# Latent Retrieval for Weakly Supervised Open Domain Question Answering

## (ORQA paper review)

2조 - 김현수

# 목차

# 목차

# Question Answering (QA)

- Closed-domain Question Answering(CDQA)
  - Only from one domain (legal, medical, engineering … )
- Open-domain Question Answering(ODQA)
  - Answering a question from any domain


*Q: Where does the energy in a nuclear explosion come from?*

*A: high-speed nuclear reaction*

*Q: Where is Einstein's house?*

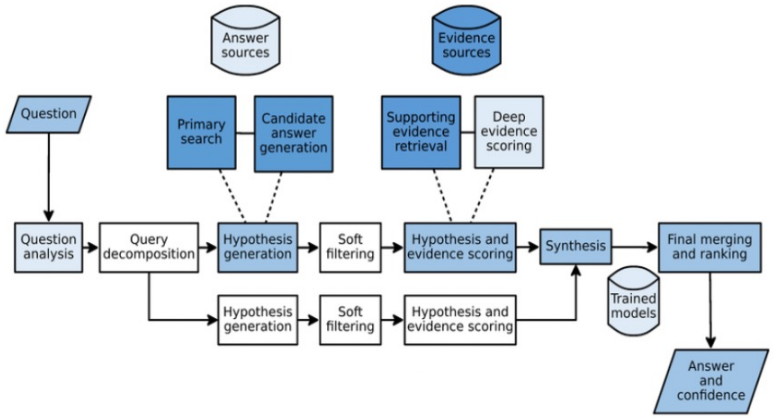
*A: 112 Mercer St, Princeton, NJ*

*Q: How many papers were accepted by ACL 2020?*

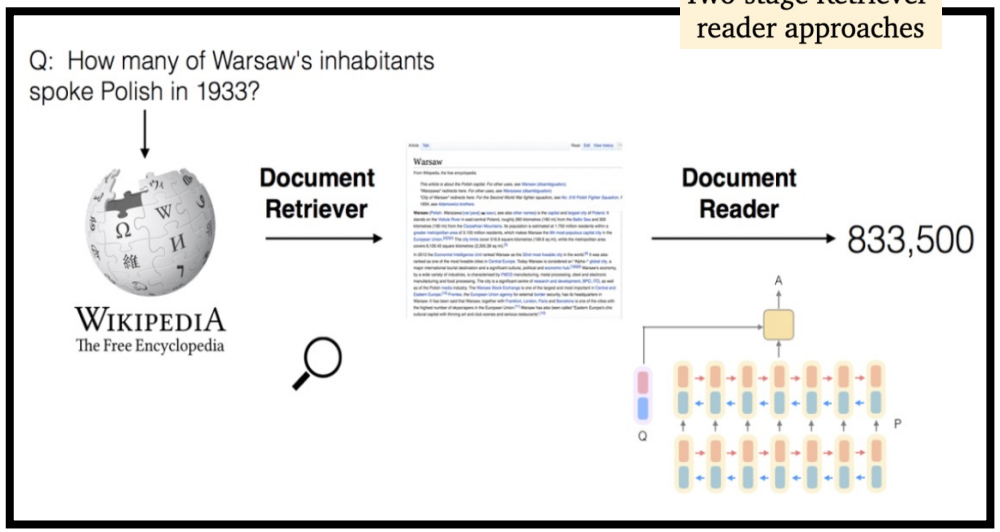
*A: 779 papers*

# Question Answering의 발전 과정

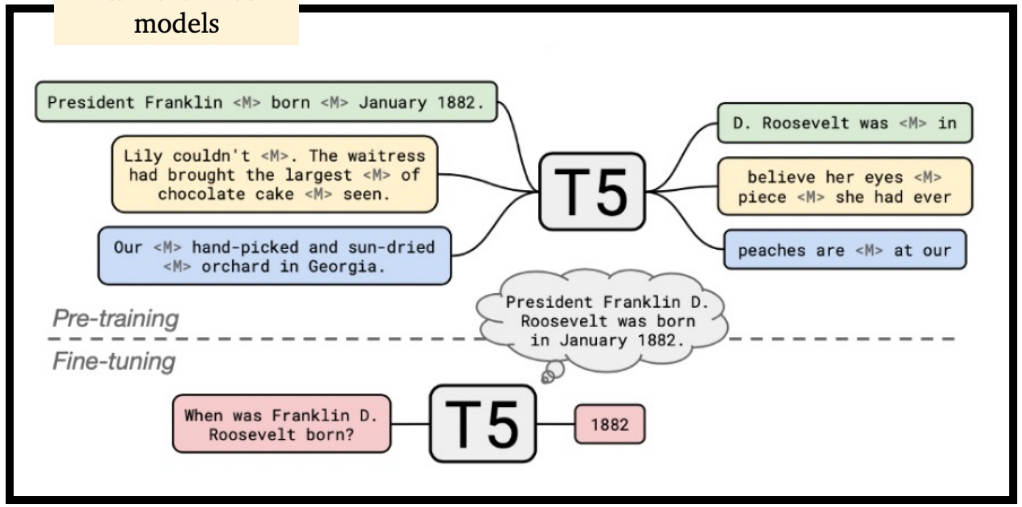
Classical QA pipeline



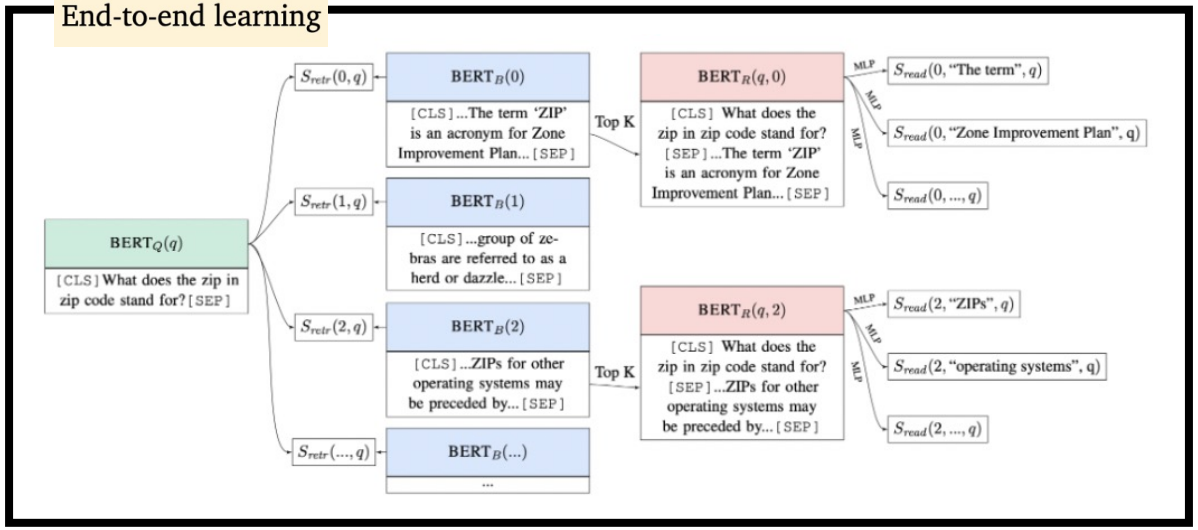

Two-stage Retriever-reader approaches




Retrieval-free models

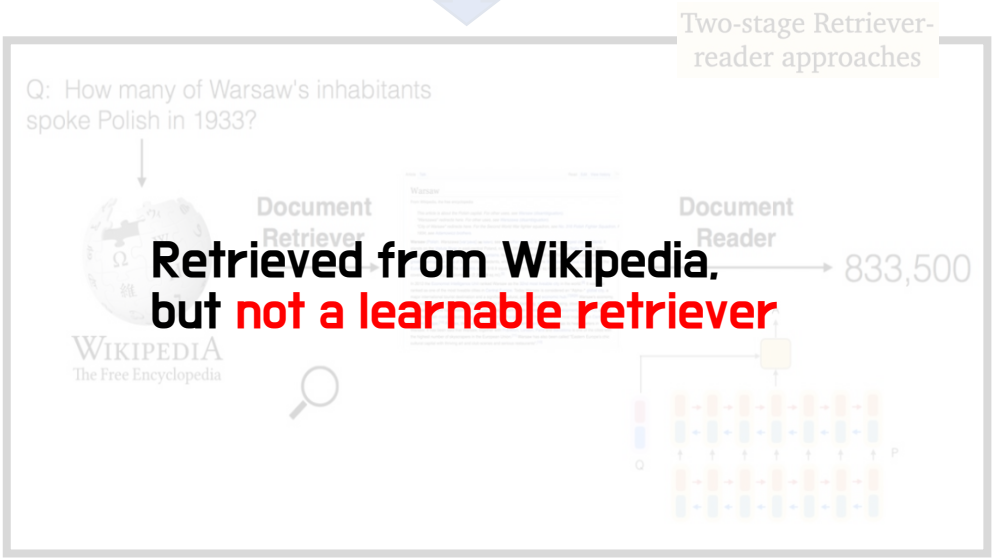



End-to-end learning



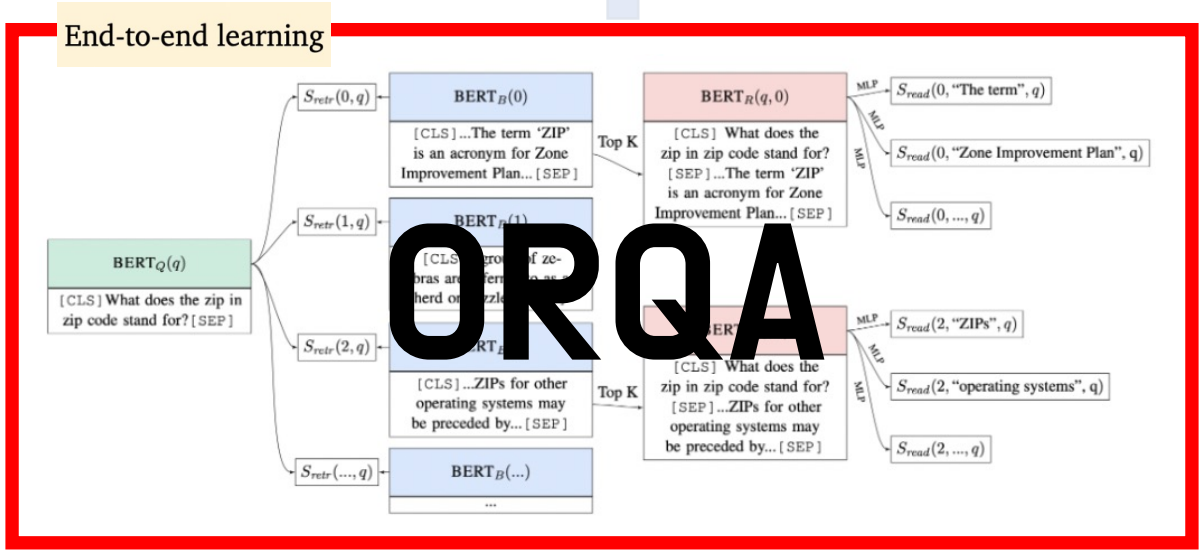
# Question Answering의 발전 과정




**Retrieved from Wikipedia, but not a learnable retriever**



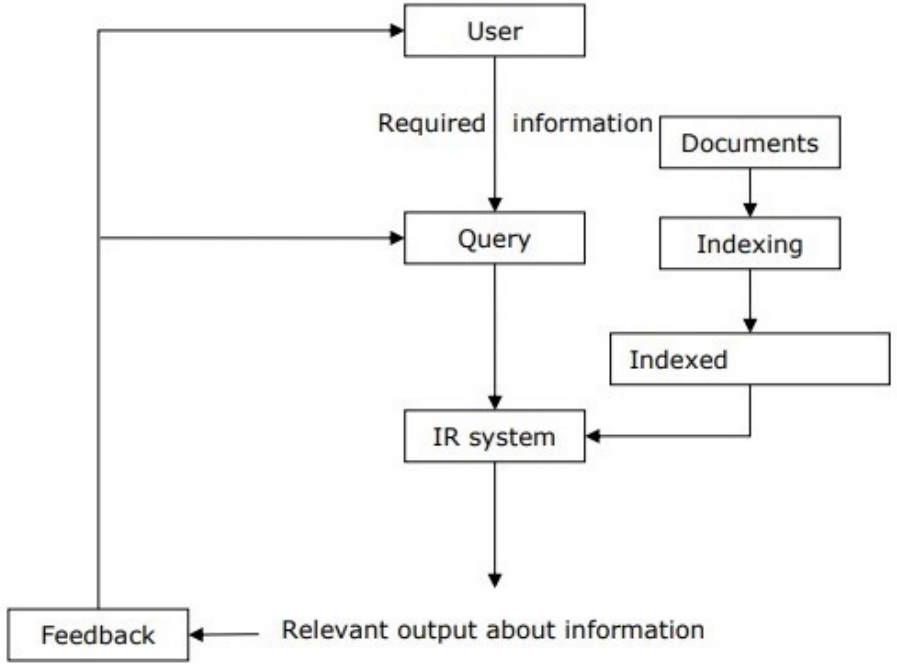




**ORQA**

기존의 연구들 .. (with IR system)

# Information retrieval (정보 탐색)

- 사용자가 입력한 키워드에 대해서 적절한 문서를 찾는 것
- 어휘 및 의미 일치와 관련(matching)
- QA 시스템과 비슷(하나 조금 다름)

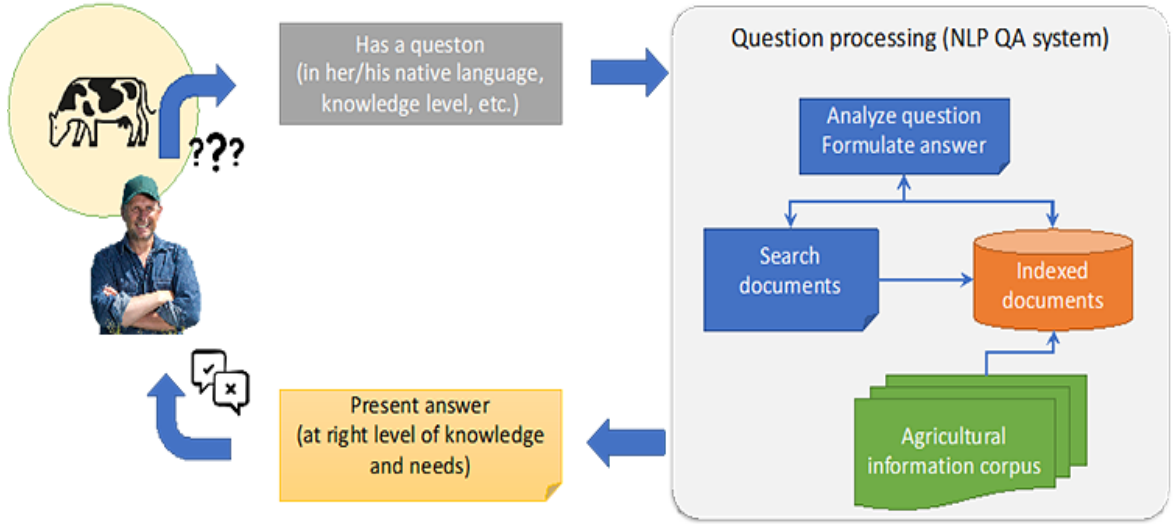



[Information retrieval pipeline]

ORQA !!

# Question Answering

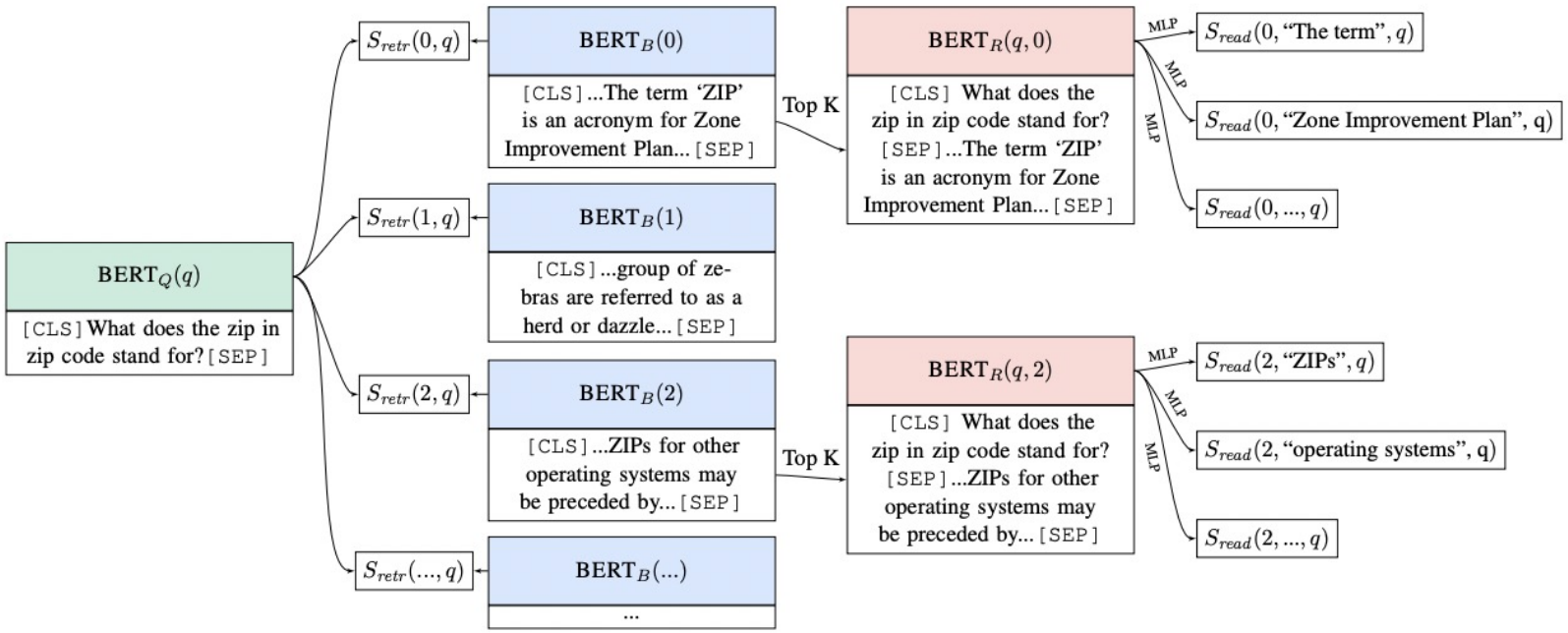
- 찾은 문서를 가져올 뿐만 아니라 문서에서 정답 부분을 반환
- 알려지지 않은 정보를 찾음(IR은 기존에 알고 있는 정보를 찾음)
- IR에 비해 더 많은 자연어 이해를 필요로 함




[Question answering pipeline]

# ORQA's contribution

- Both retriever and reader are learnable with NNs ( = BERT)
- Only learned from question-answering pairs: No reading comprehension datasets!
- A new pre-training task called Inverse Cloze Task (ICT) to address the challenging retrieval problem.



# 목차

# 3. Overview - Task (Open-Domain Question Answering)

| Task | Training | | Evaluation | | Example |
|---|---|---|---|---|---|
| | Evidence | Answer | Evidence | Answer | |
| Reading Comprehension | given | span | given | string | SQuAD (Rajpurkar et al., 2016) |
| Open-domain QA | | | | | |
| Unsupervised QA | none | none | none | string | GPT-2 (Radford et al., 2019) |
| Strongly Supervised QA | given | span | heuristic | string | DrQA (Chen et al., 2017) |
| Weakly Supervised QA | | | | | |
| Closed Retrieval QA | heuristic | string | heuristic | string | TriviaQA (Joshi et al., 2017) |
| **Open Retrieval QA** | **learned** | string | **learned** | string | **ORQA (this work)** |

Table 1: Comparison of assumptions made by related tasks, along with references to examples. Heuristic evidence refers to the typical strategy of considering only a closed set of evidence documents from a traditional IR system, which sets a strict upper-bound on task performance. In this work (ORQA), only question-answer string pairs are observed during training, and evidence retrieval is learned in a completely end-to-end manner.

**Input : question (string)**

**Output : answer (string)**

**Evidence : model로 부터 학습 (DrQA : given)** * DrQA : a first neural open-domain QA system, 2017

**Evaluation : 간단한 정규화 후 exact matching (DrQA : lowercasing)**

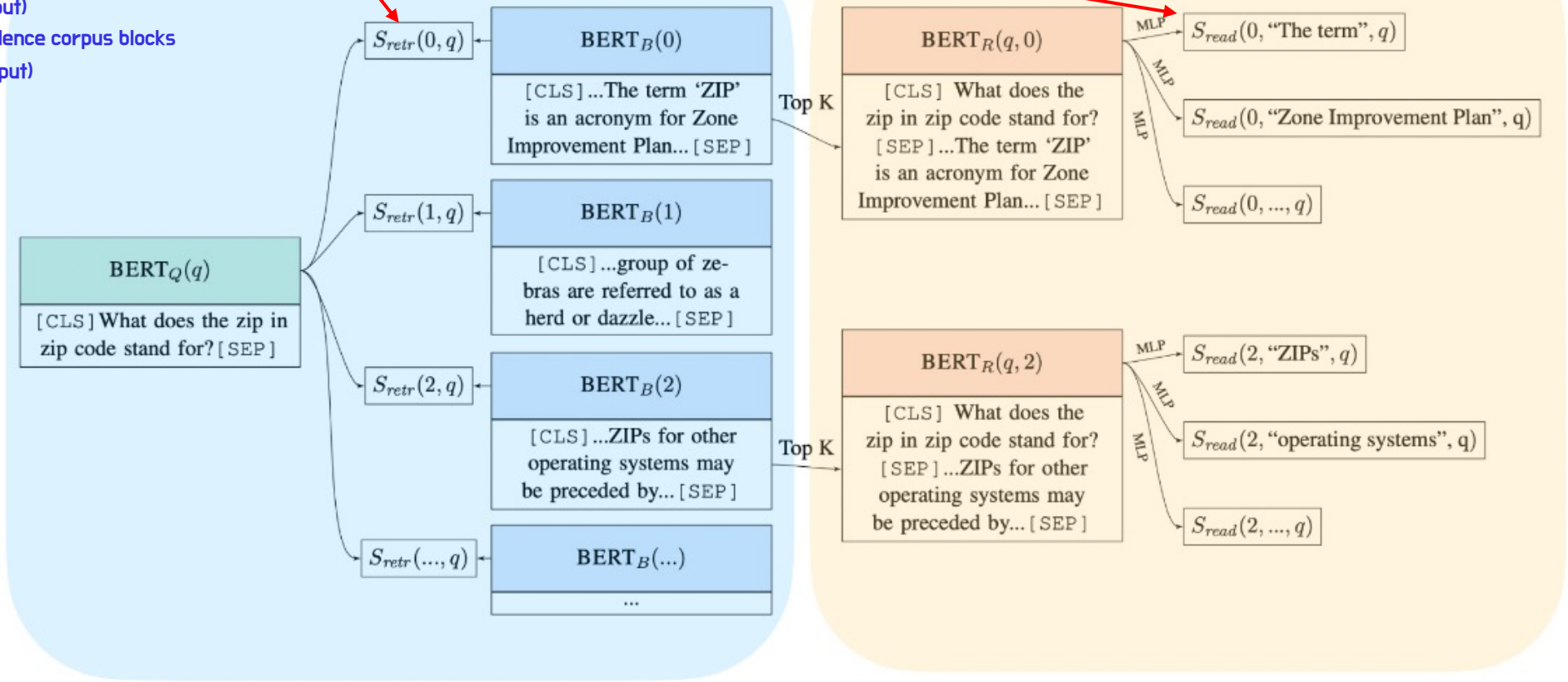
# 3. Overview - Formal Definitions & Existing Pipelined Models

**Scoring function** : $S(b, s, q) = S_{retr}(b, q) + S_{read}(b, s, q)$

**Inference output** : $a^* = \text{TEXT}(\underset{b,s}{\text{argmax}}\, S(b, s, q))$

= score를 최대화하는 b번째 인덱스 문서 내 텍스트 s

- $b$ : index of an evidence block ($1 \le b \le B$)
- $s$ : span of text within block b
- $q$ : question string (input)
- $B$ : the number of evidence corpus blocks
- $a$ : answer string (output)




Information Retrieval

**Retriever**

Reading Comprehension

**Reader**

# 목차

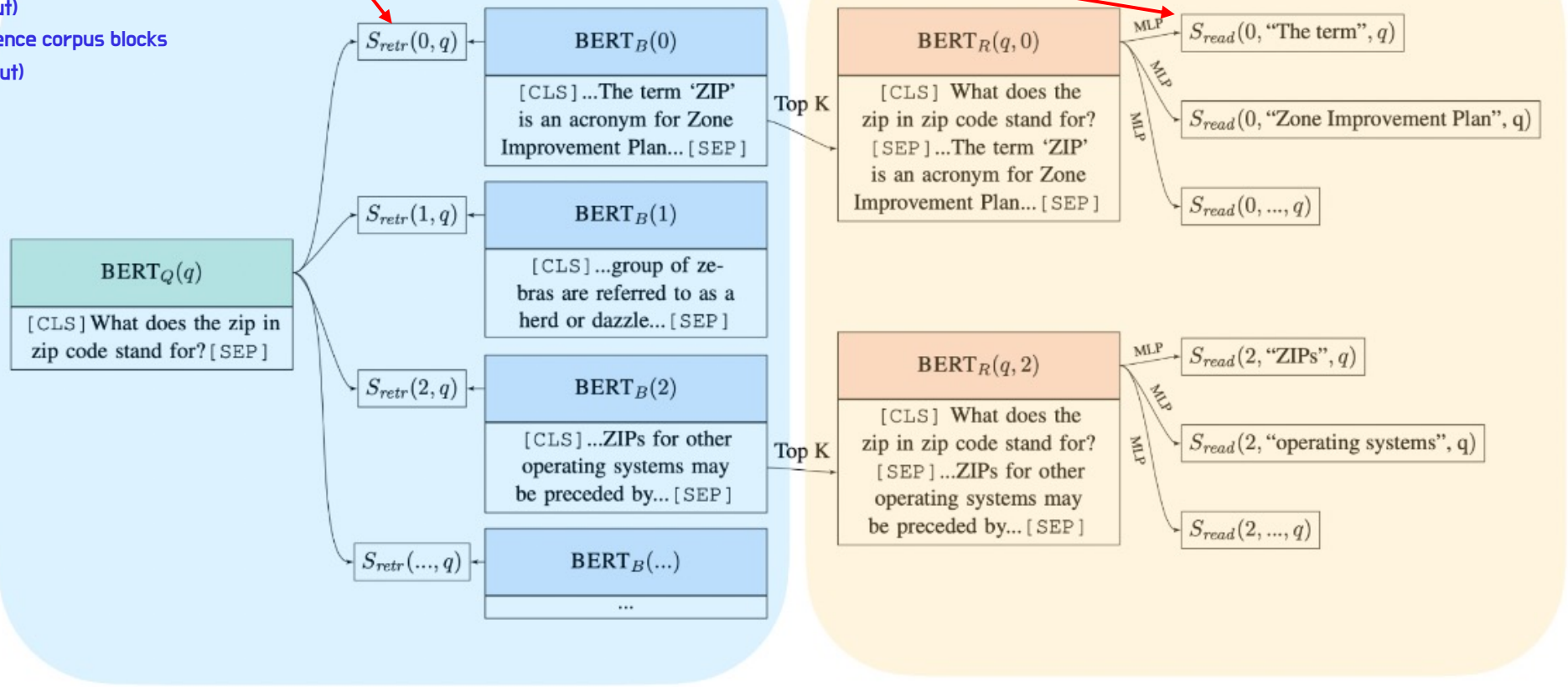
# 4. ORQA : Open-Retrival Question Answering

**Scoring function** : $S(b, s, q) = S_{retr}(b, q) + S_{read}(b, s, q)$

**Inference output** : $a^* = \text{TEXT}(\underset{b,s}{\text{argmax}}\, S(b, s, q))$

= score를 최대화하는 b번째 인덱스 문서 내 텍스트 s

- $b$ : index of an evidence block ($1 \le b \le B$)
- $s$ : span of text within block b
- $q$ : question string (input)
- $B$ : the number of evidence corpus blocks
- $a$ : answer string (output)




Information Retrieval

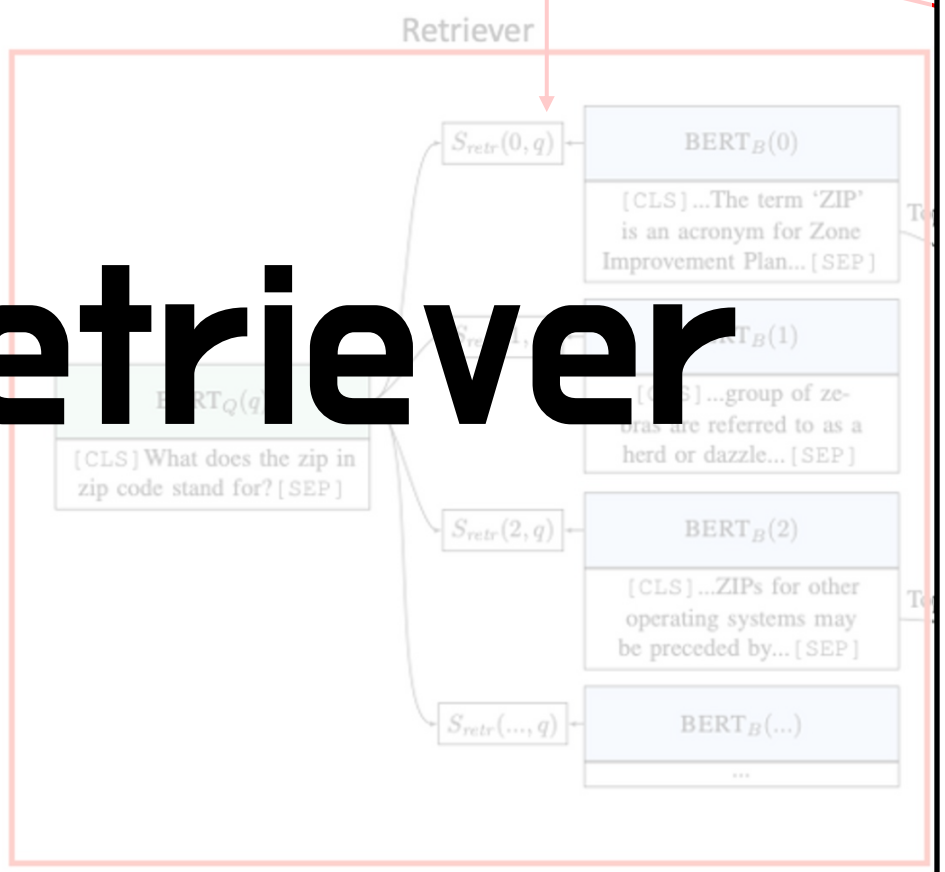
**Retriever**

Reading Comprehension

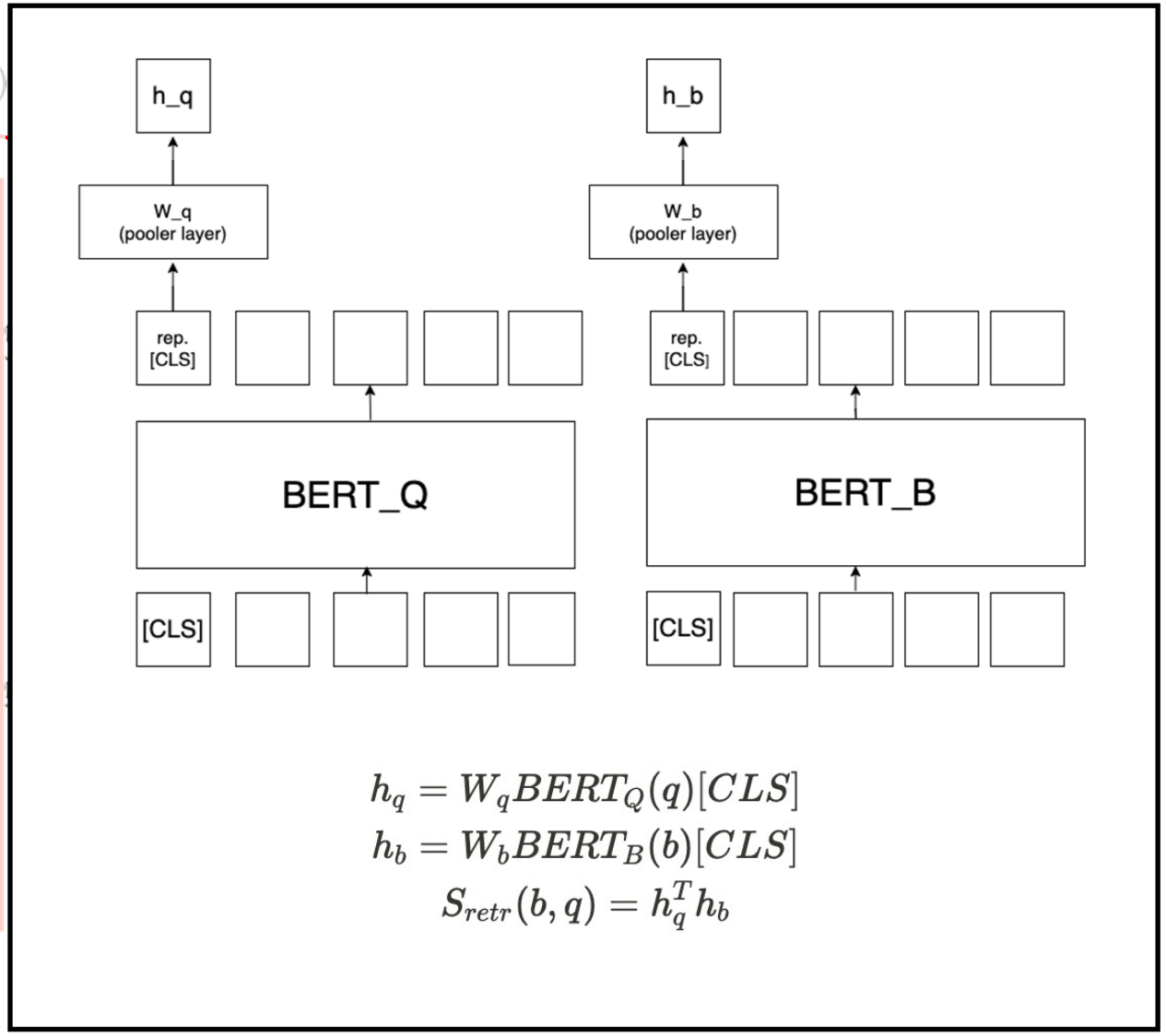
**Reader**

# 4. ORQA - Retriever

Scoring function : $S(b, s, q) = S_{retr}(b, q) + S_{read}(b, s, q)$

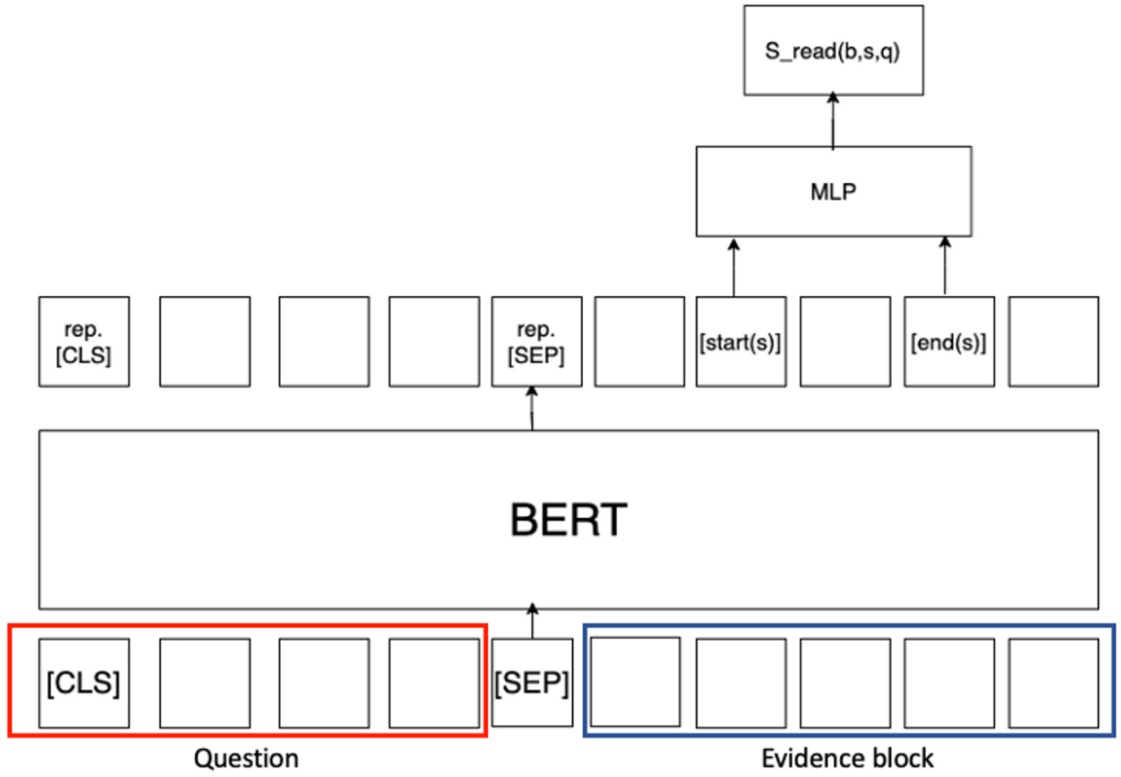



# Retriever

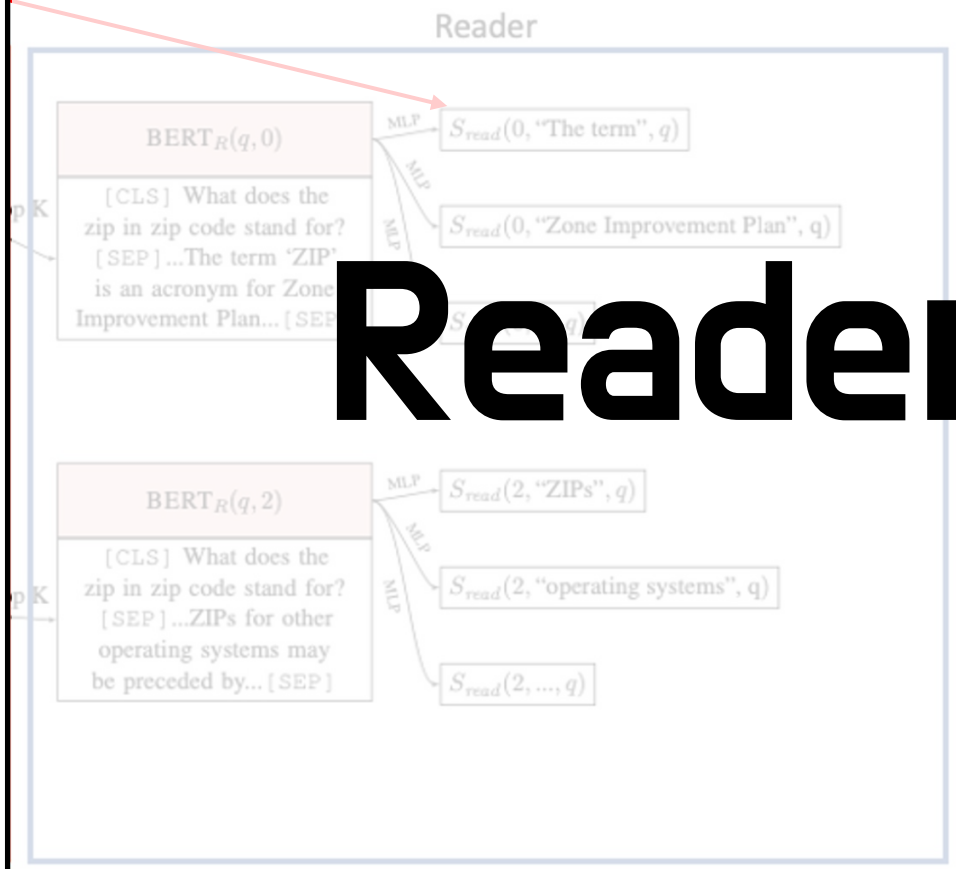



$$h_q = W_q BERT_Q(q)[CLS]$$
$$h_b = W_b BERT_B(b)[CLS]$$
$$S_{retr}(b, q) = h_q^T h_b$$

# 4. ORQA - Reader



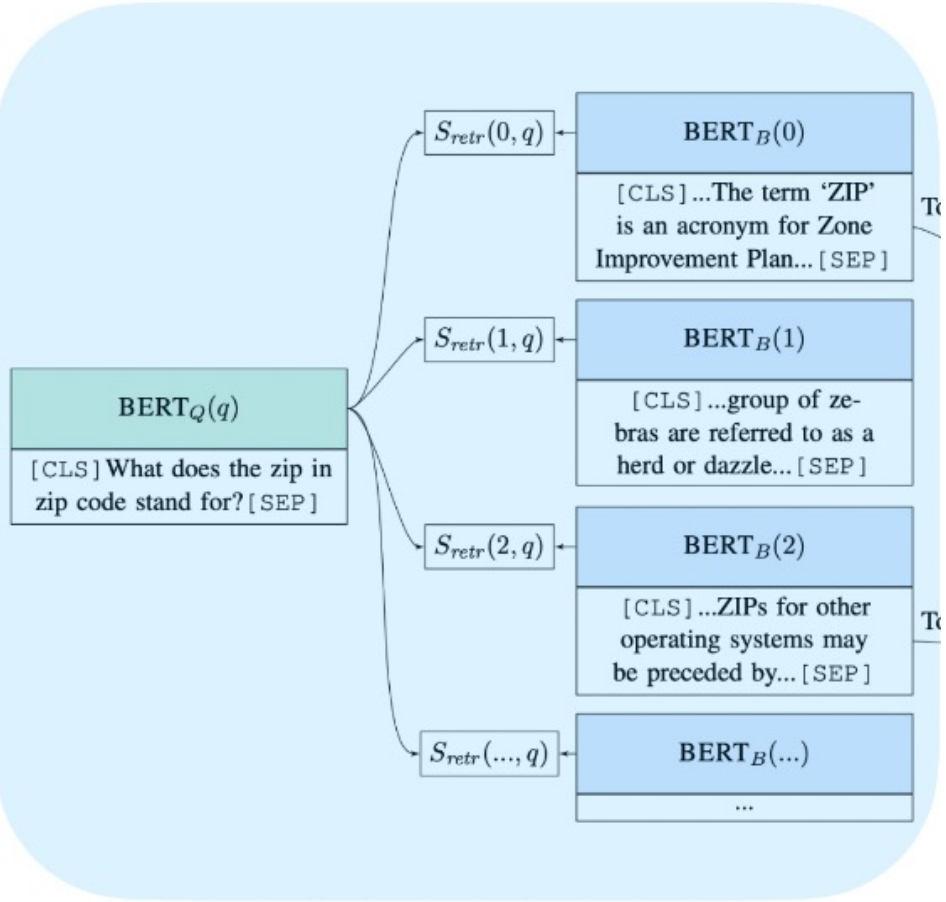

$$h_{start} = BERT_R(q, b)[START(s)]$$
$$h_{end} = BERT_R(q, b)[END(s)]$$
$$S_{read}(b, s, q) = MLP([h_{start}; h_{end}])$$



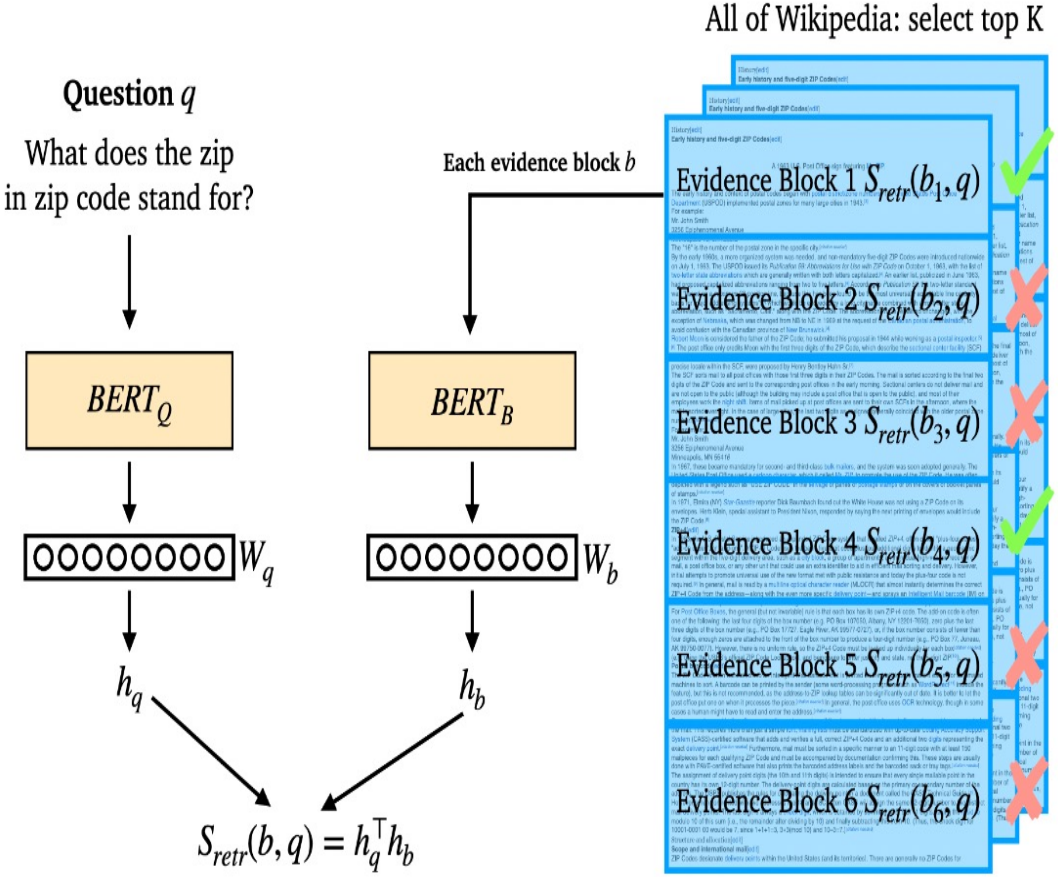

Reader

# 4. ORQA : Open-Retrival Question Answering

## Retriever

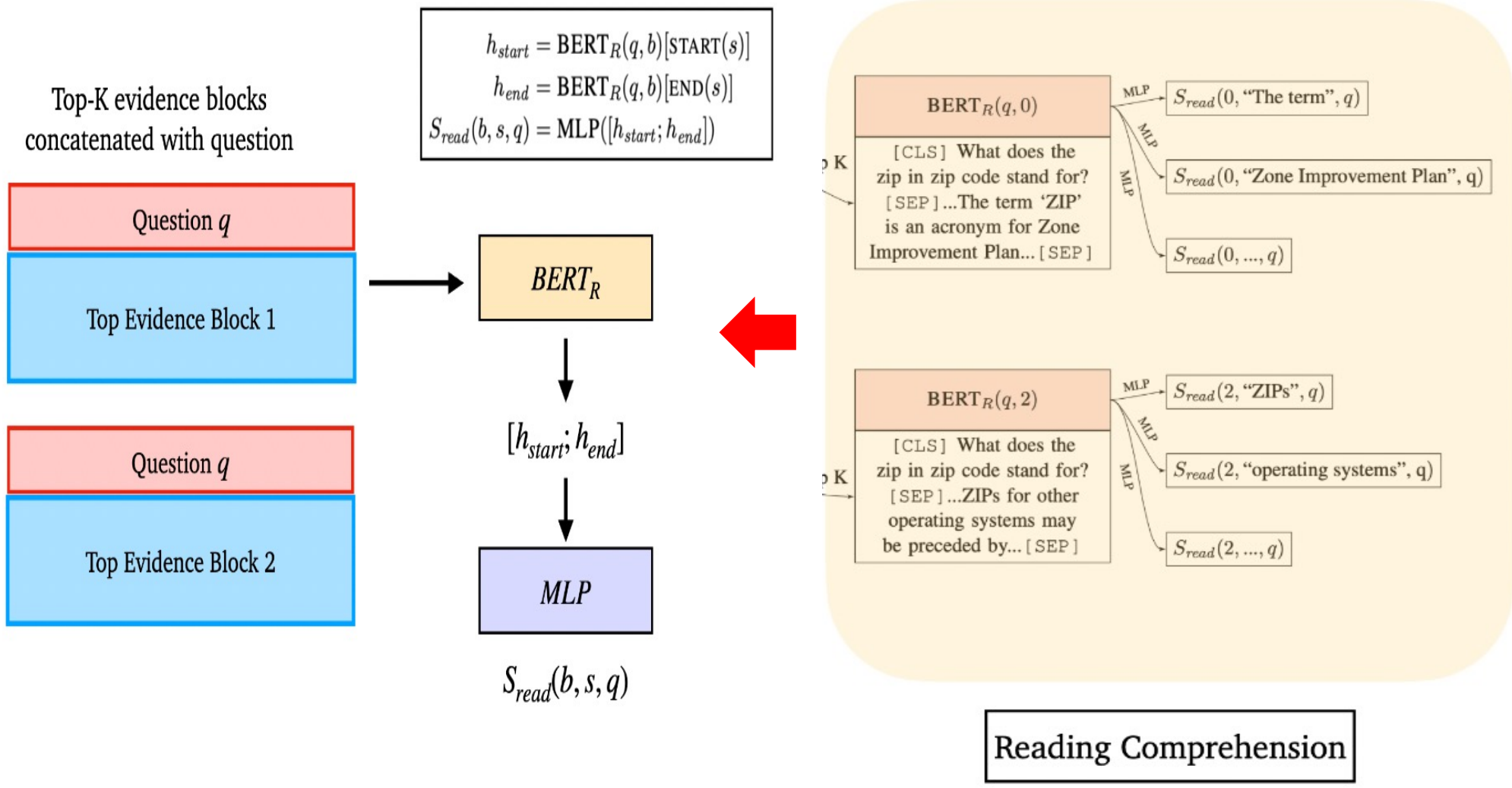



Information Retrieval

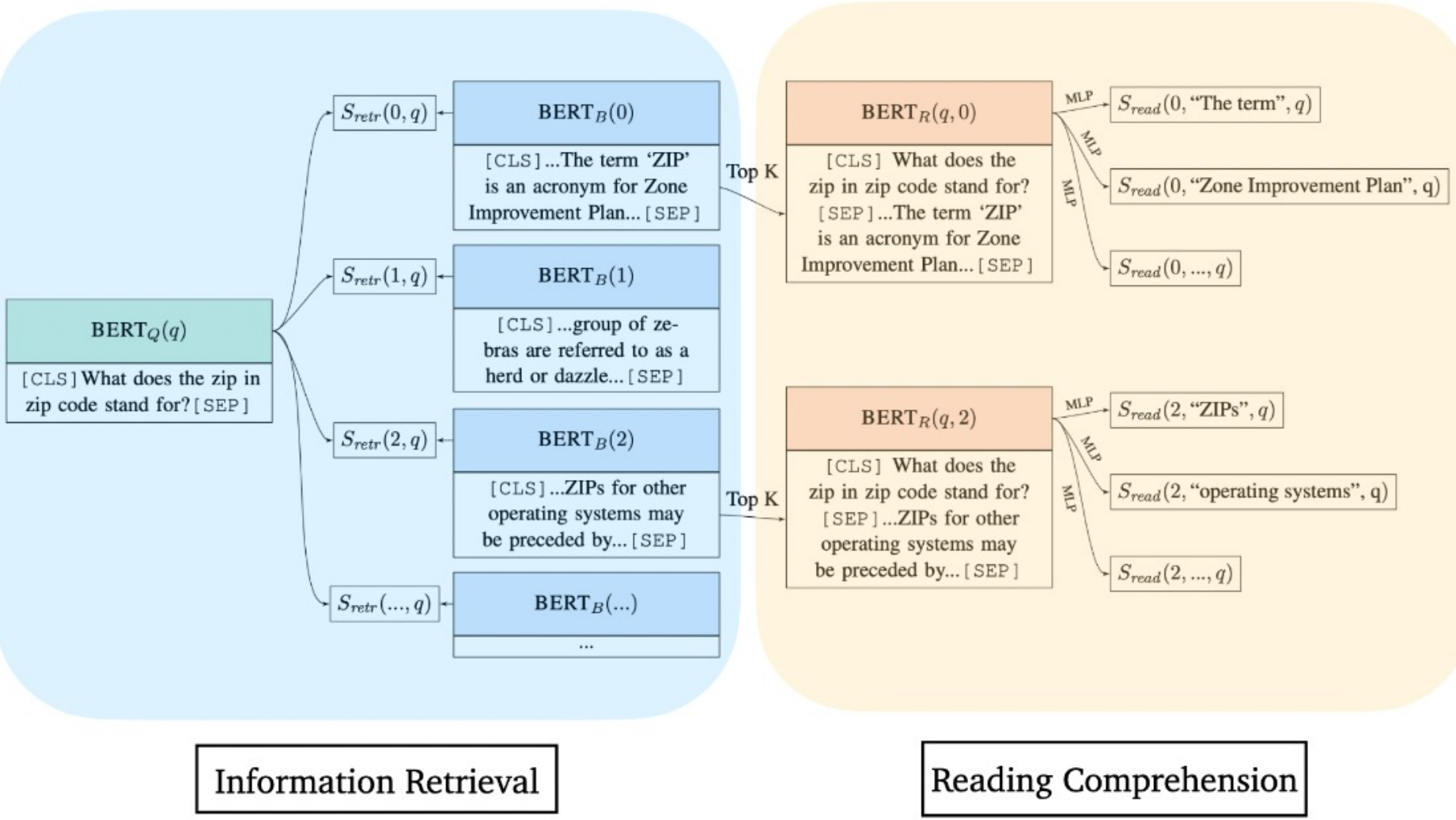


# 4. ORQA : Open-Retrival Question Answering

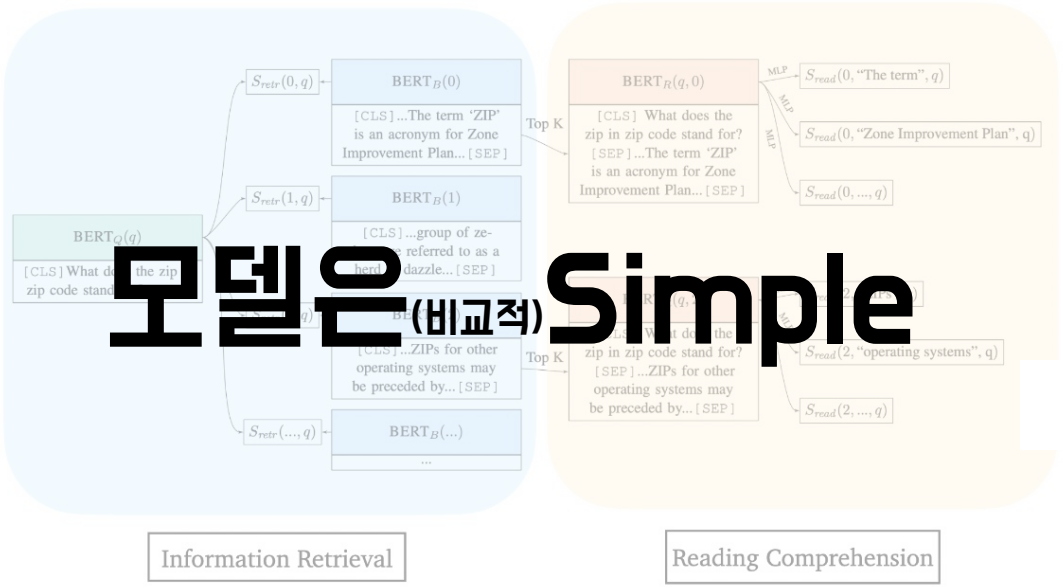
## Reader



# 4. ORQA : Open-Retrival Question Answering



# 4. ORQA - Inference & Learning Cahllenges




- Open evidence corpus가 너무 큼 (1300만개 이상의 evidence block)
- 탐색하려는 공간 전체가 latent space이므로 teacher-forcing 접근 방식을 적용하기 어려움
- Large number of spuriously ambiguous derivations(가짜 모호성)

↳ 학습은 매우 Difficult → Inverse Cloze Task

# ICT - Inverse Cloze Task

## Cloze Task : 맥락을 통해 텍스트를 유추하는 task

In this paper, we improve the fine-tuning based approaches by proposing BERT: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. BERT alleviates the previously mentioned unidirectionality constraint by using a "masked language model" (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked

[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018]

CLOZE TEST

*In across but like of to*

Every now and then I feel ____ a newly-qualified teacher who wants ____ try everything new that he comes ____ (honestly it doesn't happen that often any more ____ when it does I take advantage ____ it!) ____ one ____ these moments, I tried this amazing piece ____ software ____ create cloze texts

SCORE: ?

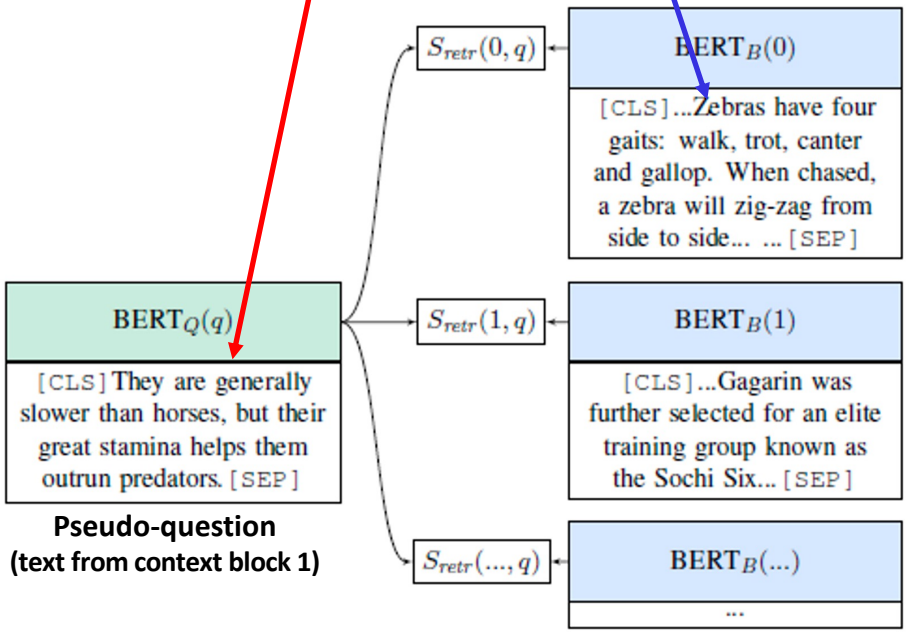
created with *the online Cloze Test Creator* © 2009 Lucy Georges

# ICT pre-training - Inverse Cloze Task

## Inverse Cloze Task : 텍스트(문장)를 통해 맥락을 유추하는 task

[Original Context]

"...Zebras have four gaits: walk, trot, canter and gallop. They are generally slower than horses, but their great stamina helps them outrun predators. When chased, a zebra will zigzag from side to side..."




Pseudo-question
(text from context block 1)

Pseudo-evidence
(context block 1)

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

q : random sentence (pseudo-question) of original context

b : text surrounding q

BATCH : set of evidence blocks in the batch (sampled negatives)

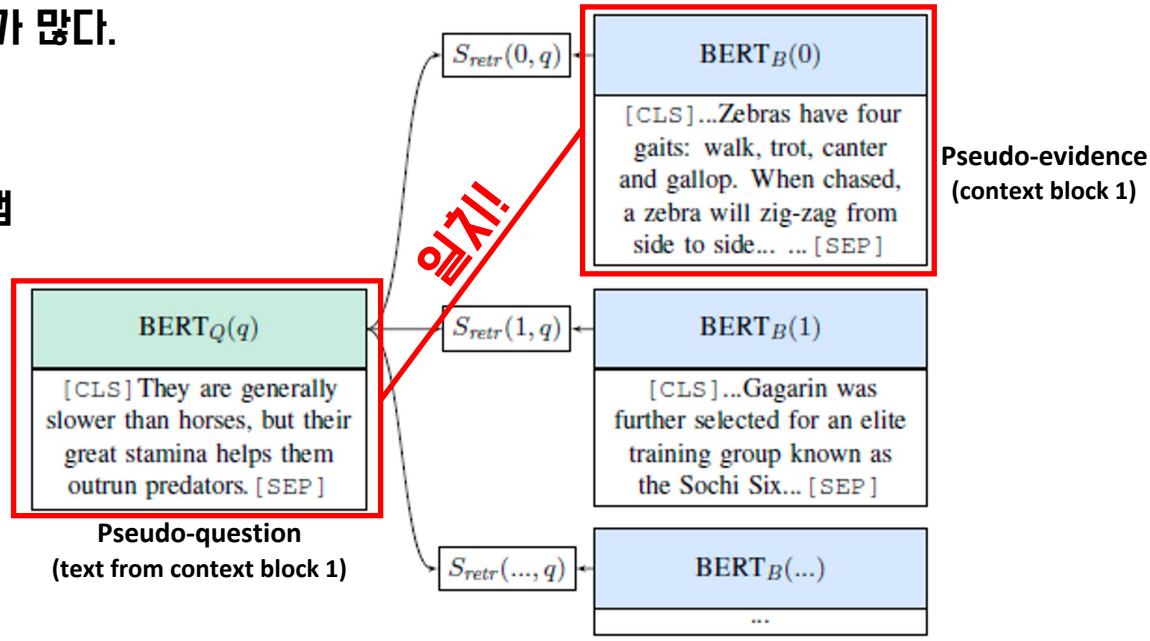
# ICT의 이점

- word matching feature 이상의 것을 학습 가능
  - 실제 QA에서는 질문에서 언급되지 않은 부분을 찾아서 답을 해주는 경우가 많다.

    (정보 검색과의 차이)
  - 예시 : qeustion에 zebra가 없어도 이것이 zebra에 대한 설명을 찾아냄
- evidence block(Wikipedia)을 인코딩 하는 BERT_B를 학습 할 필요가 없다.

  Question encoder 인 BERT_Q만 fine-tuning하면 된다. (속도 향상)
  - 이 과정에서 사전에 encoding된 evidence block을 검색하는 방법 학습
- spurious ambiguity를 피하도록 bias를 줄 수 있다.

| Example | Supportive Evidence | Spurious Ambiguity |
|---|---|---|
| **Q**: Who is credited with developing the XY coordinate plane? **A**: René Descartes | ...invention of Cartesian coordinates by **René Descartes** revolutionized... | ...**René Descartes** was born in La Haye en Touraine, France... |
| **Q**: How many districts are in the state of Alabama? **A**: seven | ...Alabama is currently divided into **seven** congressional districts, each represented by ... | ...Alabama is one of **seven** states that levy a tax on food at the same rate as other goods... |



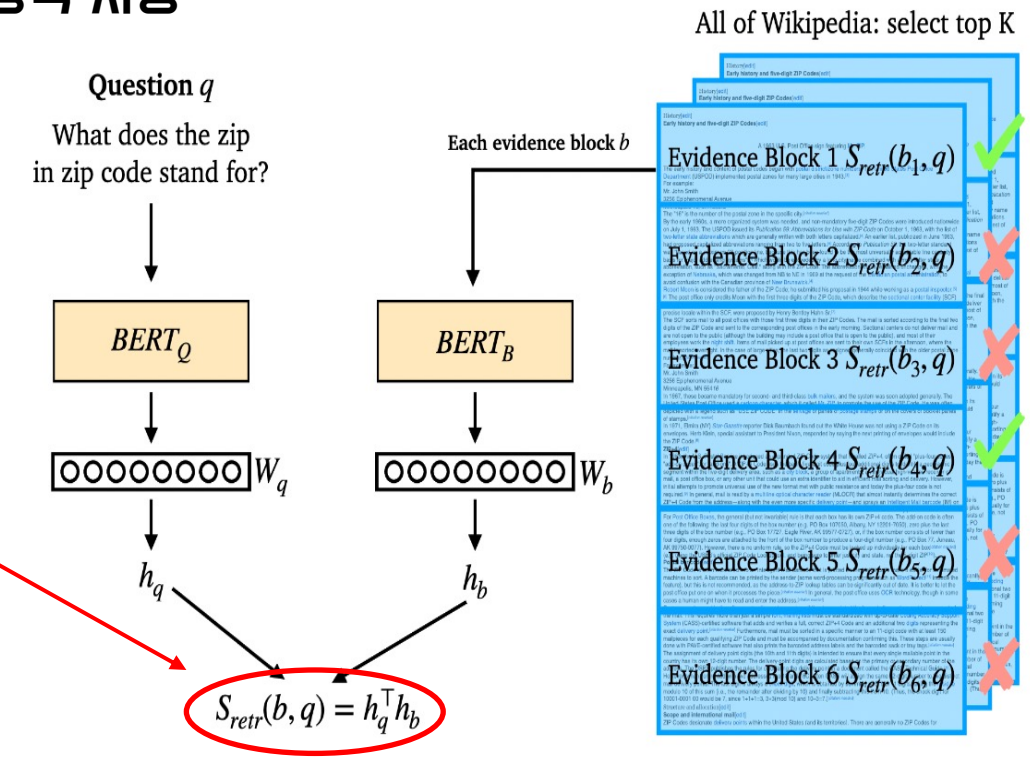

일치하는 답변이 발견되지 않으면 example을 폐기

(ICT pre-train으로 실제 예시의 10% 내외를 폐기)

# 목차

# 6. Inference

- 모든 Evidence block들은 finetuning에서 encoding할 필요 X (pre-training에서 이미 함)
  - Fixed block encoders already provide a useful representation for retrieval
- 내적값의 maximum을 빠르게 찾기 위한 index를 pre-compile 할 수 있다
- Inference는 pre-compiled index를 이용한 beam search 방식 사용
- Tok-k개의 evidence block 검색 후 reader score만 계산
  - 논문에서 k = 5



# 목차

# 7. Learning

1. Answer derivation에 대한 분포 정의
2. Gold answer a가 주어지면, beam search를 통해 모든 correct derivations를 찾고

   marginal log-likelihood를 통해 최적화
3. Top-k가 아닌 Top-c를 사용해 좀 더 포괄적으로 evidence block을 골라 계산
4. Final loss는 둘 다 업데이트

# 7. Learning

## 1. Answer derivation에 대한 분포 정의

$$P(b, s|q) = \frac{\exp(S(b, s, q))}{\sum_{b' \in \text{TOP}(k)} \sum_{s' \in b'} \exp(S(b', s', q))}$$

- q : question
- b : index of an evidence block
- s : span of text within block b
- top(k) : retriever score를 통해 검색된 k개의 evidence block (k = 5)

## 2. Gold answer a가 주어지면, beam search를 통해 모든 correct derivations를 찾고 marginal log-likelihood를 최적화

$$L_{\text{full}}(q, a) = -\log \sum_{b \in \text{TOP}(k)} \sum_{s \in b,\ a = \text{TEXT}(s)} P'(b, s|q)$$

정답이 span에 포함되어 있는 경우

- a = Text(s) : answer string a가 정확히 span s와 일치하는지 여부
- top k(=5) 문서의 span에 a가 포함되어 있는 것은 매우 적을 것이므로 top c 에서 고려 (c=5000)
  - Early learning !!

# 7. Learning

## 3. Early learning : top-k가 아닌 top-c를 사용해 좀 더 포괄적으로 evidence block을 골라 계산

$$P_{\text{early}}(b|q) = \frac{\exp(S_{retr}(b,q))}{\sum_{b' \in \text{TOP}(c)} \exp(S_{retr}(b',q))}$$

$$L_{\text{early}}(q,a) = -\log \sum_{b \in \text{TOP}(c),\ a \in \text{TEXT}(b)} P_{\text{early}}(b|q)$$

- K개(5)에서 C개(5000)로 고려하는 evidence block 개수는 늘어났지만
  (cost가 비교적 낮은) retrieval score만 update해서 계산량을 줄임

```
408    with tf.device("/cpu:0"):
409      retriever_outputs = retrieve(
410          features=features,
411          retriever_beam_size=retriever_beam_size,
412          mode=mode,                          -> C = 5000
413          params=params)
414
415    with tf.variable_scope("reader"):
416      reader_outputs = read(
417          features=features,
418          retriever_logits=retriever_outputs.logits[:reader_beam_size],
419          blocks=retriever_outputs.blocks[:reader_beam_size],        -> k = 5
420          mode=mode,
421          params=params,
422          labels=labels)
423
424    predictions = get_predictions(reader_outputs, params)
```

## 4. Final loss는 둘 다 업데이트

$$L(q,a) = L_{\text{early}}(q,a) + L_{\text{full}}(q,a)$$

(https://github.com/google-research/language/blob/60b3abae79a4199e19f28013754bd69f65ad670f/language/orqa/models/orqa_model.py)

# 8. Experimental Setup - Dataset

- **Natural Question**
  - open version의 dataset. 짧은 answer를 가진 question을 사용했고, 주어진 evidence document는 제거.
  - 많은 token을 가진 answer에서 역시 5개의 token보다 더 많이 삭제함. 왜냐면 긴 토큰의 answer는 extractive snippet과 비슷하기 때문.
- **WebQuestions**
  - Google Suggest API의 샘플 question을 포함하고 있음
  - annotated answer(string type의 representation만 취급함)
- **CuratedTrec**
  - TREC QA data의 question-answer pair로 이루어진 corpus
  - question은 MSNSearch나 AskJeeves logs 같은 real queries
- **TriviaQA**
  - trivia QA pair(from web)
  - unfiltered set 활용, supervised evidence는 버림
- **SQuAD**
  - ODQA 보다는 reading comprehension에 더 적합한 dataset
  - Wikipedia 문단에서 선택된 answer spans 와 annotator들에 의해 쓰여진 question

# 8. Experimental Setup - Dataset

- 다양한 QA pairs를 평가하는 것이 중요
  - 모든 존재하는 dataset들이 내부에 bias들을 가지고 있을 수 있기 때문
- Natural Questions, WebQuestions, CuratedTrec
  - 질문자들이 정답을 모른 채로 질문한 것들임
  - 진짜 정보를 찾는 질문
  - 그래서 moderate bias(중간정도의 bias)가 있다고 볼 수 있음
- TriviaQA, SQuAD
  - 정보가 필요해서 한 질문들이 아님

| Dataset | Question writer knows answer | Question writer knows evidence | Tool-assisted answer |
|---|---|---|---|
| Natural Questions | | | ✓ |
| WebQuestions | | | ✓ |
| CuratedTrec | | | ✓ |
| TriviaQA | ✓ | | |
| SQuAD | ✓ | ✓ | |

Table 4: A breakdown of biases in existing QA datasets. These biases are associated with either the question or the answer.

# 8. Experimental Setup - Implementation Details

## Hyperparameters

- 12 transformer layers (hidden size 768)
- 128 dimensions
- BERT 와 동일한 optimizer사용
- pre-training 할 때 (ICT로 retriever를)
  - lr: 1e-4
  - batch size: 4096
  - 100k steps
- fine-tunning 할 때
  - lr: 1e-5
  - batch size: 1
  - larger dataset(nq, TQA, SQuAd)에는 2 epoch을 돌렸지만 smaller dataset(WebQuestions, CuratedTrec)에서는 20 epoch 돌림

## Evidence corpus

- English Wikipedia snapshot(Dec. 20, 2018)
- BERT tokenizer에 기반하여 최대 288 wordpieces로 자름
- 13 million evidence block 이상

# 목차

# 9. Main Results - Baselines

Retrieval score을 활용해 다른 retrieval methods 와 비교해보자.

## 1) BM25

- 기존의 SOTA (unsupervised retrieval method)
- IR task 와 evidence retrieval for QA 둘 다에 튼튼한 method이기에 비교군으로 사용함
- BM25가 학습 요소가 아니기 때문에(not trainable) fine-tuning 하는 동안의 retrieved evidence는 고정되어 있다.
- final score : BM25 와 reader score의 가중합

## 2) Language Models

- 또 다른 비교군으로 unsupervised neural retrieval은 traditional IR을 뛰어넘기 어렵기로 유명
- 그래서 비교 baseline으로 LM의 unsupervised pooled representation을 실험
  - 두 개의 128 dim representation (널리 사용됨)
    - NNLM(문맥과 독립적인 embeddings)
    - ELMo(small, 문맥과 관련된 양방향의 LSTM)

정답 X
(ORQA win)

정답 O
(BM25 + BERT win)

| | Model | BM25 +BERT | NNLM +BERT | ELMo +BERT | ORQA |
|---|---|---|---|---|---|
| Dev | Natural Questions | 24.8 | 3.2 | 3.6 | **31.3** |
| | WebQuestions | 20.8 | 9.1 | 17.7 | **38.5** |
| | CuratedTrec | 27.1 | 6.0 | 8.3 | **36.8** |
| | TriviaQA | **47.2** | 7.3 | 6.0 | 45.1 |
| | SQuAD | **28.1** | 2.8 | 1.9 | 26.5 |
| Test | Natural Questions | 26.5 | 4.0 | 4.7 | **33.3** |
| | WebQuestions | 17.7 | 7.3 | 15.6 | **36.4** |
| | CuratedTrec | 21.3 | 4.5 | 6.8 | **30.1** |
| | TriviaQA | **47.1** | 7.1 | 5.7 | 45.0 |
| | SQuAD | **33.2** | 3.2 | 2.3 | 20.2 |

Table 5: **Main results**: End-to-end exact match for open-domain question answering from question-answer pairs only. Datasets where question askers know the answer behave differently from datasets where they do not.

# 9. Main Results - Results

- BM25는 powerful 한 retrieval system
  - word matching에 있어서 LM은 해당 task를 잘 캐치하지 못함.
- ORQA는 진짜 정보를 찾고자 하는 datasets (Natural Questions WebQuestions. CuratedTrec)에서 BM25를 능가했다. (6~19 points)
- 그러나 question asker가 정답을 이미 알고 있는 데이터셋인 SQuAD 와 TriviaQA의 경우 retrieval problem은 traditinal IR과 비슷하다.
- 128 차원으로 압축된 벡터는 evidence의 모든 단어들을 정확하게 표현하는 BM25보다 안 좋음.

정답 X
(ORQA win)

정답 O
(BM25 + BERT win)

| | Model | BM25 +BERT | NNLM +BERT | ELMo +BERT | ORQA |
|---|---|---|---|---|---|
| Dev | Natural Questions | 24.8 | 3.2 | 3.6 | **31.3** |
| | WebQuestions | 20.8 | 9.1 | 17.7 | **38.5** |
| | CuratedTrec | 27.1 | 6.0 | 8.3 | **36.8** |
| | TriviaQA | **47.2** | 7.3 | 6.0 | 45.1 |
| | SQuAD | **28.1** | 2.8 | 1.9 | 26.5 |
| Test | Natural Questions | 26.5 | 4.0 | 4.7 | **33.3** |
| | WebQuestions | 17.7 | 7.3 | 15.6 | **36.4** |
| | CuratedTrec | 21.3 | 4.5 | 6.8 | **30.1** |
| | TriviaQA | **47.1** | 7.1 | 5.7 | 45.0 |
| | SQuAD | **33.2** | 3.2 | 2.3 | 20.2 |

Table 5: **Main results**: End-to-end exact match for open-domain question answering from question-answer pairs only. Datasets where question askers know the answer behave differently from datasets where they do not.

- SQuAD dataset은 Dev data와 Test data 점수 차이가 크다.
- 536개의 적은 지문에서 10만개의 많은 질문을 뽑아냈기 때문 (data bias)
- 좋은 retrieval target(점수 하락이 크지 않은)을 위해서는
  - 1) training example
  - 2) IID assumption을 위반하는지
  - 3) 학습된 retrieval과 적합한가
- 등을 신경써야하고 이러한 이유에서 앞으로의 ODQA 모델들은 SQuAD dataset을 사용하지 않기를 제안

# 10. Analysis

- ICT 사전학습 시 모든 example(100%)을 마스킹하면 전혀 학습되지 않는 양상을 볼 수 있다.
- 90%만 마스킹해서 word-matching으로서의 역할도 할 수 있음
- 아예 masking을 하지 않았을 경우(0% masking), memory 하는 것에 문제는 없지만 (masking을 하지 않았으니까 정보 손실은 없다) QA에 일반화하지는 못한다.

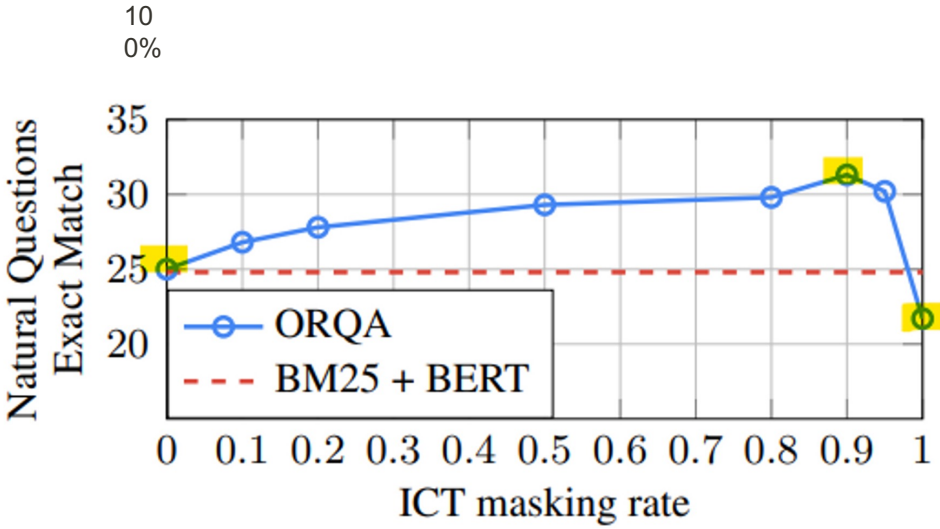



Figure 3: **Analysis**: Performance on our open version of the Natural Questions dev set with various masking rates for the ICT pre-training. Too much masking prevents the model from learning to exploit exact n-gram overlap. Too little masking makes language understanding unnecessary.

전체 example의 90%만 masking

Q. They are generally slower than horses, but their great stamina helps them outrun predators.

B. "...Zebras have four gaits: walk, trot, canter and gallop. (They are generally slower than horses, but their great stamina helps them outrun predators.) When chased, a zebra will zigzag from side to side..."

# 10. Analysis

| Example | ORQA | BM25 + BERT |
|---|---|---|
| **Q**: what is the new orleans saints symbol called<br>**A**: fleur-de-lis | ...The team's primary colors are old gold and black; their logo is a simplified **fleur-de-lis**. They played their home games in Tulane Stadium through the 1974 NFL season.... | ...the **SkyDome** was owned by Sportsco at the time... the sale of the New Orleans Saints with team owner Tom Benson... the Saints became a symbol for that community... |
| **Q**: how many senators per state in the us<br>**A**: two | ...powers of the Senate are established in Article One of the U.S. Constitution. Each U.S. state is represented by **two** senators... | ...The Georgia Constitution mandates a maximum of **56** senators, elected from single-member districts... |
| **Q**: when was germany given a permanent seat on the council of the league of nations<br>**A**: 1926 | ...Under the Weimar Republic, Germany (in fact the "Deutsches Reich" or German Empire) was admitted to the League of Nations through a resolution passed on September 8 **1926**. An additional 15 countries joined later... | ...the accession of the German Democratic Republic to the Federal Republic of Germany, it was effective on **3 October 1990**...Germany has been elected as a non-permanent member of the United Nations Security Council... |
| **Q**: when was diary of a wimpy kid double down published<br>**A**: November 1, 2016 | ..."Diary of a Wimpy Kid" first appeared on FunBrain in 2004, where it was read 20 million times. The abridged hardcover adaptation was released on **April 1, 2007**... | Diary of a Wimpy Kid: Double Down is the eleventh book in the "Diary of a Wimpy Kid" series by Jeff Kinney... The book was published on **November 1, 2016**... |

Table 7: **Analysis**: Example predictions on our open version of the Natural Questions dev set. We show the highest scoring derivation, consisting of the evidence block and the predicted answer in bold. ORQA is more robust at separating semantically distinct text that have high lexical overlap. However, the limitation of the 128-dimensional vectors is that extremely specific concepts are less precisely represented.

# 목차

# 12. Conclusion - ORQA's contribution

- Both retriever and reader are learnable with NNs ( = BERT)
  - First model to learn retriever and reader jointly
- Only learned from question-answering pairs; No reading comprehension datasets!
- A new pre-training task called Inverse Cloze Task (ICT) to address the challenging retrieval problem.

# References


- https://www.notion.so/yukyunglee/ORQA-fca9f63ad536463ca2c4b453d4f7ccfa
- https://github.com/danqi/acl2020-openqa-tutorial/blob/master/slides/part1-introduction.pdf
- https://lilianweng.github.io/posts/2020-10-29-odqa/
- https://www.youtube.com/watch?v=_K6SVN_ygzk
- https://github.com/google-research/language/blob/60b3abae79a4199e19f28013754bd69f65ad670f/language/orqa/models/orqa_model.py

# 감사합니다