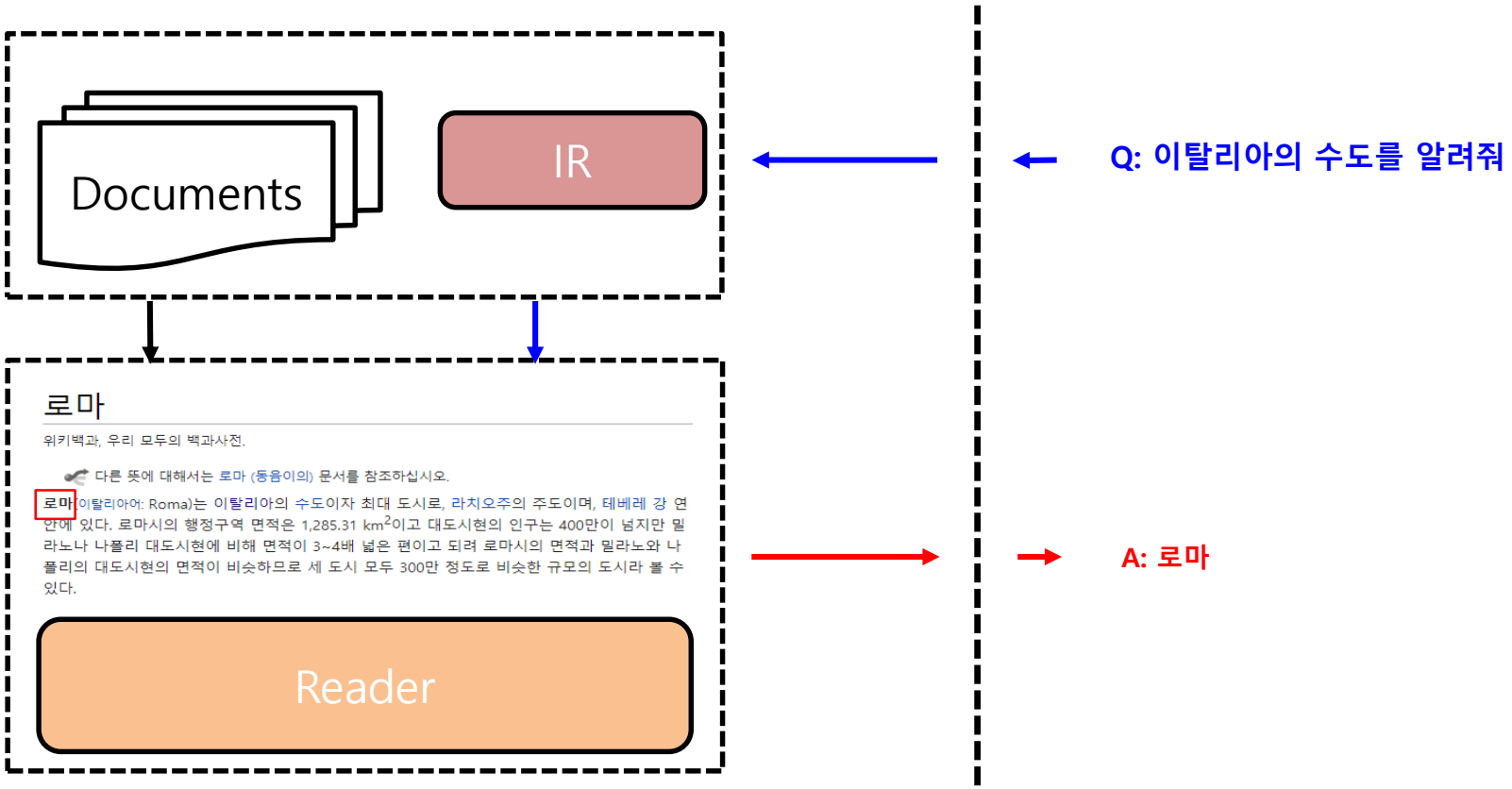
Background: Open-domain QA

- 오픈 도메인 질의 응답은 위키피디아나 웹과 같은 방대한 데이터 자원을 기반으로 질문자의 자연어 질문에 답변하는 것을 목표로 한다.
- 데이터 자원을 질의 응답 시스템에서 어떻게 활용할 것인가에 대해서는 다양한 방식이 존재한다. (e.g. KBQA, IRQA, Encoder-decoder ...)



Background: Information Retrieval QA

- 그 중 IRQA는 기계 독해 모델과, 검색 시스템을 결합한 형태

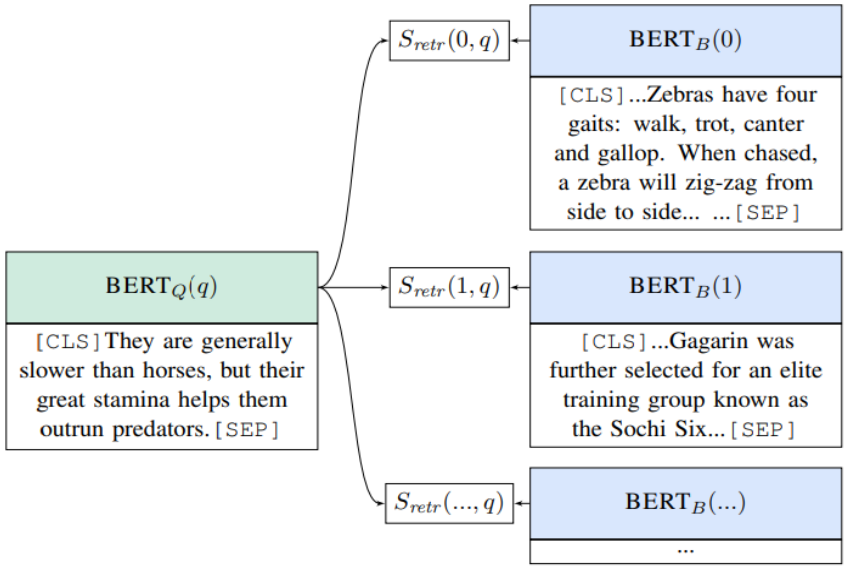


Background: Information Retrieval

- **IR system**
  - **키워드 기반 검색 모듈: BM25, TF-IDF**
    - 질문과 문서를 질문과 **문서에 나타난 단어에 기반**하여 질문과 문서 사이의 유사도를 계산하기 때문에, 동의어, 유의어의 처리가 어렵다.
    - 또한 적용될 도메인에 알맞게 튜닝하기 까다로움.
    - 이러한 한계에도 불구하고 다양한 방식으로 보완되어 최근까지 쓰여왔다.
  - **Dense retrieval**
    - 질문과 문서를 BERT와 같은 인코더를 통해 연속 벡터로 인코딩 한 후, 내적 혹은 코사인 유사도 기반으로 연관도를 계산함.
    - 최근의 연구결과, BM25를 뛰어넘는 검색 성능을 뛰어넘는 결과를 보이고 있음

Related Works: ORQA

- "Latent Retrieval for Weakly Supervised Open Domain Question Answering", Lee et al., ACL, 2019
- 문서에서 빠져 나온 문장을 쿼리로 하고, 문장이 빠져나온 문서를 정답으로 하는 방식의 사전학습 (ICT; Inverse Cloze Task)을 제시함
- 사전학습, 파인튜닝을 거쳐 질문과 정답 만을 가지고 기존의 IRQA의 성능을 개선하는 실험결과



$$h_q = W_q BERT_Q(q)[CLS]$$
$$h_b = W_b BERT_B(b)[CLS]$$
$$S_{retr}(b, q) = h_q^{\mathrm{T}} h_b$$

$$P_{ICT}(b|q) = \frac{\exp(S_{retr}(b, q))}{\sum_{b' \in BATCH} \exp(S_{retr}(b', q))}$$

Related Works: ICT

거미

위키백과, 우리 모두의 백과사전.

다른 뜻에 대해서는 거미 (동음이의) 문서를 참조하십시오.

q_i

거미(영어: spider)는 거미강 거미목의 절지동물이다. 여덟 개의 다리와 독을 주사할 수 있는 송곳니가 달린 집게발이 있으며 공기 호흡을 한다. 곤충은 머리 가슴 배(이하 두흉부)지만 거미는 머리와 배 부분으로 나뉜다. 또한 날개가 없어 날 수 없다. 대부분의 거미는 점액을 만드는 특수한 기관을 이용하여 거미줄을 만드는 정주성(定住性) 거미이다. 하지만 물거미, 게거미, 깡충거미, 농발거미처럼 거미줄을 만들지 않는 배회성(徘徊性) 거미도 있다.

거미

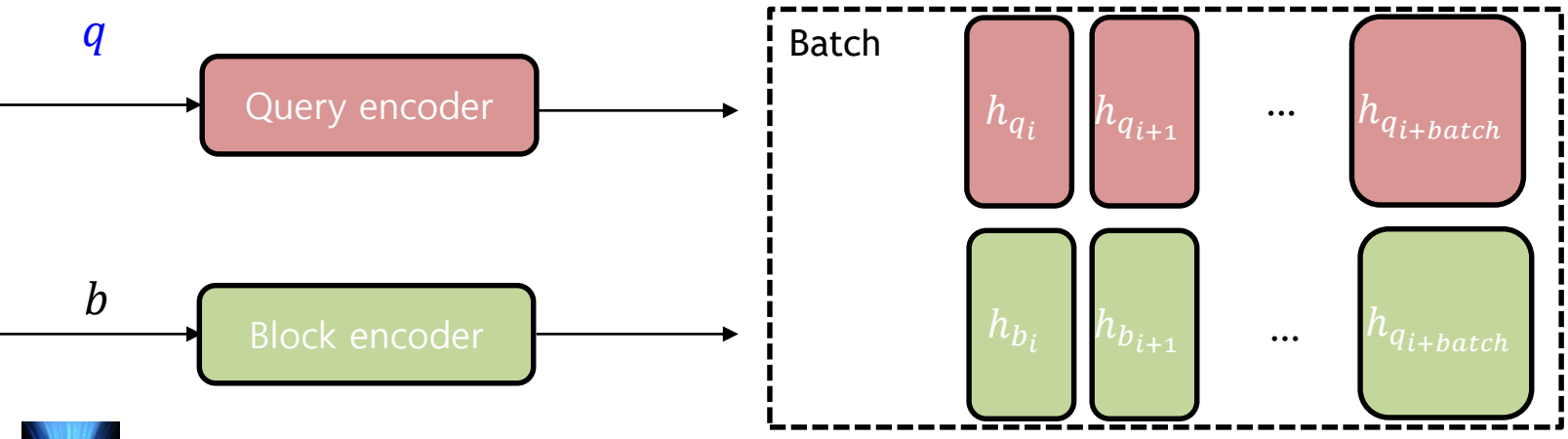
위키백과, 우리 모두의 백과사전.

다른 뜻에 대해서는 거미 (동음이의) 문서를 참조하십시오.

b_i

거미(영어: spider)는 거미강 거미목의 절지동물이다.
곤충은 머리 가슴 배(이하 두흉부)지만 거미는 머리와 배 부분으로 나뉜다. 또한 날개가 없어 날 수 없다. 대부분의 거미는 점액을 만드는 특수한 기관을 이용하여 거미줄을 만드는 정주성(定住性) 거미이다. 하지만 물거미, 게거미, 깡충거미, 농발거미처럼 거미줄을 만들지 않는 배회성(徘徊性) 거미도 있다.

$$P_{ICT}(b|q) = \frac{\exp(S_{retr}(b,q))}{\sum_{b' \in BATCH} \exp(S_{retr}(b',q))}$$



Related Works: ICT

Query encoder → MLM Pretrained BERT

Block encoder → MLM Pretrained BERT

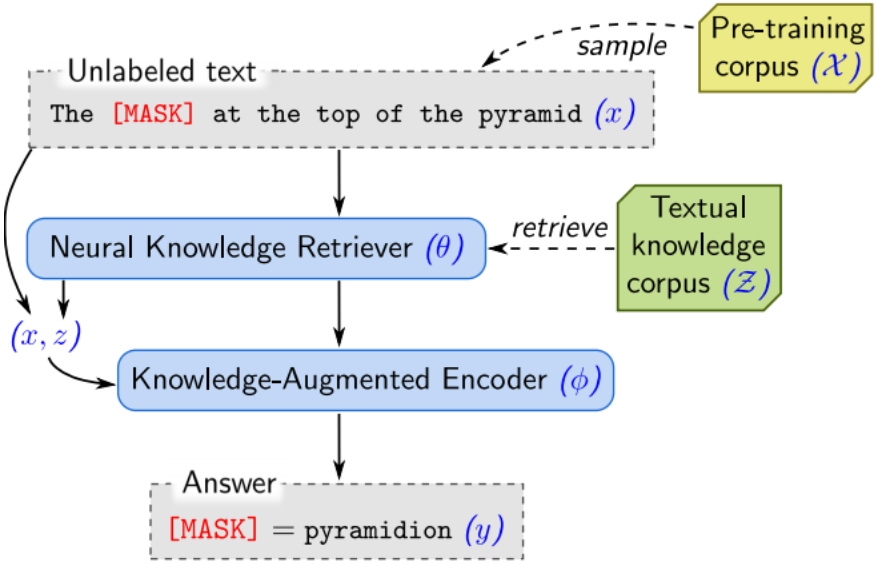
Ours

Query encoder → MLM Pretrained **DistilRoberta**

Block encoder → MLM Pretrained **DistilRoberta**

Related Works: REALM

- "REALM: Retrieval-Augmented Language Model Pre-Training", Guu et al., ICML, 2020
- ICT 사전 학습 된 모델부터 시작하여, 마스킹 된 문장을 쿼리로 하여, 가져온 문서를 엮어 MLM까지 더 사전학습
- ORQA와 동일한 파인튜닝 과정을 거쳐 IRQA의 성능과 검색성능을 더 개선하는 실험결과.



$$p(y|x) = \sum_{z \in Z} p(y|z,x)p(z|x)$$

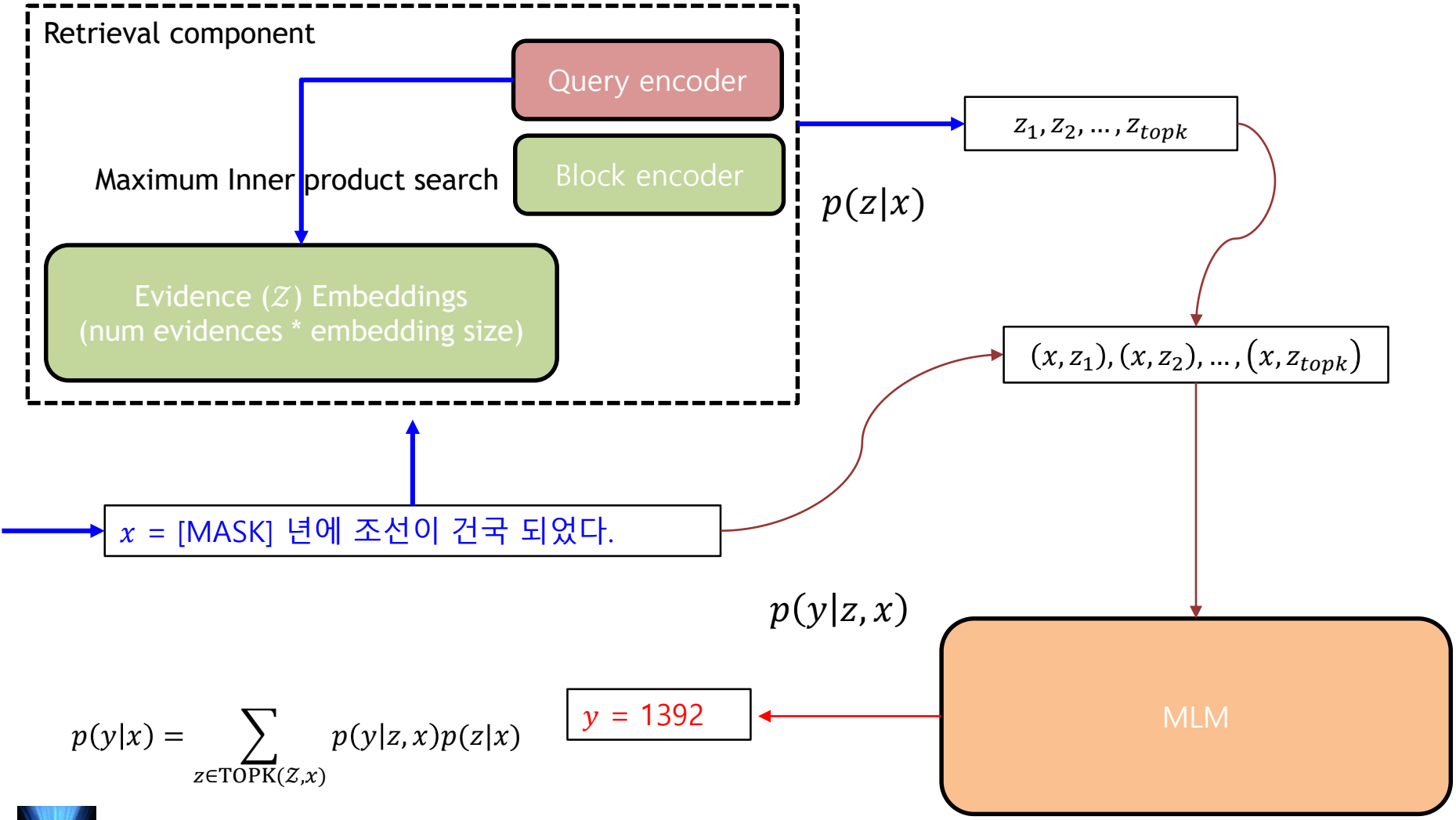
$where$:

$x = masked\ input$

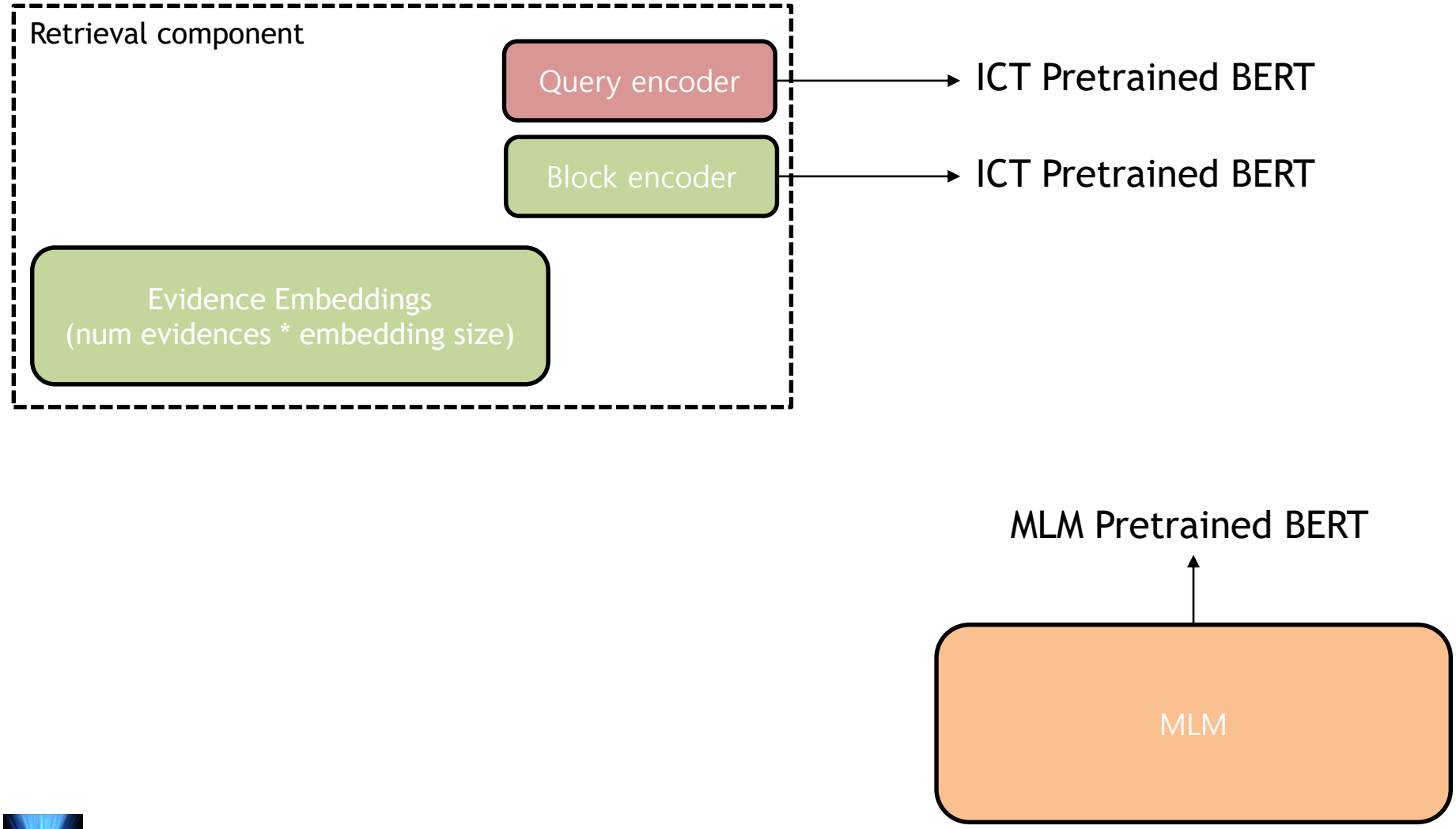
$y = replaced\ tokens$

$z = evidence$

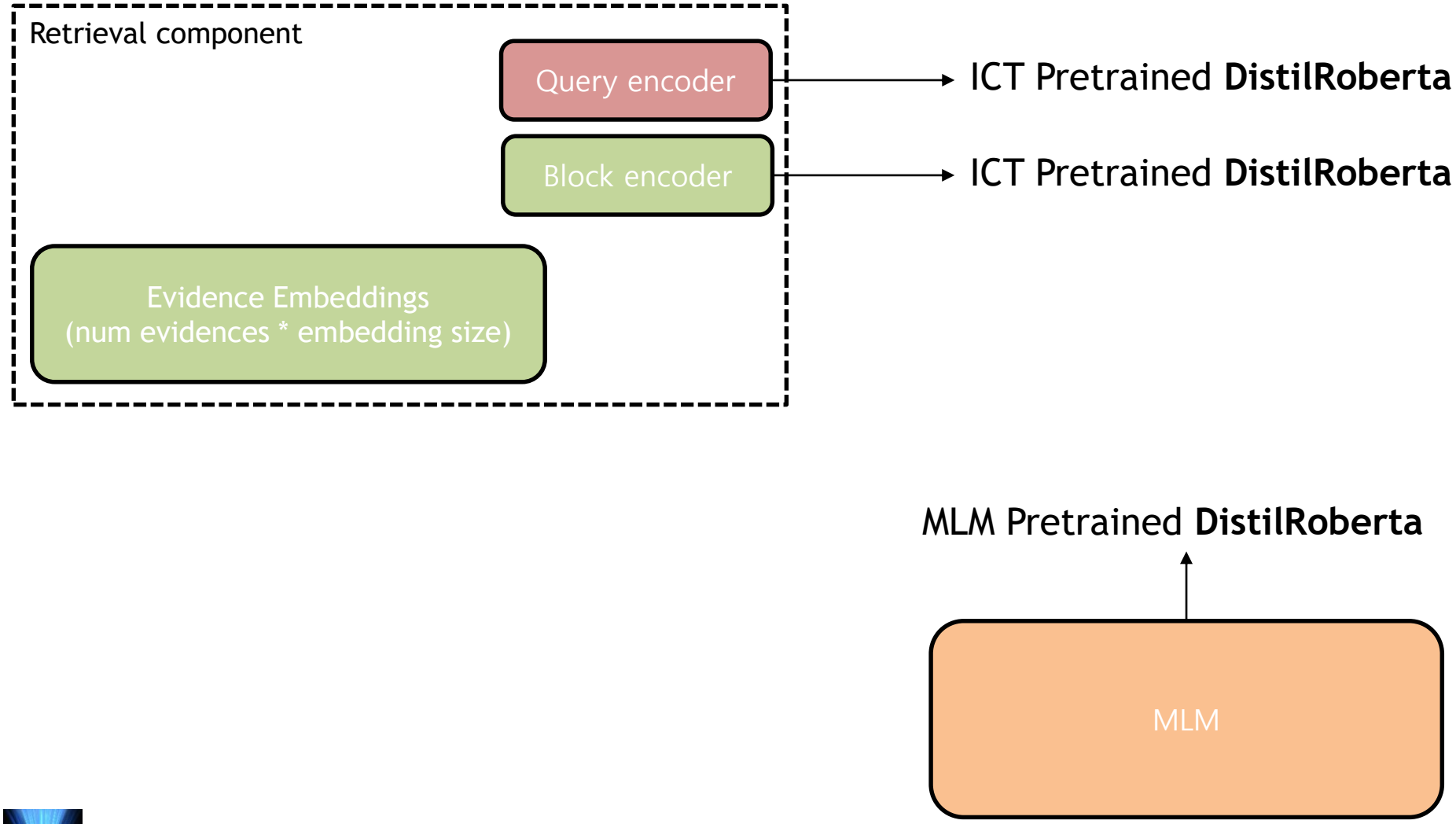
Related Works: REALM



Related Works: REALM



Related Works: REALM



Experiment Setting

ORQA

Query encoder	ICT Pretrained DistilRoberta
Block encoder	ICT Pretrained DistilRoberta
MLM	Pretrained DistilRoberta

REALM

Query encoder	REALM Pretrained DistilRoberta
Block encoder	REALM Pretrained DistilRoberta
MLM	REALM Pretrained DistilRoberta

Experiment Setting: Data

- 오픈 도메인 질의 응답 데이터 셋

  KT에서 제공한 질문과 정답 셋 페어로 구성된 오픈 도메인 질의 응답 셋

	질문 개수
Train	15900
Dev	900
Test	1800

- 문서 셋 (Evidence set)

  Roberta 토크나이저를 이용해 블록 단위로 나뉘어진 한국어 Wikipedia 문서들

	문서 개수	**블록 개수**
Wikipedia	437,321	**872,233**

Experiment Setting: Baseline

- **Baseline**
  - [1]에서 사용한 R3 모델
  - **IR system:** BM25를 이용한 검색 엔진
  - **Reader:** LSTM
  - **Evidence:** 한국어 위키피디아 + 한겨레 뉴스 + 네이버 포스트 / 지식 백과

- **Ours**
  - ORQA, REALM 모델
  - **IR system:** DistilRoberta 기반의 Dense Retrieval
  - **Reader:** DistilRoberta, REALM pretrained DistilRoberta
  - **Evidence:** 한국어 위키피디아

[1] 이영훈, 나승훈, 최윤수, 장두성, "NIL을 고려한 한국어 오픈 도메인 질의 응답", 한국정보과학회 학술 발표논문집, 2019

Experiment result: QA performance

Data	Model	Retriever	Reader	Evaluation script	ALL		Has Answer	
					EM	F1	EM	F1
Dev	R3[1]	BM25	LSTM	[1]	47.22	53.30	71.19	76.17
	ORQA	ORQA	DistilRoberta	[1]	52.55	58.09	78.96	82.24
				KorQuAD 1.0	51.33	63.71	77.12	86.39
	REALM	REALM	DistilRoberta (REALM pretrained)	[1]	**55.22**	**61.20**	**80.16**	**83.24**
				KorQuAD 1.0	**55.0**	**66.28**	**79.67**	**86.61**
Test	R3[1]	BM25	LSTM	[1]	44.39	51.51	66.53	72.78
	ORQA	ORQA (DistilRoberta)	DistilRoberta	[1]	48.80	55.15	77.20	80.49
				KorQuAD 1.0	47.91	61.59	75.79	85.15
	REALM	REALM (DistilRoberta)	DistilRoberta (REALM pretrained)	[1]	**51.25**	**57.38**	**77.46**	**80.92**
				KorQuAD 1.0	**50.80**	**63.61**	**76.78**	**85.06**

* [1]에서는 형태소 분석기에 기반하여 SQuAD 2.0과 동일한 평가 방식을 사용했기 때문에, 평가 스크립트에 따라 EM과 F1 성능의 차이가 있다.

Experiment result: Retrieval performance

	IR	Dev		Test	
		K=1	K=5	K=1	K=5
baseline	BM25	23.67	46.67	23.87	43.31
before finetuing (zero-shot)	ORQA	9.11	22.77	8.68	23.31
	REALM	**25.22**	**38.33**	**23.42**	**35.94**
after finetuing	ORQA	51.77	66.88	50.41	63.77
	REALM	**57.55**	**68.88**	**55.03**	**66.16**

Metric: Top-k accuracy

Experiment result: Retrieval performance

	IR	Dev		Test	
		K=1	K=5	K=1	K=5
baseline	BM25	23.67	20.02	23.87	19.89
before finetuing (zero-shot)	ORQA	9.11	10.28	8.68	10.50
	REALM	**25.22**	**23.44**	**23.42**	**22.68**
after finetuing	ORQA	51.77	46.75	50.41	45.31
	REALM	**57.55**	**51.17**	**55.03**	**49.01**

Metric: Top-k precision

Conclusion

- **결론**
- 파인 튜닝 된 Dense Retrieval은 전통적인 키워드 기반 IR 시스템인 BM25보다 더 나은 검색성능을 보였으며, 이를 기반으로 더 작은 문서 셋으로도 이전 실험결과보다 개선된 QA 성능과 검색 성능을 이끌어 낼 수 있었다.

- **향후 연구**
- 추후에는 훈련된 REALM 모델을 이용해 다양한 자연어 처리 태스크에 적용해 볼 예정이다.

Q/A