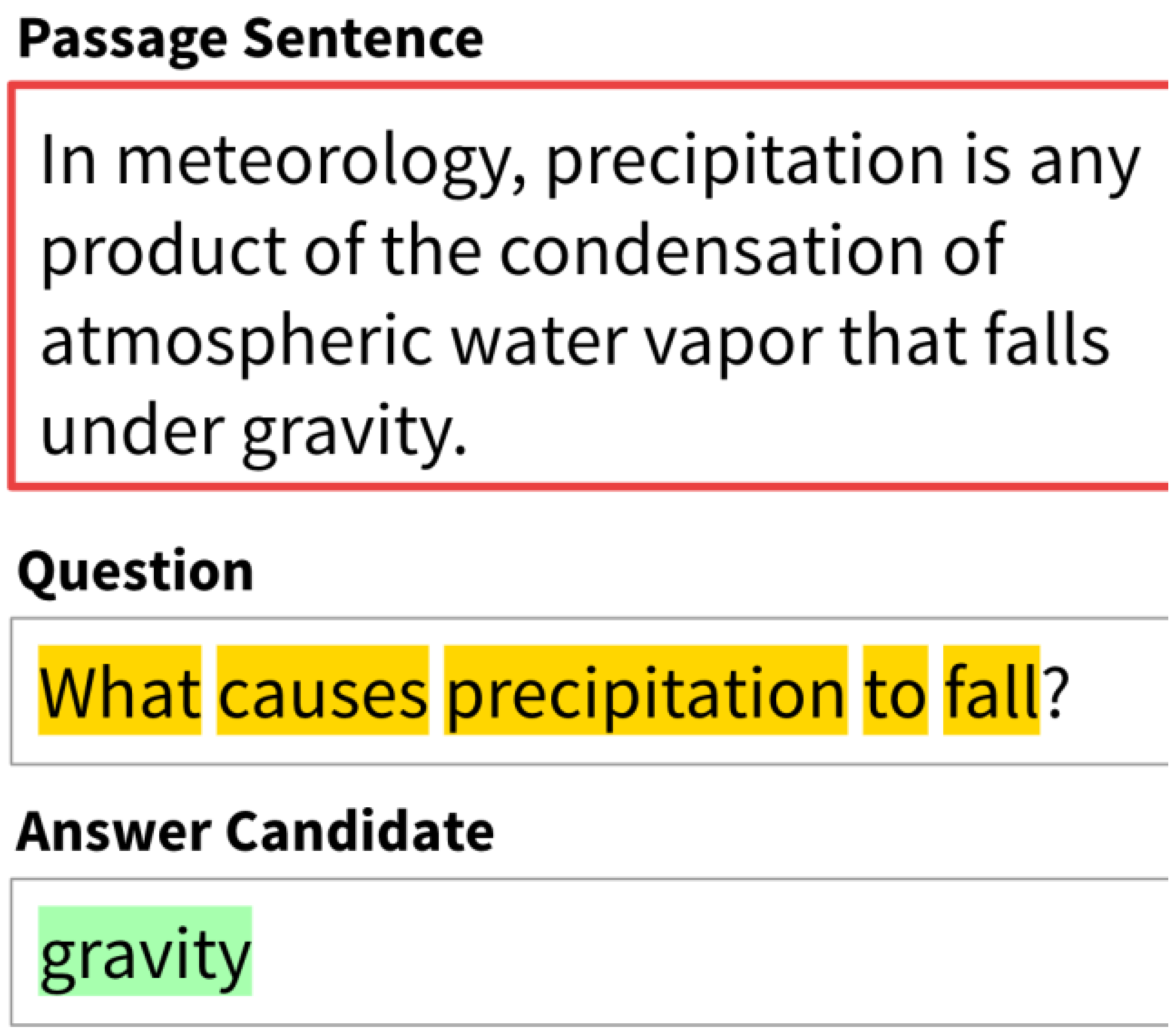
Open Domain Question Answering

김한성

Question Answering



Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

What causes precipitation to fall?

Answer Candidate

gravity

- 주어진 지문 Passage에서 입력 받은 Question에 대한 답을 찾아내는 시스템입니다.

- INPUT : Passage, Question
- OUTPUT : Answer

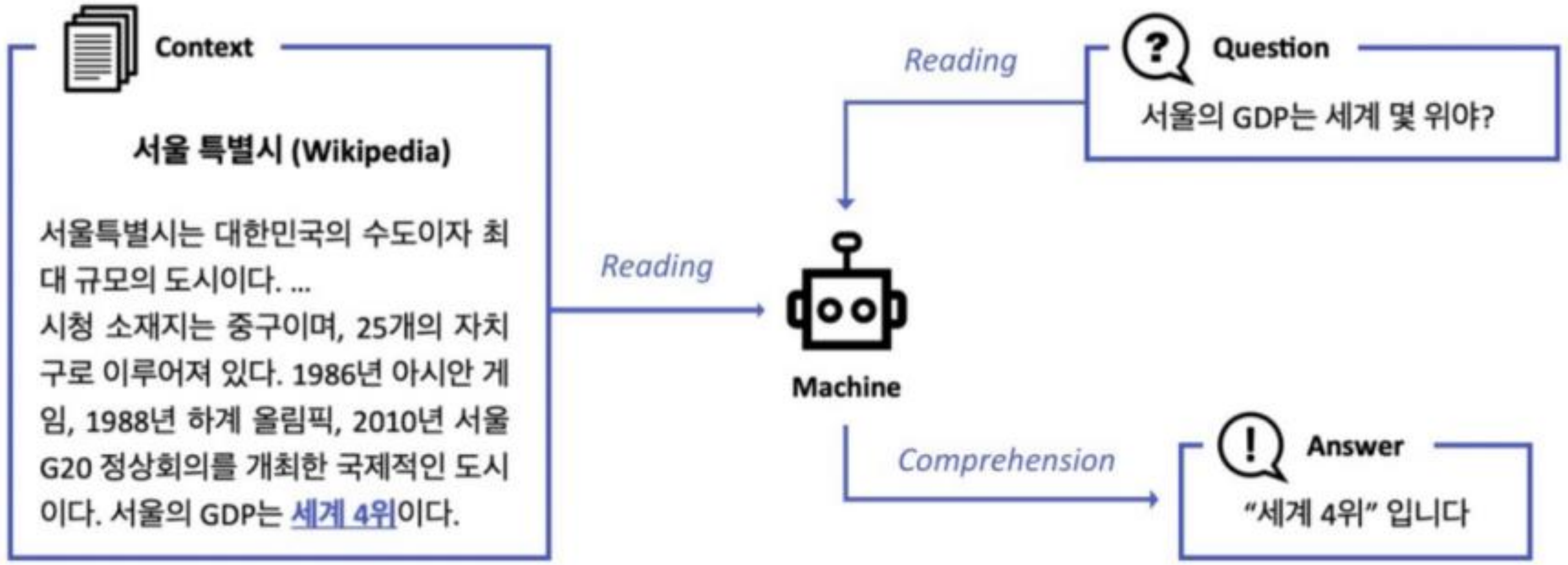
Answer와 Passage의 형태에 따라 Task 상이

Open-Domain QA

Definition

QA에서 Passage가 정해져 있지 않고 정답이 있을 것 같은 Passage를 직접 찾고 답을 추출

Open-domain : Domain 즉, 데이터의 분야가 2개 이상으로 질문에 대한 적절한 question, passage가 존재

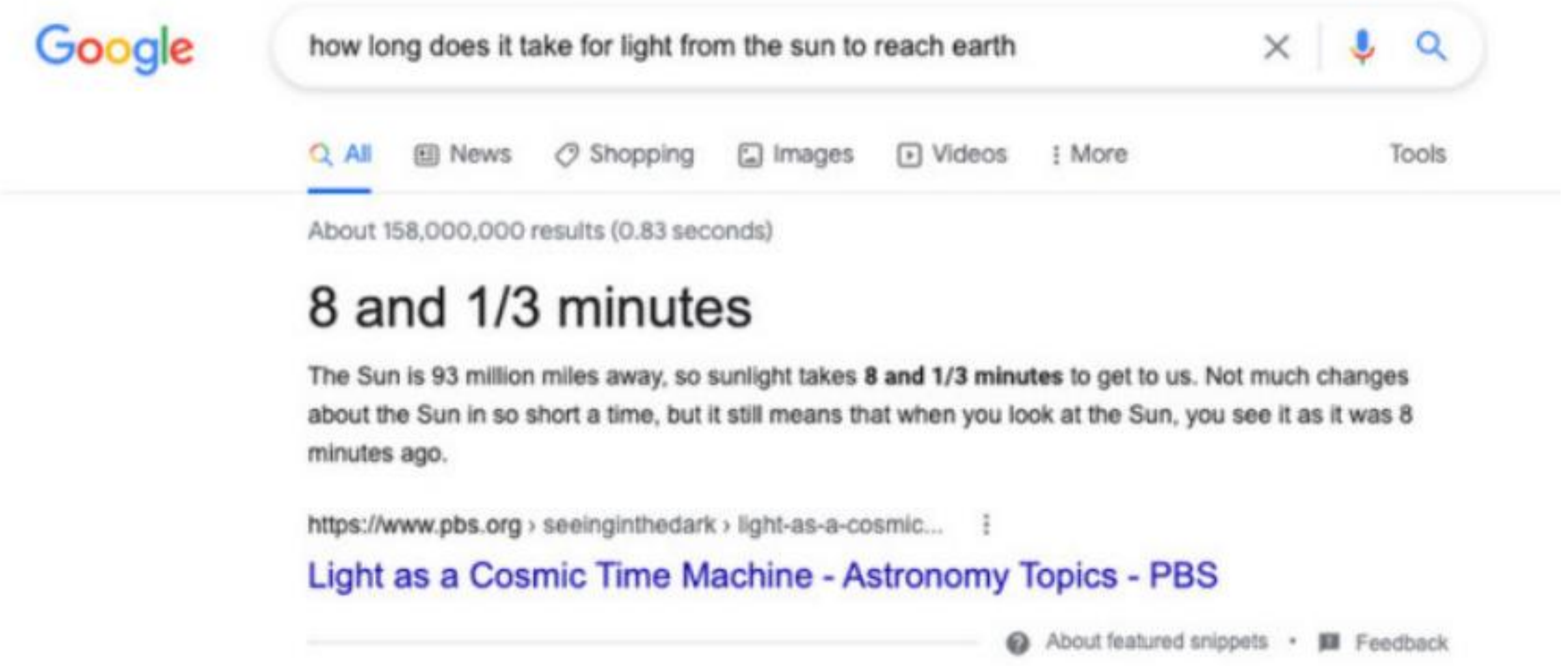


서울 특별시 (Wikipedia)

서울특별시는 대한민국의 수도이자 최대 규모의 도시이다. ...
시청 소재지는 중구이며, 25개의 자치구로 이루어져 있다. 1986년 아시안 게임, 1988년 하계 올림픽, 2010년 서울 G20 정상회의를 개최한 국제적인 도시이다. 서울의 GDP는 세계 4위이다.

서울의 GDP는 세계 몇 위야?

"세계 4위" 입니다



how long does it take for light from the sun to reach earth

About 158,000,000 results (0.83 seconds)

8 and 1/3 minutes

The Sun is 93 million miles away, so sunlight takes **8 and 1/3 minutes** to get to us. Not much changes about the Sun in so short a time, but it still means that when you look at the Sun, you see it as it was 8 minutes ago.

https://www.pbs.org › seeinginthedark › light-as-a-cosmic...

Light as a Cosmic Time Machine - Astronomy Topics - PBS

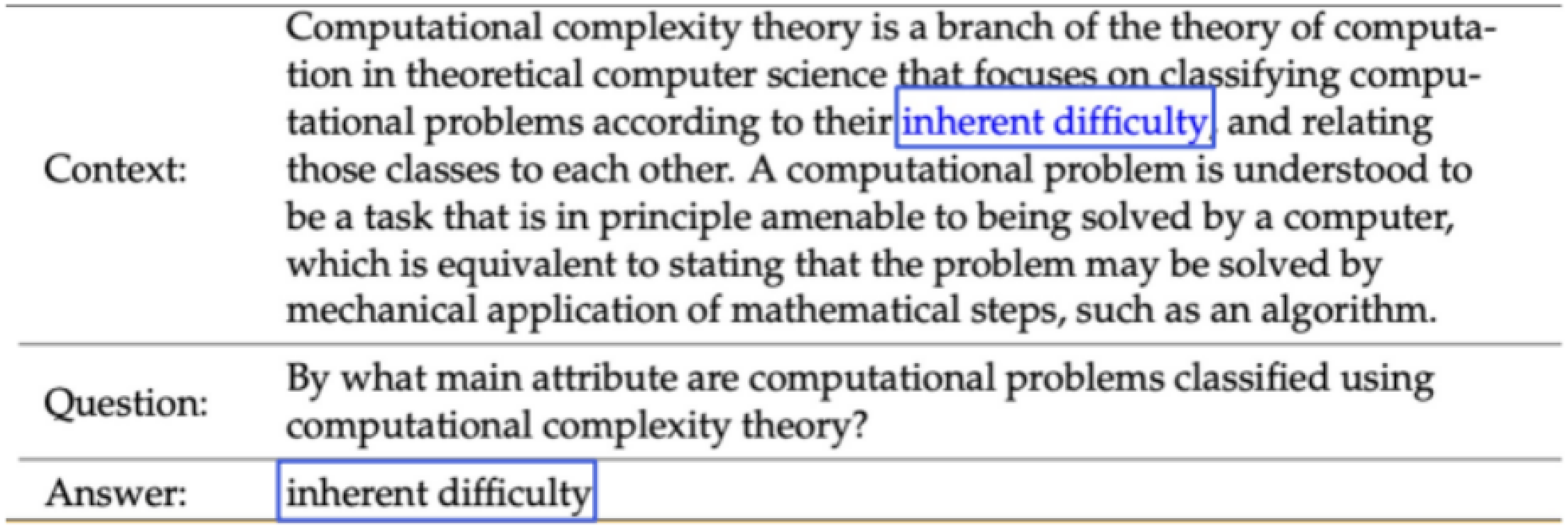
구글의 대표적인 예시

Open-Domain QA

종류 Extractive Answer

질의(question)에 대한 답이 지문(Document)의 하나의 span으로 존재

Span Extraction
ex) SQuAD, KorQuAD,
NewsQA, Natural Questions,
etc



Context:	Computational complexity theory is a branch of the theory of computation in theoretical computer science that focuses on classifying computational problems according to their inherent difficulty, and relating those classes to each other. A computational problem is understood to be a task that is in principle amenable to being solved by a computer, which is equivalent to stating that the problem may be solved by mechanical application of mathematical steps, such as an algorithm.
Question:	By what main attribute are computational problems classified using computational complexity theory?
Answer:	inherent difficulty

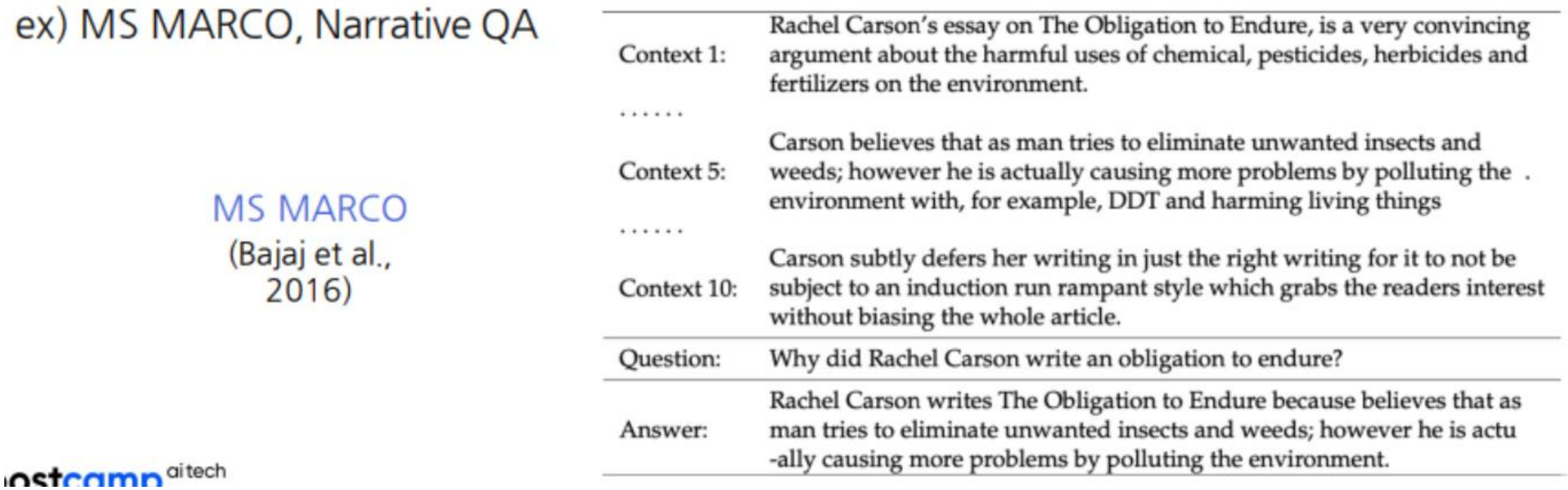
SQuAD, TriviaQA가 대표적인 Extractive answer task dataset
(한국어 : KLUE MRC, KorQuaD, AI hub 기계 독해)

OUTPUT : Answer span의 start point, end point

Open-Domain QA

종류 Descriptive Answer

답이 지문 내에서 추출한 span이 아니라, 질의를 보고 생성된 sentence (or free-form) 형태



ex) MS MARCO, Narrative QA

MS MARCO
(Bajaj et al.,
2016)

Context 1:	Rachel Carson's essay on The Obligation to Endure, is a very convincing argument about the harmful uses of chemical, pesticides, herbicides and fertilizers on the environment.
......	
Context 5:	Carson believes that as man tries to eliminate unwanted insects and weeds; however he is actually causing more problems by polluting the . environment with, for example, DDT and harming living things
......	
Context 10:	Carson subtly defers her writing in just the right writing for it to not be subject to an induction run rampant style which grabs the readers interest without biasing the whole article.
Question:	Why did Rachel Carson write an obligation to endure?
Answer:	Rachel Carson writes The Obligation to Endure because believes that as man tries to eliminate unwanted insects and weeds; however he is actu -ally causing more problems by polluting the environment.

MS MARCO 또는 Narrative QA에서 강점

OUTPUT : Answer를 eos token을 만날때 까지 생성

Open-Domain QA

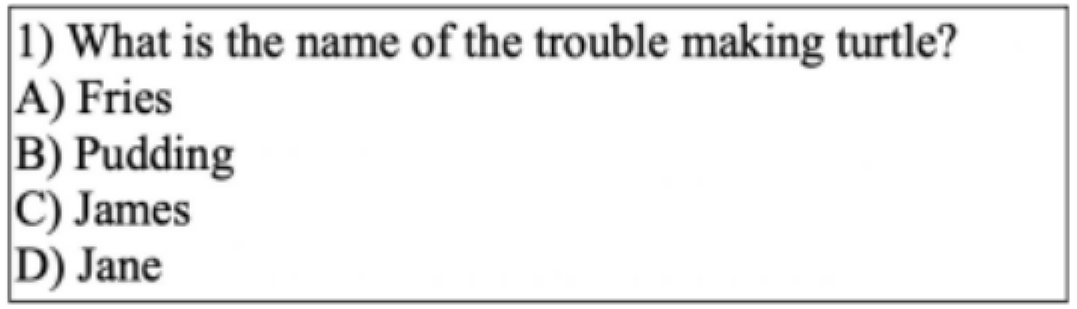
종류 Multiple-choice

질의에 대한 답을 여러 개의 answer candidates 중 하나로 고르는 형태

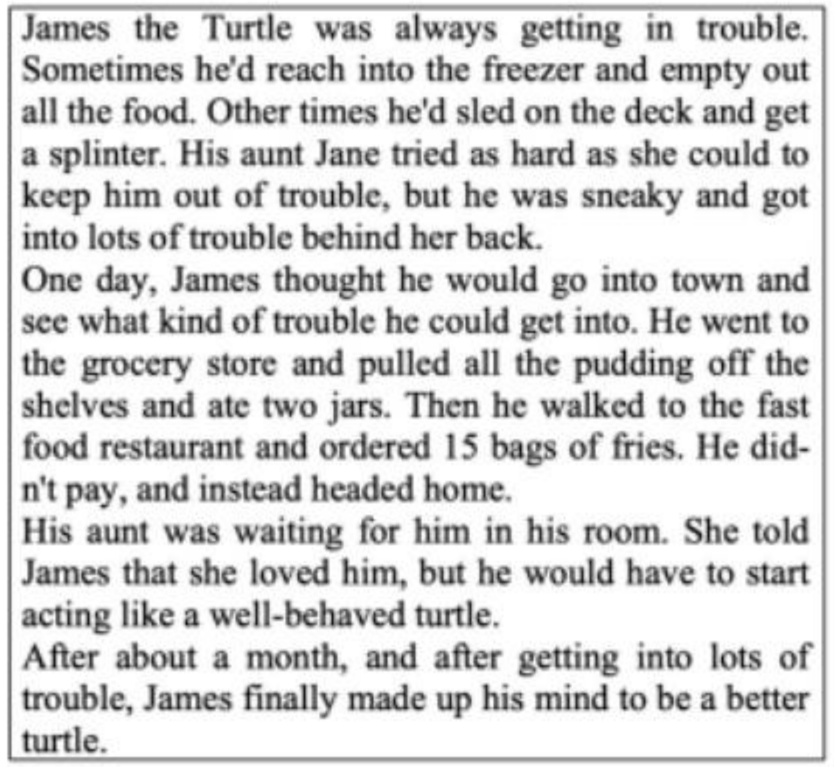
ex) MCTest, RACE, ARC, etc.

MCTest
(Richardson et al., 2013)

Question:



1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

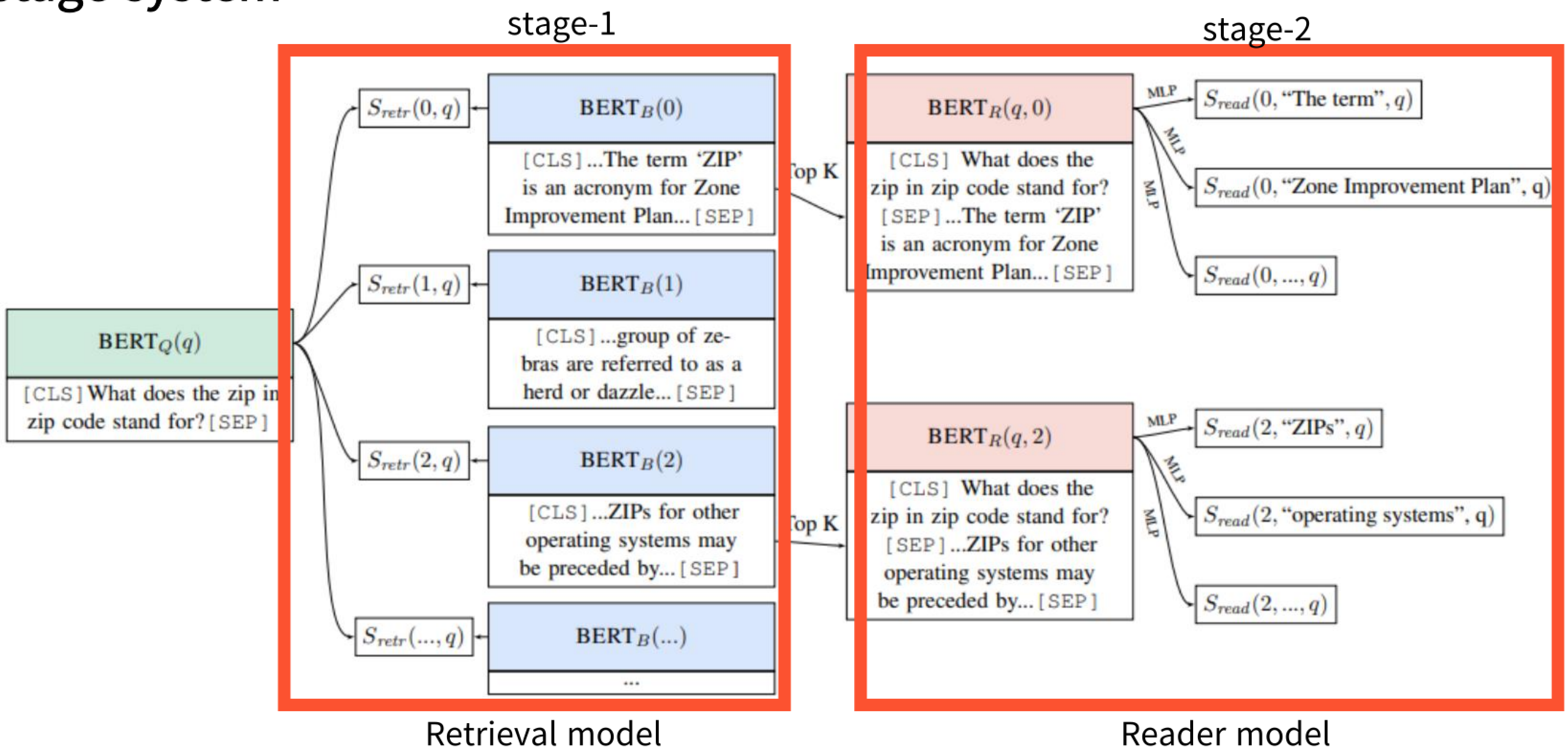


James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

OUTPUT : Answer의 번호를 예측

Open-Domain QA

2 stage system



stage-1
stage-2
BERT$_Q(q)$
[CLS] What does the zip in zip code stand for? [SEP]
$S_{retr}(0,q)$
BERT$_B(0)$
[CLS] ...The term 'ZIP' is an acronym for Zone Improvement Plan... [SEP]
$S_{retr}(1,q)$
BERT$_B(1)$
[CLS] ...group of zebras are referred to as a herd or dazzle... [SEP]
$S_{retr}(2,q)$
BERT$_B(2)$
[CLS] ...ZIPs for other operating systems may be preceded by... [SEP]
$S_{retr}(...,q)$
BERT$_B(...)$
...
Top K
Top K
BERT$_R(q,0)$
[CLS] What does the zip in zip code stand for? [SEP] ...The term 'ZIP' is an acronym for Zone Improvement Plan... [SEP]
MLP
MLP
MLP
$S_{read}(0, \text{"The term"}, q)$
$S_{read}(0, \text{"Zone Improvement Plan"}, q)$
$S_{read}(0, ..., q)$
BERT$_R(q,2)$
[CLS] What does the zip in zip code stand for? [SEP] ...ZIPs for other operating systems may be preceded by... [SEP]
MLP
MLP
MLP
$S_{read}(2, \text{"ZIPs"}, q)$
$S_{read}(2, \text{"operating systems"}, q)$
$S_{read}(2, ..., q)$
Retrieval model
Reader model

Open-Domain QA

Stage-1 : Information Retrieval (IR)

Sparse Embedding

BM25

$$score(D,Q)=\sum_{i=1}^{n} IDF(q_i) * \frac{\overbrace{f(q_i,D)}^{\text{문서 } D\text{에서 } q_i\text{의 term frequency}} * (k_1+1)}{f(q_i,D)+\underbrace{k_1}_{\text{파라메터}} * (1-\underbrace{b}_{\text{파라메터}}+b * \frac{\overset{\text{문서 } D \text{ 의 길이}}{|D|}}{\underbrace{avgdl}_{\text{문서 집합의 평균 문서 길이}}})}$$

$$IDF(q_i)=\ln(1+\frac{(\overbrace{docCount}^{\text{총 문서의 개수}} - f(q_i)+0.5)}{\underbrace{f(q_i)}_{\text{해당 단어를 포함하는 문서의 개수}}+0.5})$$

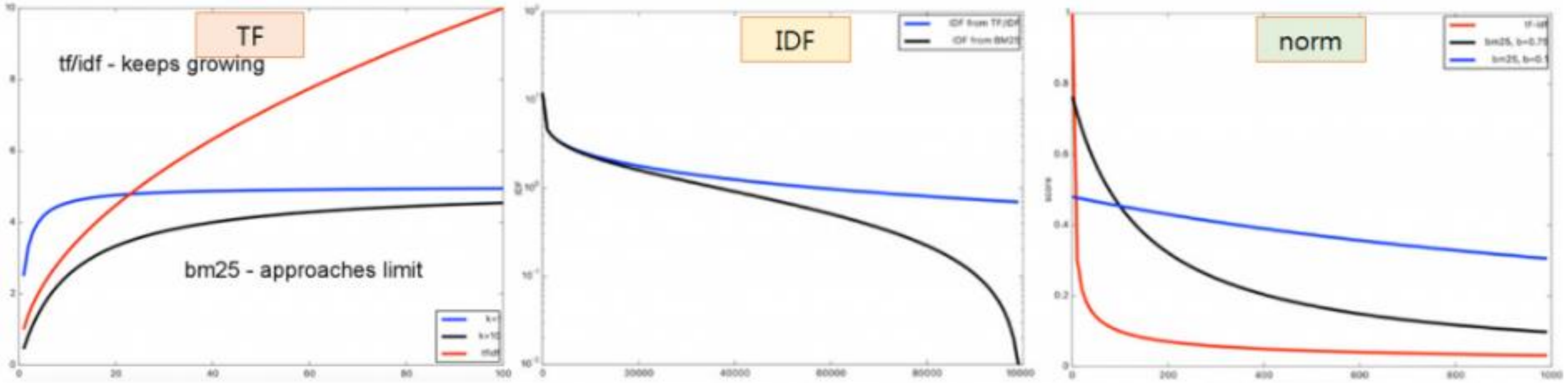
https://www.elastic.co/kr/elasticon/conf/2016/sf/improved-text-scoring-with-bm25

BM25가 TF-IDF보다 좋은 이유

1. TF(단어 빈도)가 일정 수준에서 수렴한다.
2. IDF의 영향이 커져 불용어가 검색점수에 덜 미친다.
3. 문서 길이의 영향(외부 요인)이 줄어든다.

$$score_{q,d} = norm(q) \times \sum_{t \ in \ q} \sqrt{tf_{t,d}} \times idf_t^2 \times norm(d,field) \times boost(t)$$

$$\text{bm25}(d) = \sum_{t\in q, f_{t,d}>0} \log\left(1+\frac{N-df_t+0.5}{df_t+0.5}\right) \cdot \frac{f_{t,d}}{f_{t,d}+k\cdot(1-b+b\frac{l(d)}{avgdl})}$$



Open-Domain QA

Stage-1 : Information Retrieval (IR)

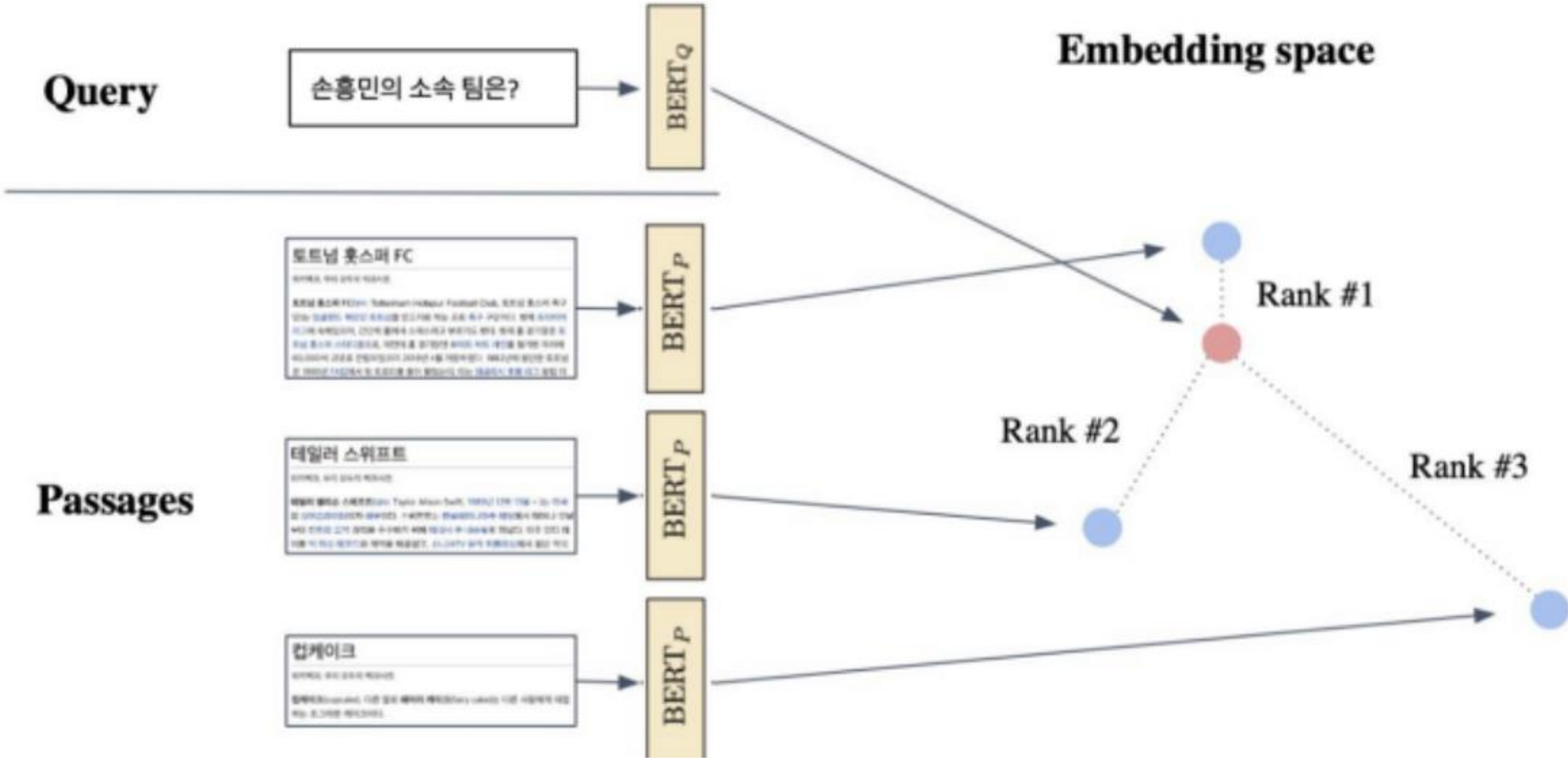
Dense Embedding

Sparse Embedding의 단점

1. 단어의 수 만큼 필요한 차원의 크기
2. 유사성을 고려하지 못함.
   (벡터 공간상 비교가 아닌 단어가 일치 하는지 비교하기 때문)

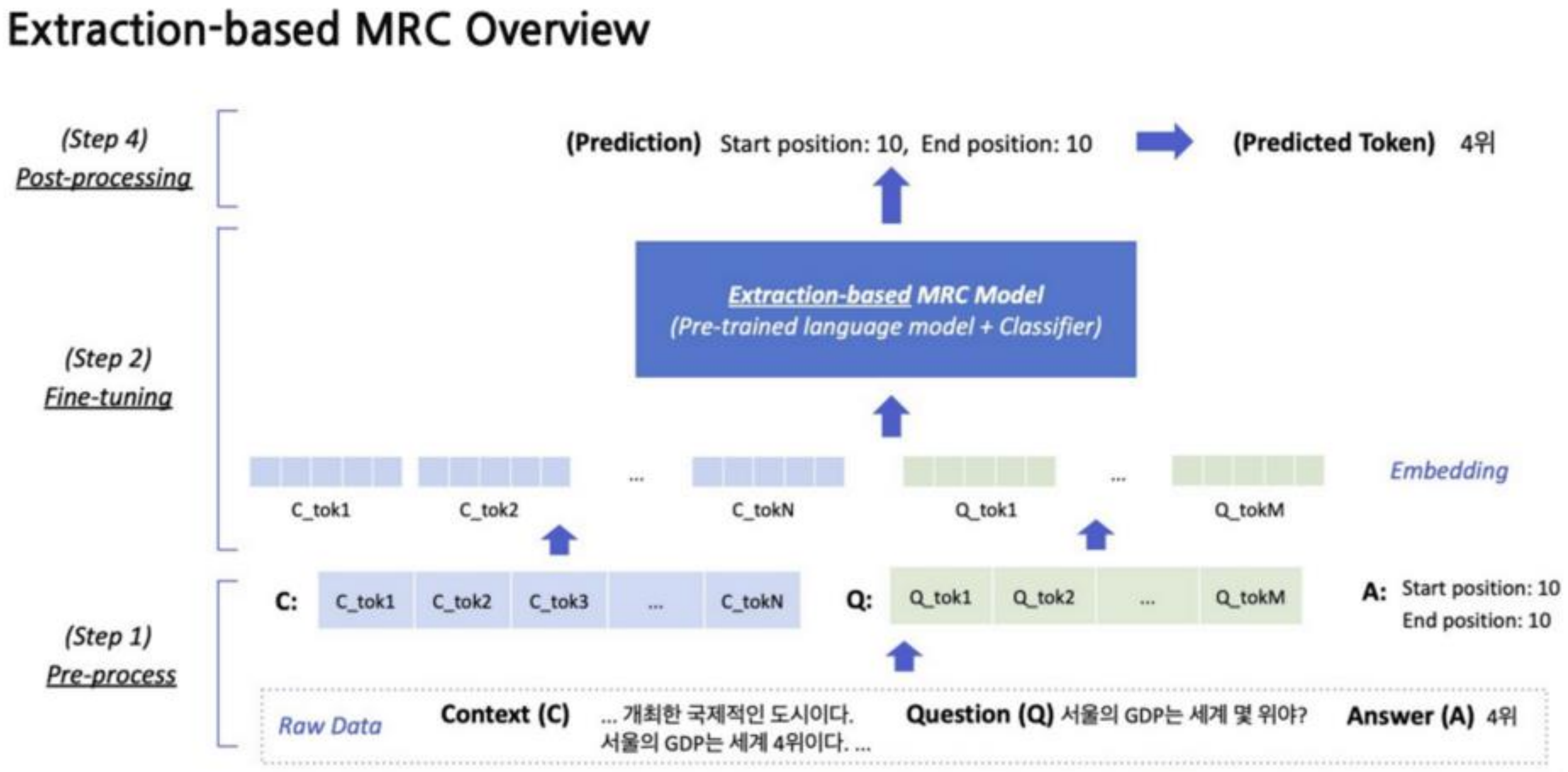
Dense Embedding 이란?

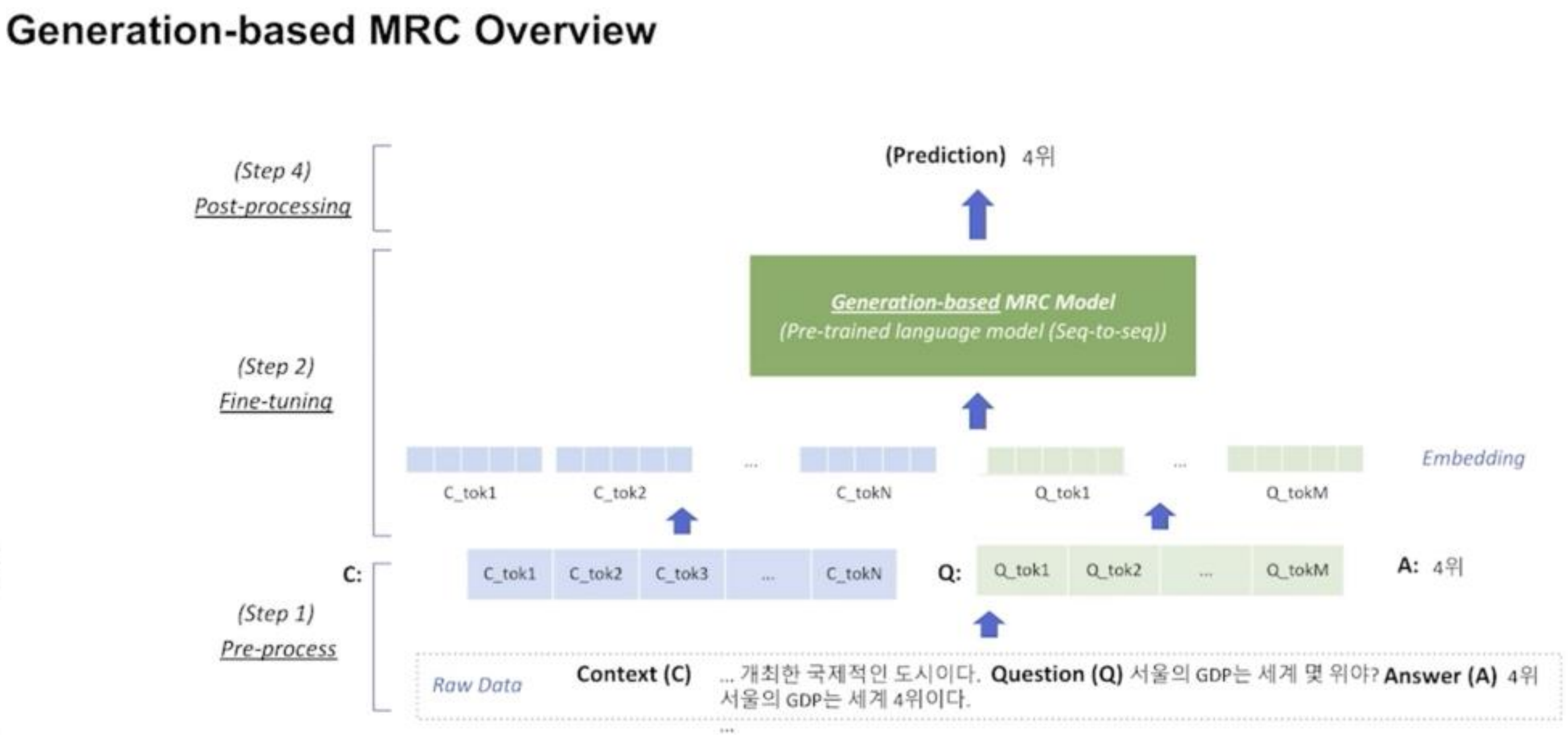
1. 고밀도 벡터로 문장을 임베딩
   (차원이 단어의 수에 의존 없이 고정)
2. 실수 표현이기 때문에 메모리 효율적



Open-Domain QA

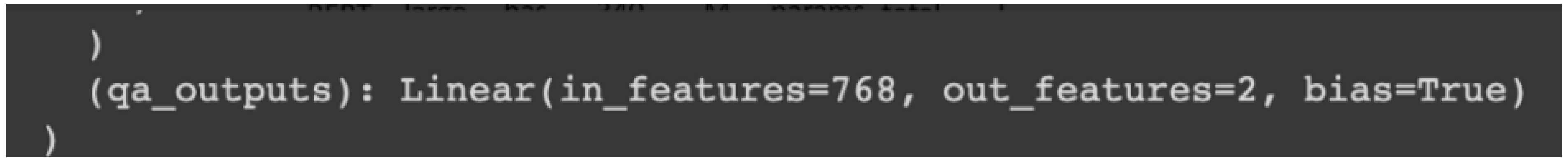
Stage-2 : Reader


Extraction-based MRC Overview
(Step 4)
Post-processing
(Prediction) Start position: 10, End position: 10
(Predicted Token) 4위
Extraction-based MRC Model
(Pre-trained language model + Classifier)
(Step 2)
Fine-tuning
C_tok1
C_tok2
...
C_tokN
Q_tok1
...
Q_tokM
Embedding
(Step 1)
Pre-process
C:
C_tok1
C_tok2
C_tok3
...
C_tokN
Q:
Q_tok1
Q_tok2
...
Q_tokM
A: Start position: 10
End position: 10
Raw Data
Context (C) ... 개최한 국제적인 도시이다. 서울의 GDP는 세계 4위이다. ...
Question (Q) 서울의 GDP는 세계 몇 위야?
Answer (A) 4위


Generation-based MRC Overview
(Step 4)
Post-processing
(Prediction) 4위
Generation-based MRC Model
(Pre-trained language model (Seq-to-seq))
(Step 2)
Fine-tuning
C_tok1
C_tok2
...
C_tokN
Q_tok1
...
Q_tokM
Embedding
(Step 1)
Pre-process
C:
C_tok1
C_tok2
C_tok3
...
C_tokN
Q:
Q_tok1
Q_tok2
...
Q_tokM
A: 4위
Raw Data
Context (C) ... 개최한 국제적인 도시이다. 서울의 GDP는 세계 4위이다. ...
Question (Q) 서울의 GDP는 세계 몇 위야?
Answer (A) 4위

PLM + classifier

Seq2Seq PLM



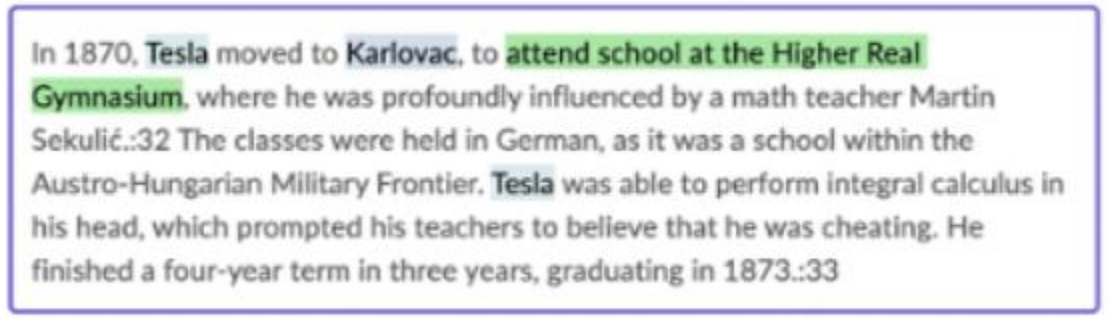
```
  )
  (qa_outputs): Linear(in_features=768, out_features=2, bias=True)
)
```

Open-Domain QA

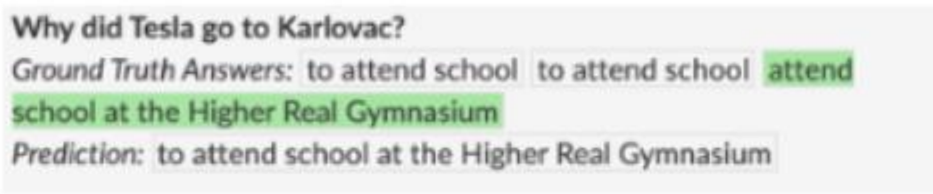
Metric

Extraction-based MRC

Exact Match (EM)



In 1870, Tesla moved to Karlovac, to attend school at the Higher Real Gymnasium, where he was profoundly influenced by a math teacher Martin Sekulić.:32 The classes were held in German, as it was a school within the Austro-Hungarian Military Frontier. Tesla was able to perform integral calculus in his head, which prompted his teachers to believe that he was cheating. He finished a four-year term in three years, graduating in 1873.:33

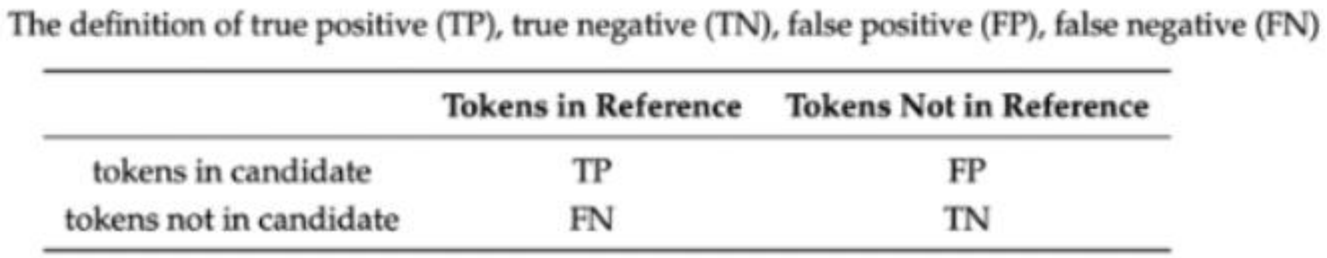


Why did Tesla go to Karlovac?

Ground Truth Answers: to attend school to attend school attend school at the Higher Real Gymnasium

Prediction: to attend school at the Higher Real Gymnasium

F1-score



The definition of true positive (TP), true negative (TN), false positive (FP), false negative (FN)

	Tokens in Reference	Tokens Not in Reference
tokens in candidate	TP	FP
tokens not in candidate	FN	TN

$$Precision = \frac{num(same\_token)}{num(pred\_tokens)}$$

$$Recall = \frac{num(same\_token)}{num(groud\_tokens)}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Generation-based MRC

ROUGE

ROUGE-L은 두 문장의 LCS (Longest Common Subsequence)를 이용해 얼마나 유사한지 측정합니다.

$Reference$: 실제 라벨

$Hypothesis$: 모델이 예측한 값

$$ROUGE - N_{recall} = \frac{\sum n\_gram_{in-match}}{\sum n\_gram_{inReference}} \quad ...(A)$$

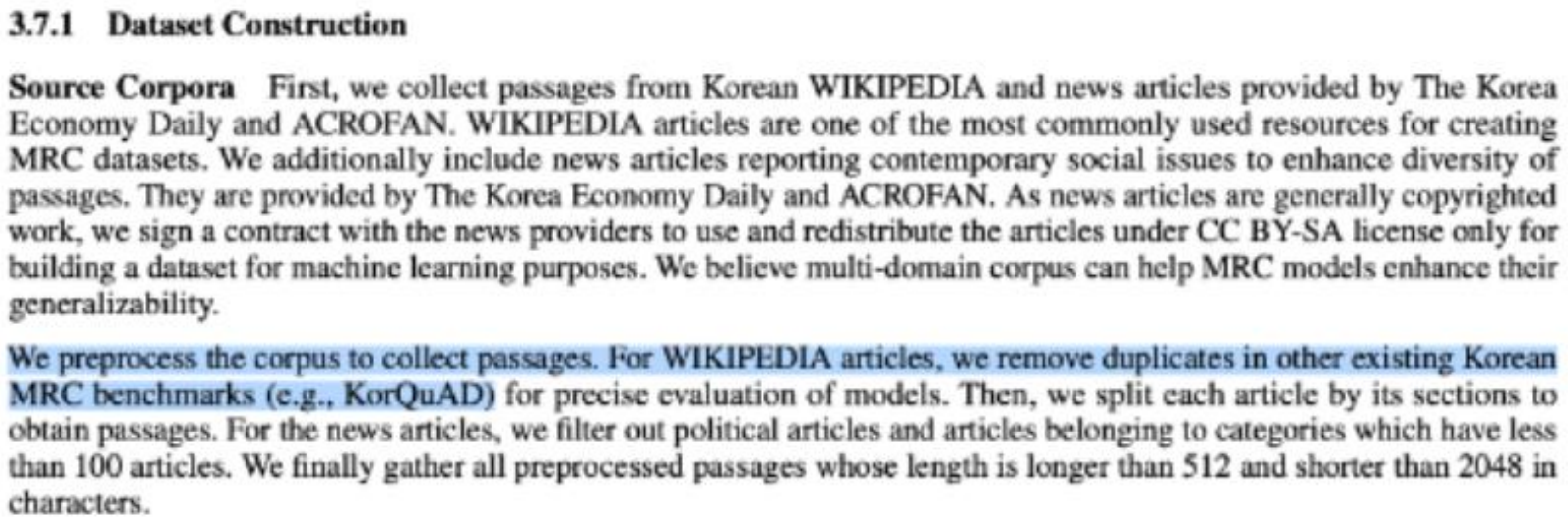
$$ROUGE - N_{precision} = \frac{\sum n\_gram_{in-match}}{\sum n\_gram_{inHypothesis}} \quad ...(B)$$

$$ROUGE - N_{F1} = \frac{2AB}{A + B} \quad ...(C)$$

What i do?

Task Define KLUE MRC 데이터셋 리뷰 (Extractive-based ODQA)

1.Source Corpora



3.7.1 Dataset Construction

Source Corpora First, we collect passages from Korean WIKIPEDIA and news articles provided by The Korea Economy Daily and ACROFAN. WIKIPEDIA articles are one of the most commonly used resources for creating MRC datasets. We additionally include news articles reporting contemporary social issues to enhance diversity of passages. They are provided by The Korea Economy Daily and ACROFAN. As news articles are generally copyrighted work, we sign a contract with the news providers to use and redistribute the articles under CC BY-SA license only for building a dataset for machine learning purposes. We believe multi-domain corpus can help MRC models enhance their generalizability.

We preprocess the corpus to collect passages. For WIKIPEDIA articles, we remove duplicates in other existing Korean MRC benchmarks (e.g., KorQuAD) for precise evaluation of models. Then, we split each article by its sections to obtain passages. For the news articles, we filter out political articles and articles belonging to categories which have less than 100 articles. We finally gather all preprocessed passages whose length is longer than 512 and shorter than 2048 in characters.

1. WIKIPEDIA와 뉴스 기사 데이터(Korea Economy Daily and ACROFAN)를 출처로 활용
2. 뉴스데이터에서 정치적 이슈를 제거하고 총 Passage의 단위를 ***512에서 2048로 한정***
3. 특이한 점은 ***KorQuAD와 겹치는 passage는 MRC데이터 셋에서 제거***한 것을 알 수 있음

2. Question type

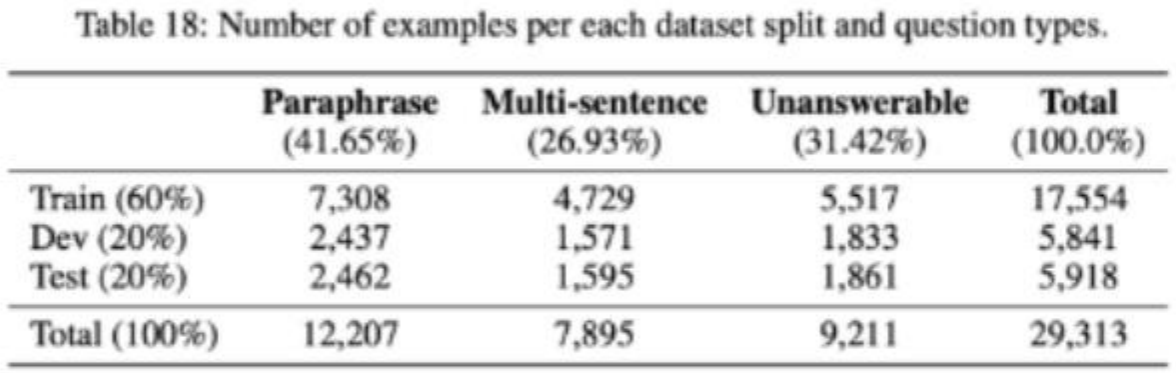
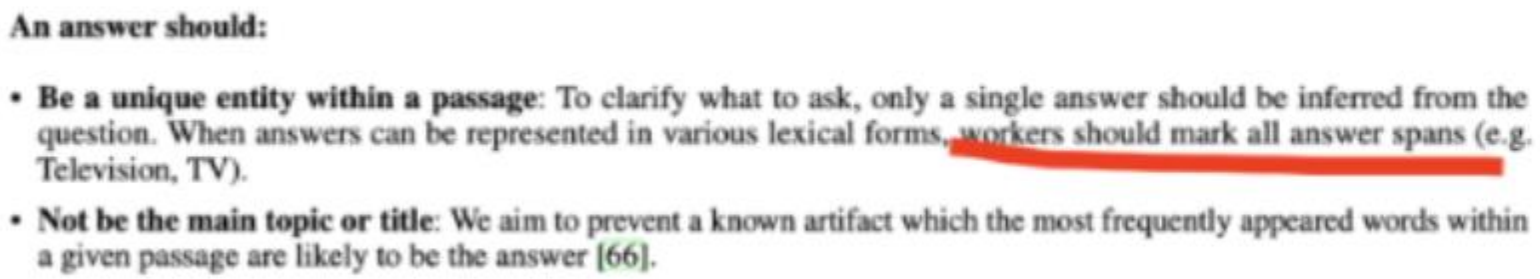


Table 18: Number of examples per each dataset split and question types.

	Paraphrase (41.65%)	Multi-sentence (26.93%)	Unanswerable (31.42%)	Total (100.0%)
Train (60%)	7,308	4,729	5,517	17,554
Dev (20%)	2,437	1,571	1,833	5,841
Test (20%)	2,462	1,595	1,861	5,918
Total (100%)	12,207	7,895	9,211	29,313

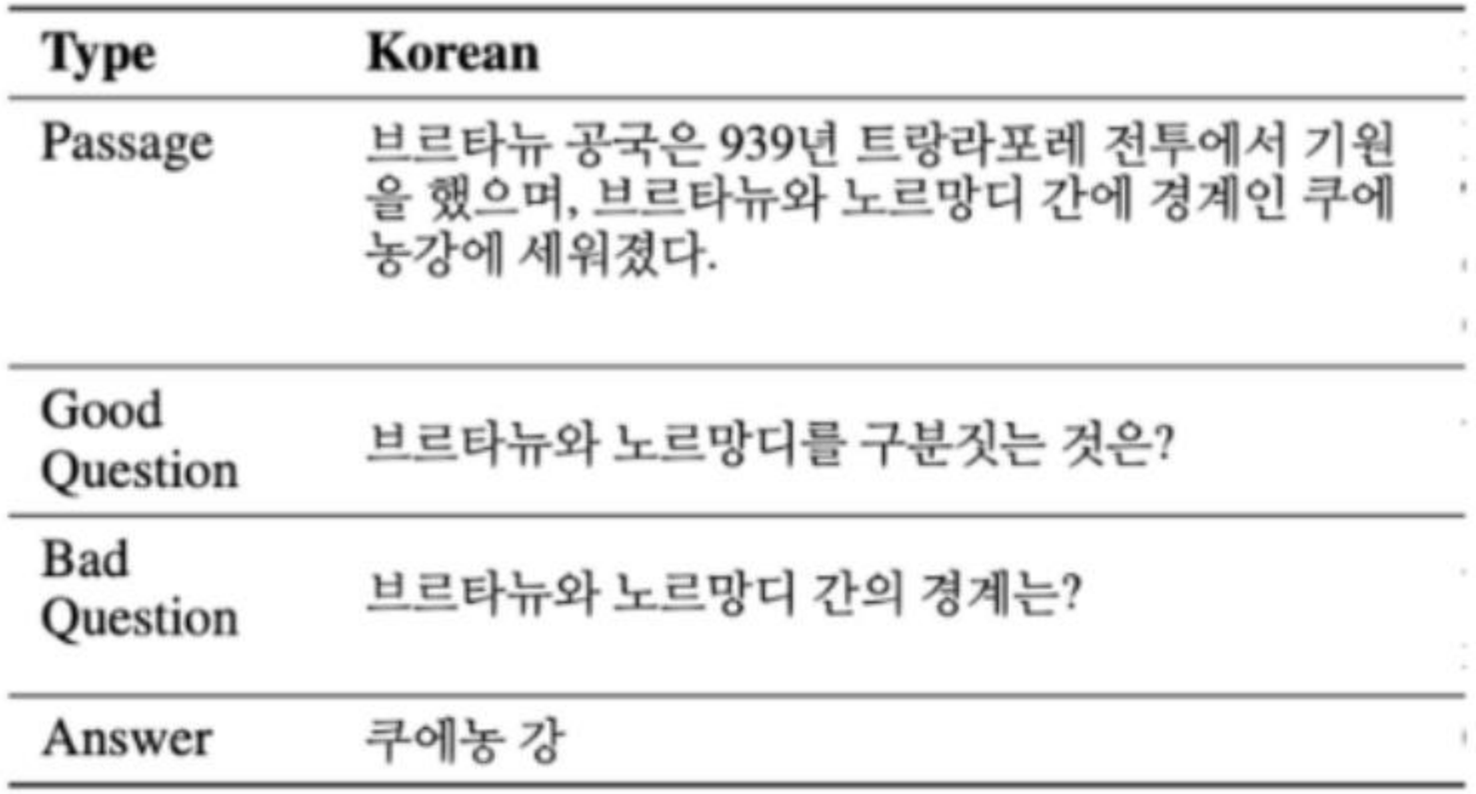
질문은 총 3가지 형식으로 구성되어 있음을 알 수 있었다. EDA를 통해 본 대회는 어떤 데이터로 구성되었을 지 확인

2. Answers type



An answer should:

- **Be a unique entity within a passage**: To clarify what to ask, only a single answer should be inferred from the question. When answers can be represented in various lexical forms, workers should mark all answer spans (e.g. Television, TV).
- **Not be the main topic or title**: We aim to prevent a known artifact which the most frequently appeared words within a given passage are likely to be the answer [66].



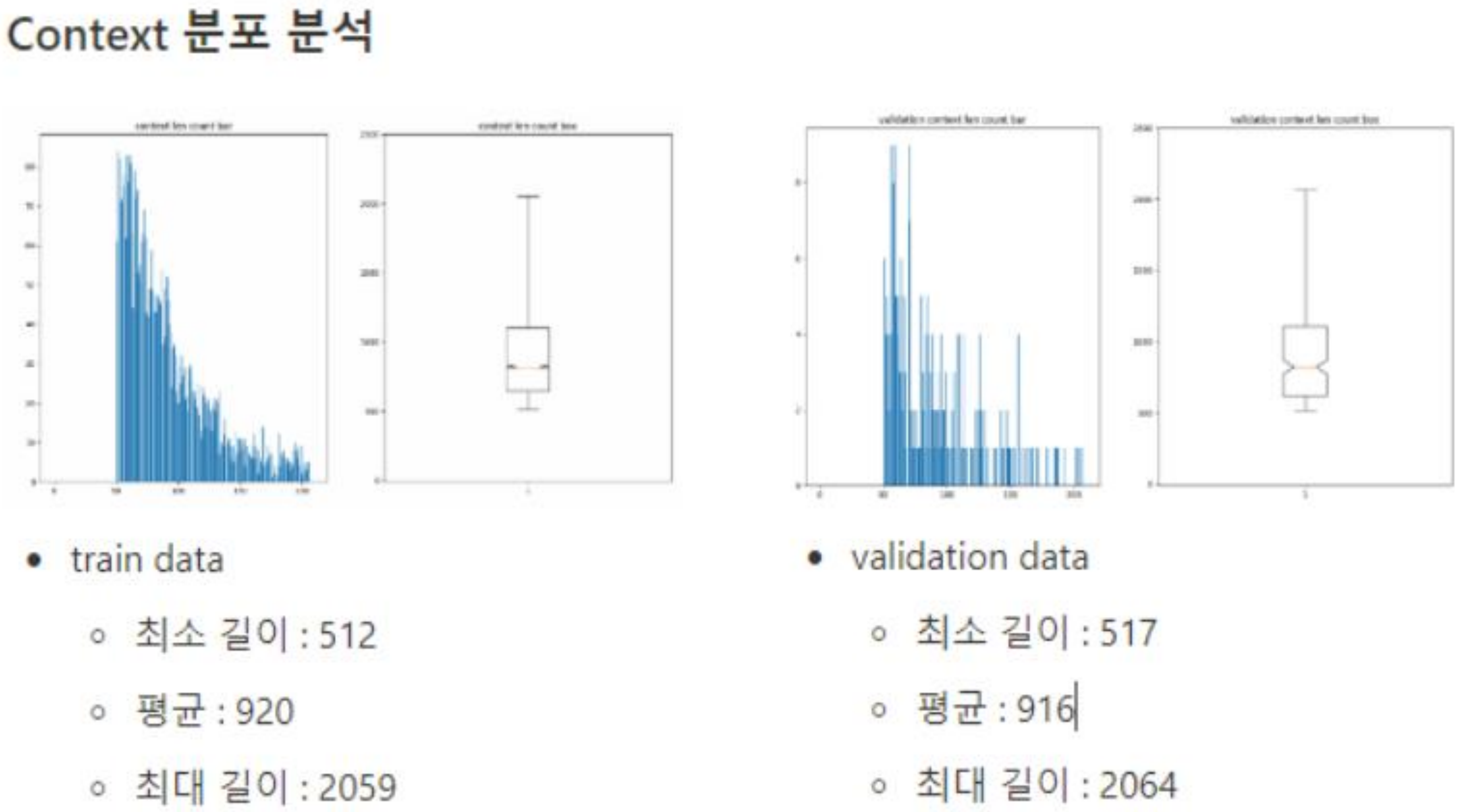
Type	Korean
Passage	브르타뉴 공국은 939년 트랑라포레 전투에서 기원을 했으며, 브르타뉴와 노르망디 간에 경계인 쿠에농강에 세워졌다.
Good Question	브르타뉴와 노르망디를 구분짓는 것은?
Bad Question	브르타뉴와 노르망디 간의 경계는?
Answer	쿠에농 강

주석자로 하여금 ***꼭 Passage안에 정답 Span이 존재***하는 질문으로 구성하도록 설계하는 것을 강조했다고 저자는 말함.

What i do?

Solution Long Context Problem

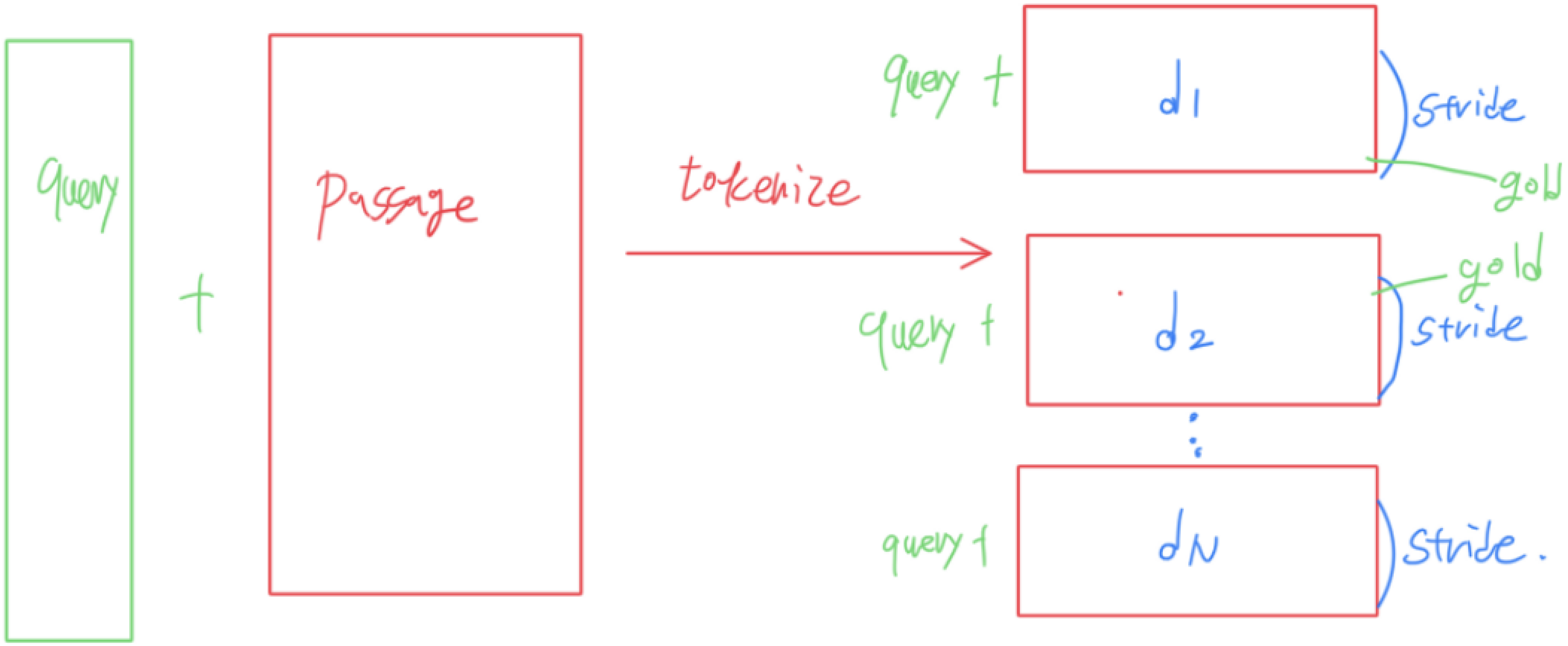
Document의 평균길이



Context 분포 분석

- train data
  - 최소 길이 : 512
  - 평균 : 920
  - 최대 길이 : 2059
- validation data
  - 최소 길이 : 517
  - 평균 : 916
  - 최대 길이 : 2064

Document의 평균길이



Korbigbird ->roberta-large변경 stride를 늘려 offset mapping을 통해 하나의 데이터를 n개로 분리

장점 : 데이터 증강 효과

<korbigbird>

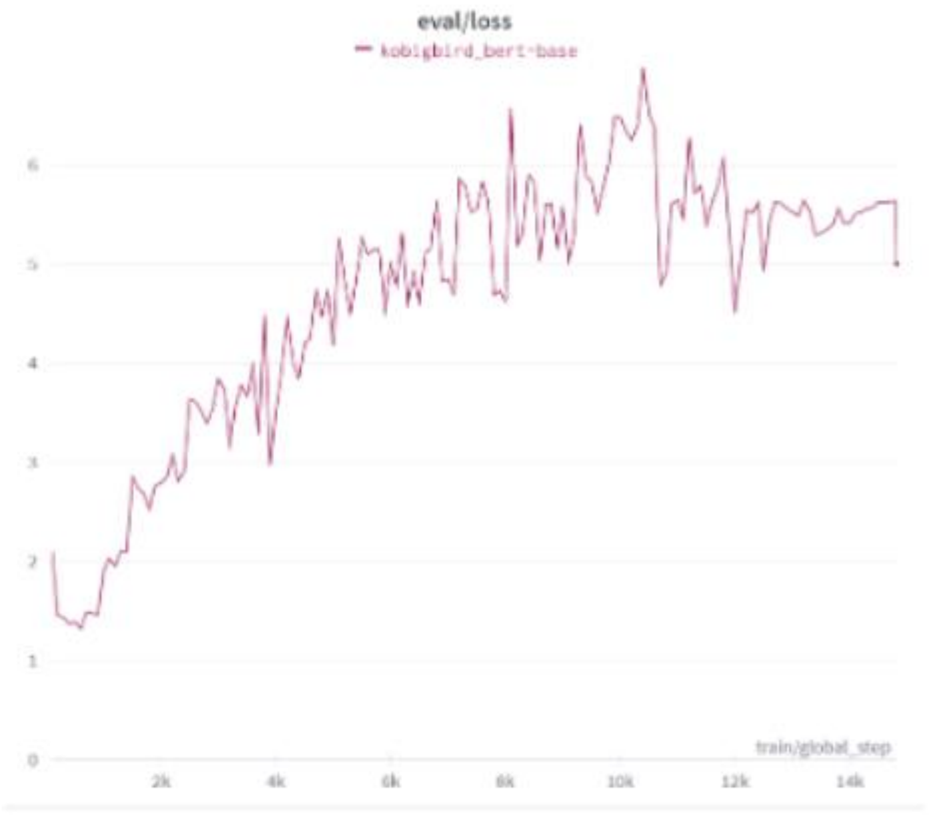
V_EM	V_F1
60.41	69.29

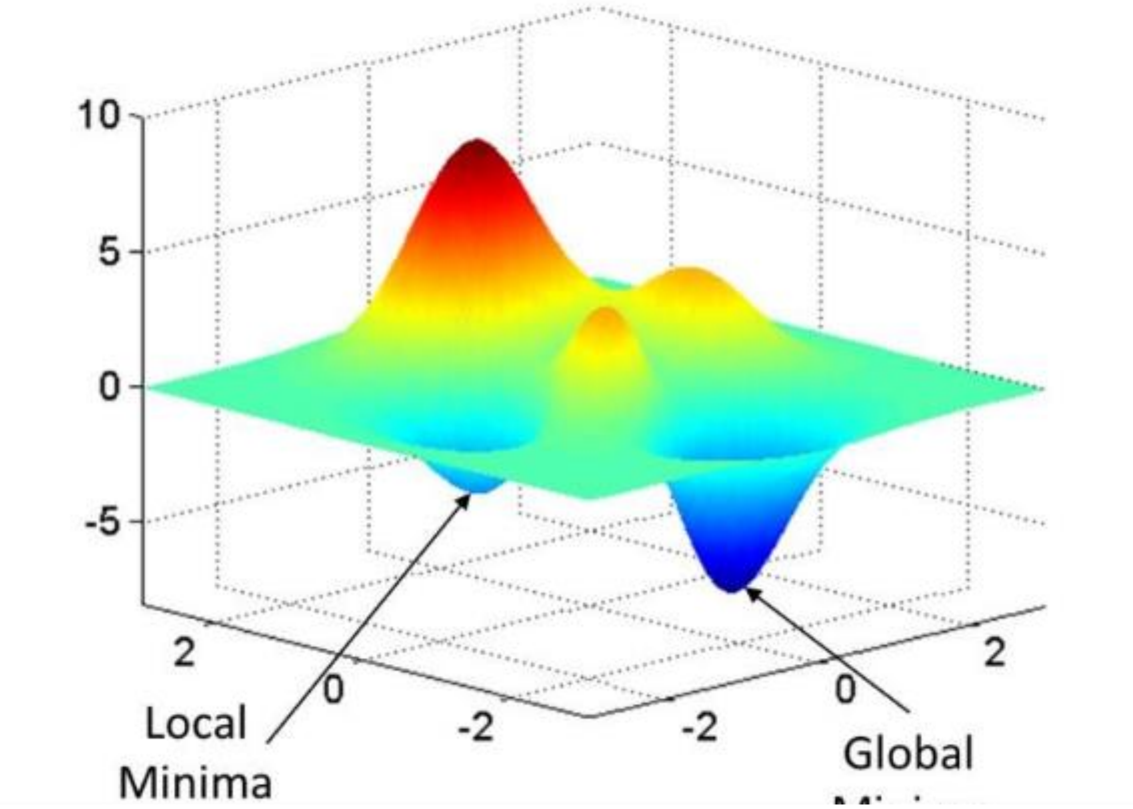
<roberta-large +256>

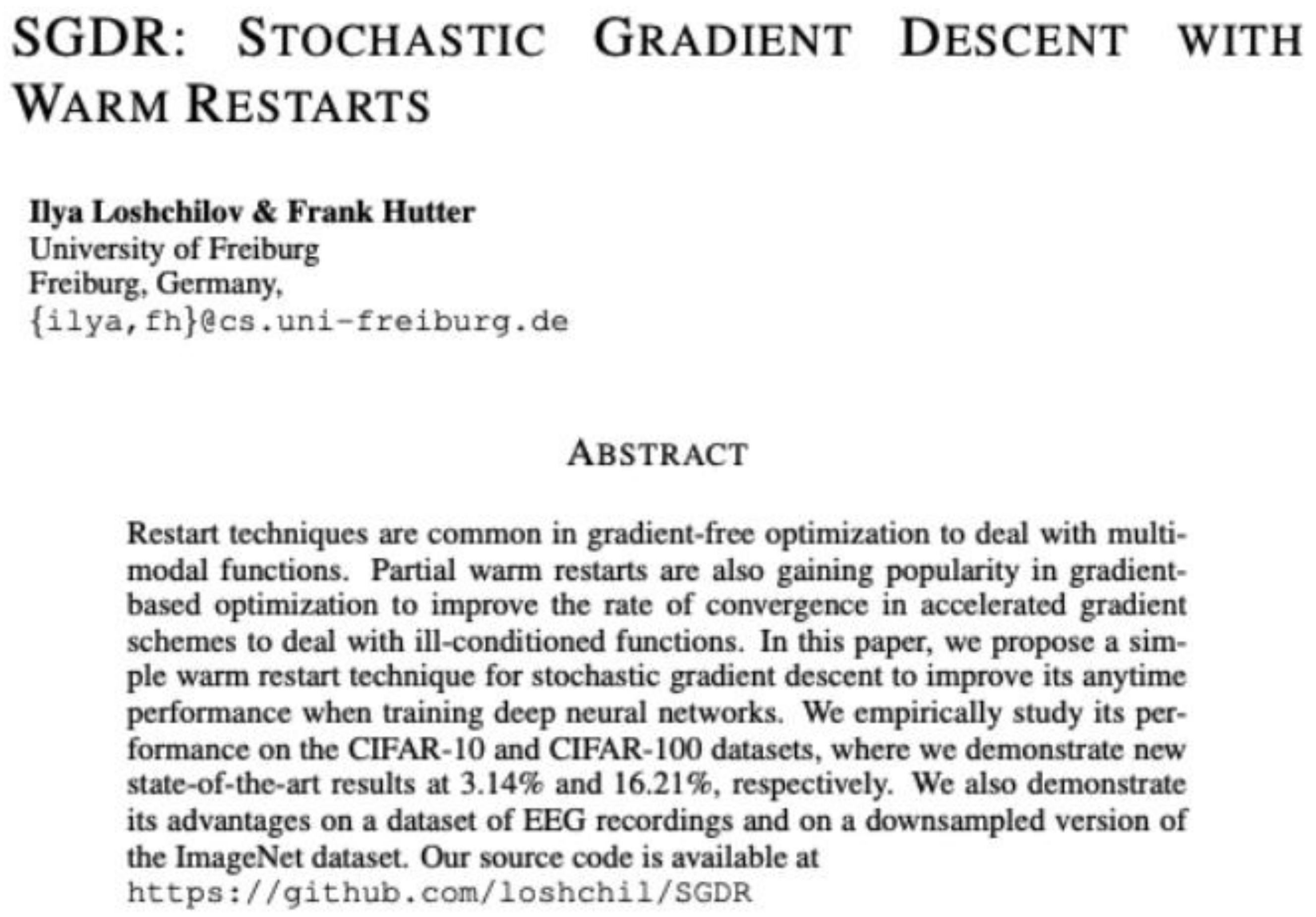
V_EM	V_F1
55->66.66	62.34->74.04

What i do?

Solution 일반화 성능 문제







SGDR: Stochastic Gradient Descent with Warm Restarts

Ilya Loshchilov & Frank Hutter
University of Freiburg
Freiburg, Germany,
{ilya,fh}@cs.uni-freiburg.de

Abstract

Restart techniques are common in gradient-free optimization to deal with multimodal functions. Partial warm restarts are also gaining popularity in gradient-based optimization to improve the rate of convergence in accelerated gradient schemes to deal with ill-conditioned functions. In this paper, we propose a simple warm restart technique for stochastic gradient descent to improve its anytime performance when training deep neural networks. We empirically study its performance on the CIFAR-10 and CIFAR-100 datasets, where we demonstrate new state-of-the-art results at 3.14% and 16.21%, respectively. We also demonstrate its advantages on a dataset of EEG recordings and on a downsampled version of the ImageNet dataset. Our source code is available at
https://github.com/loshchil/SGDR

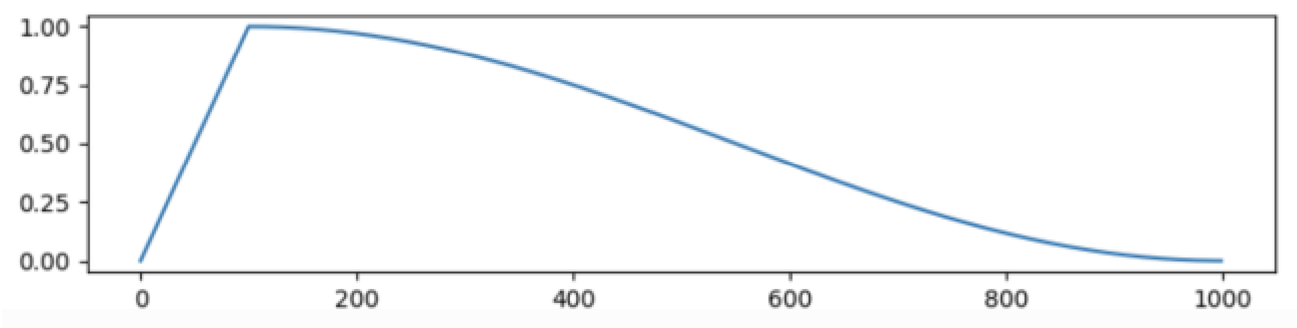
Train loss 감소 but evaluation loss 증가
Local minimum에 수렴 했다고 판단.

Cosine annealing Learning rate를 빠르게
변화를 주어 Local minimum에 빠지지 않도록

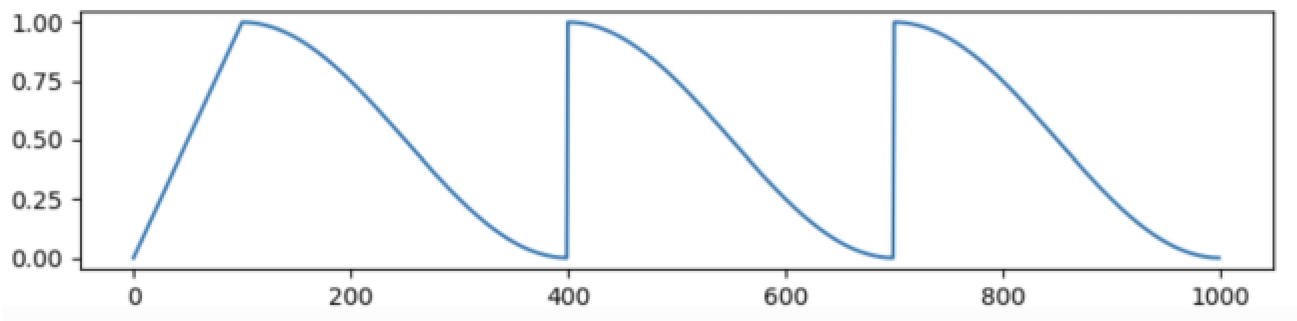
What i do?

Solution　　일반화 성능 문제

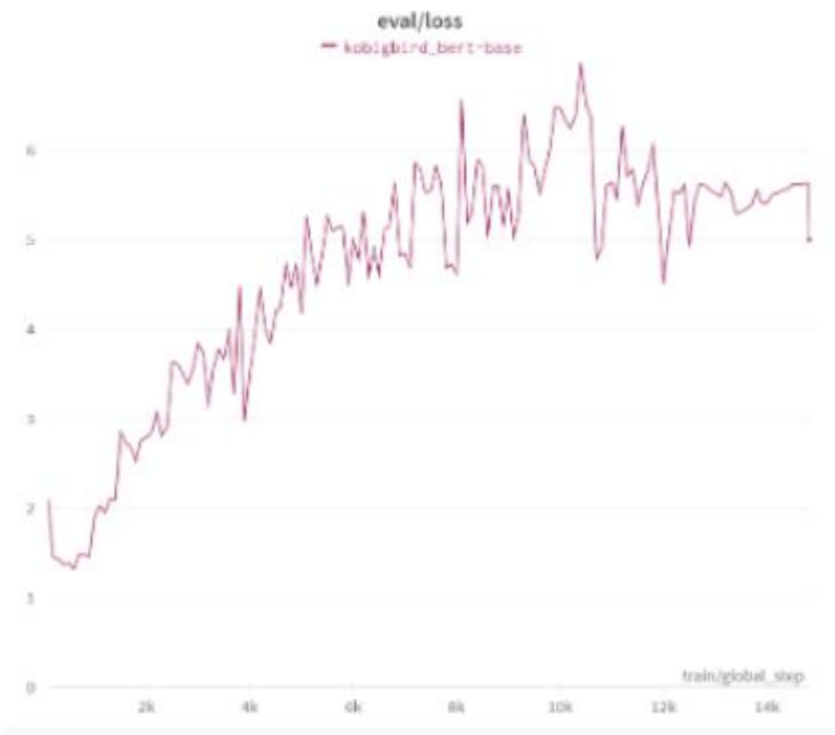
lr_scheduler 방법론 회고



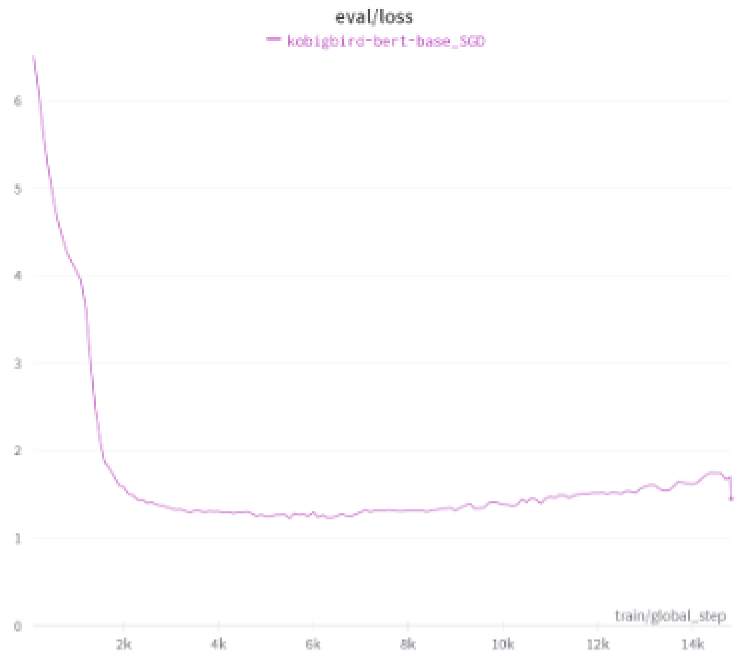
max_lr scheduler



CosineAnnealingLR scheduler

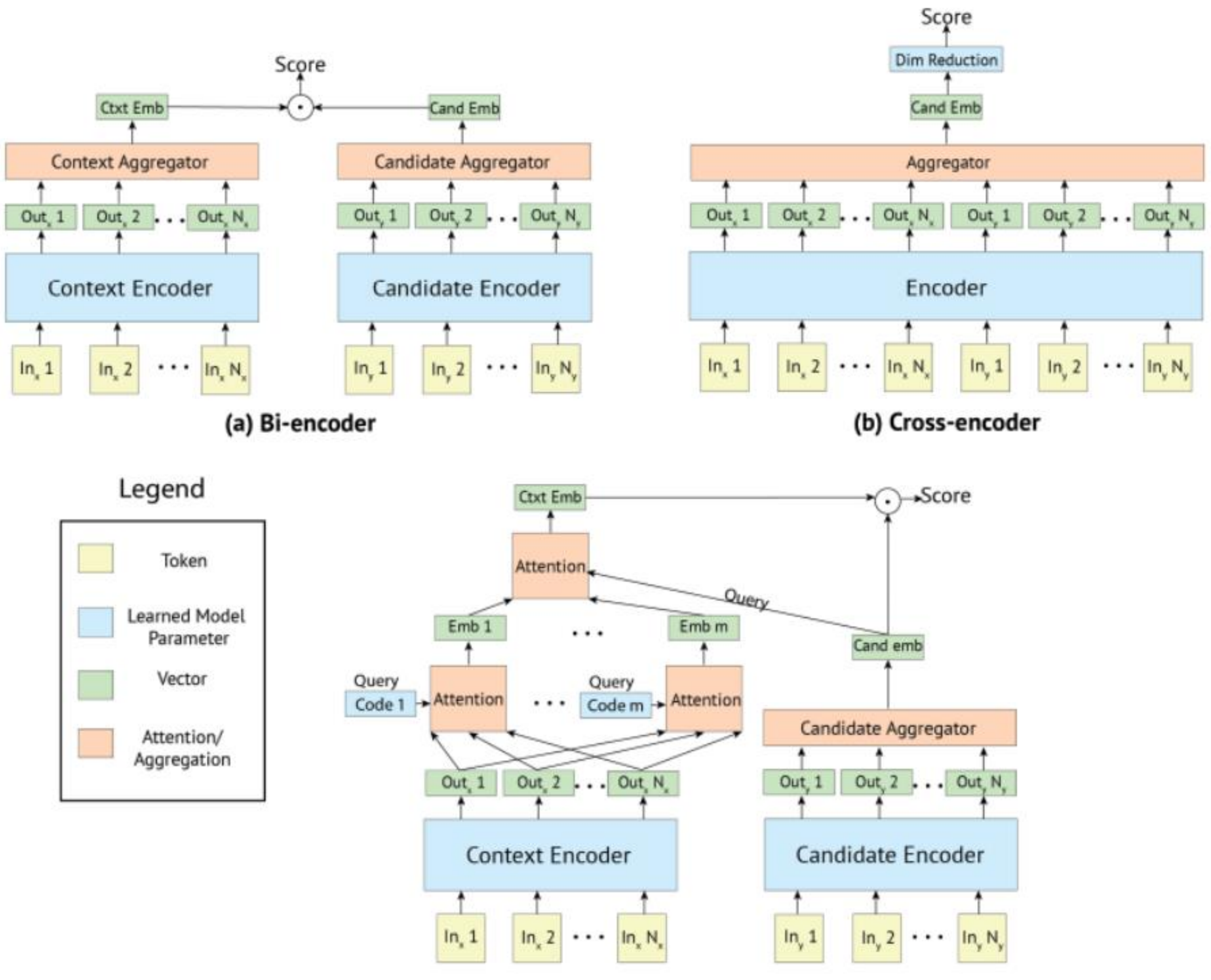


V_EM	V_F1
66.66->74.16	74.04->83.24



What i do?

Solution Poly-encoder

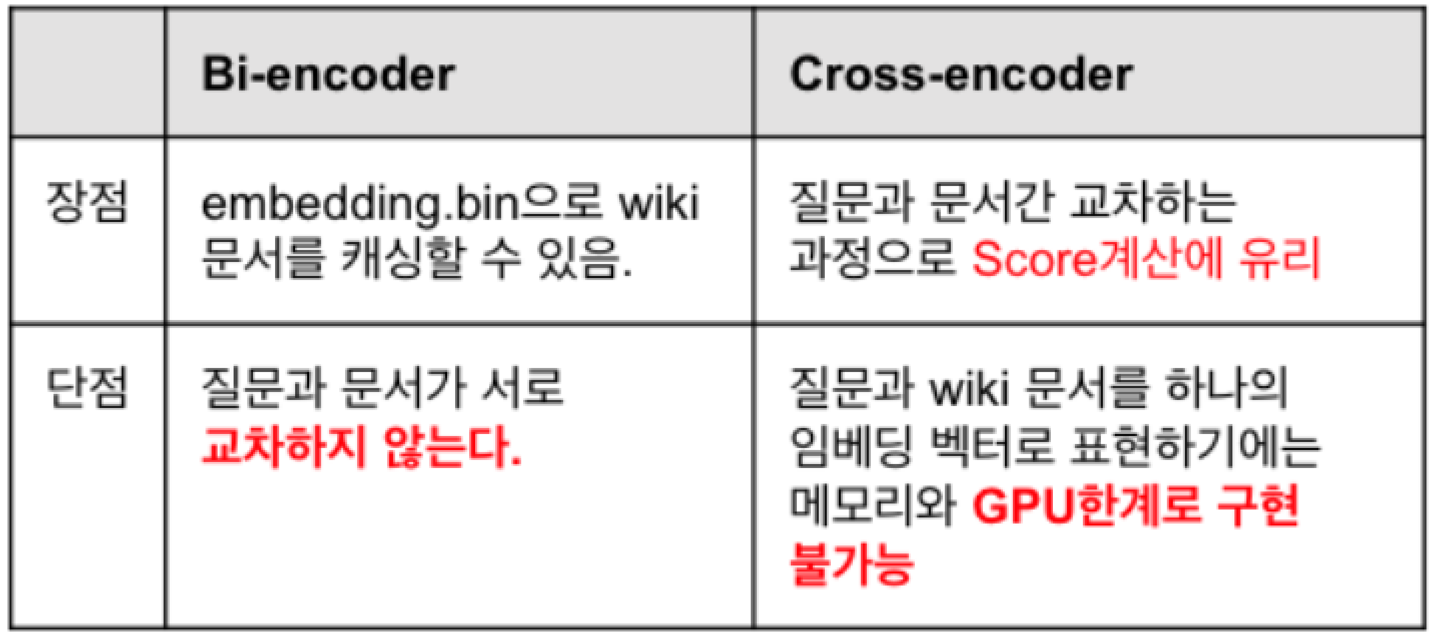


(a) Bi-encoder

(b) Cross-encoder

(c) Poly-encoder

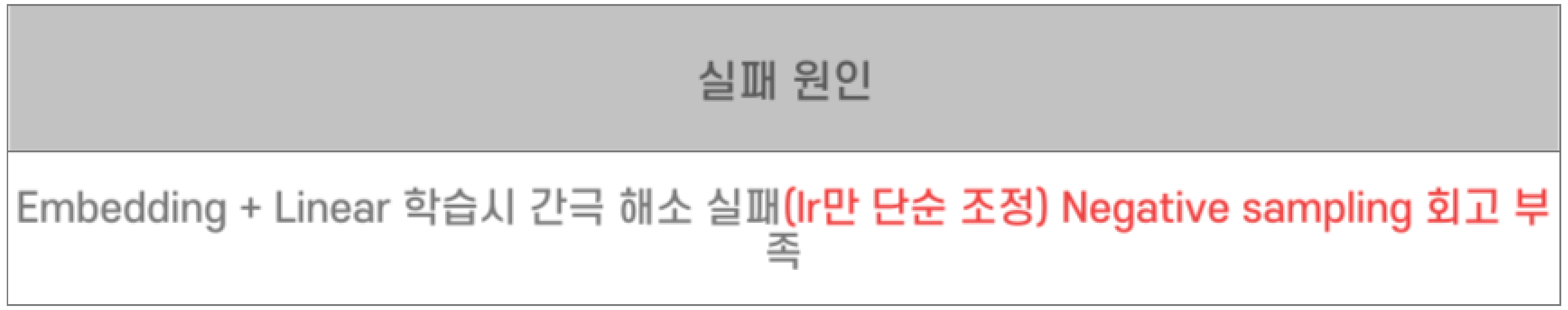
기존 Dense Retrieval의 구조를 비교



	Bi-encoder	Cross-encoder
장점	embedding.bin으로 wiki 문서를 캐싱할 수 있음.	질문과 문서간 교차하는 과정으로 Score계산에 유리
단점	질문과 문서가 서로 **교차하지 않는다.**	질문과 wiki 문서를 하나의 임베딩 벡터로 표현하기에는 메모리와 **GPU한계로 구현 불가능**

code m의 Learnable params를 추가

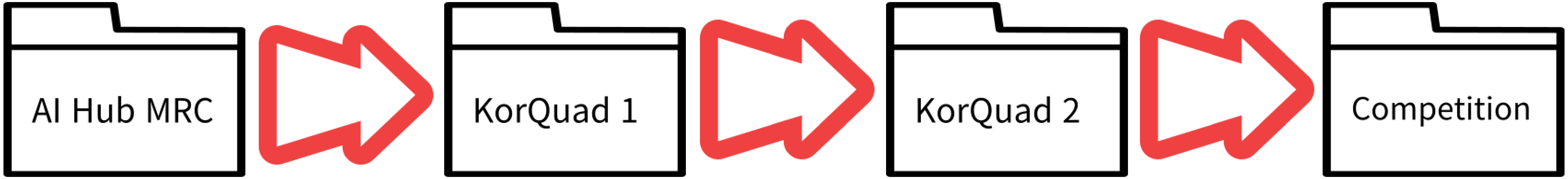
Positive, Negative pair를 만들어
code m은 벡터상 거리를 조정 Negative log likelihood



실패 원인
Embedding + Linear 학습시 간극 해소 실패(lr만 단순 조정) Negative sampling 회고 부족

What i do?

Solution 데이터 추가

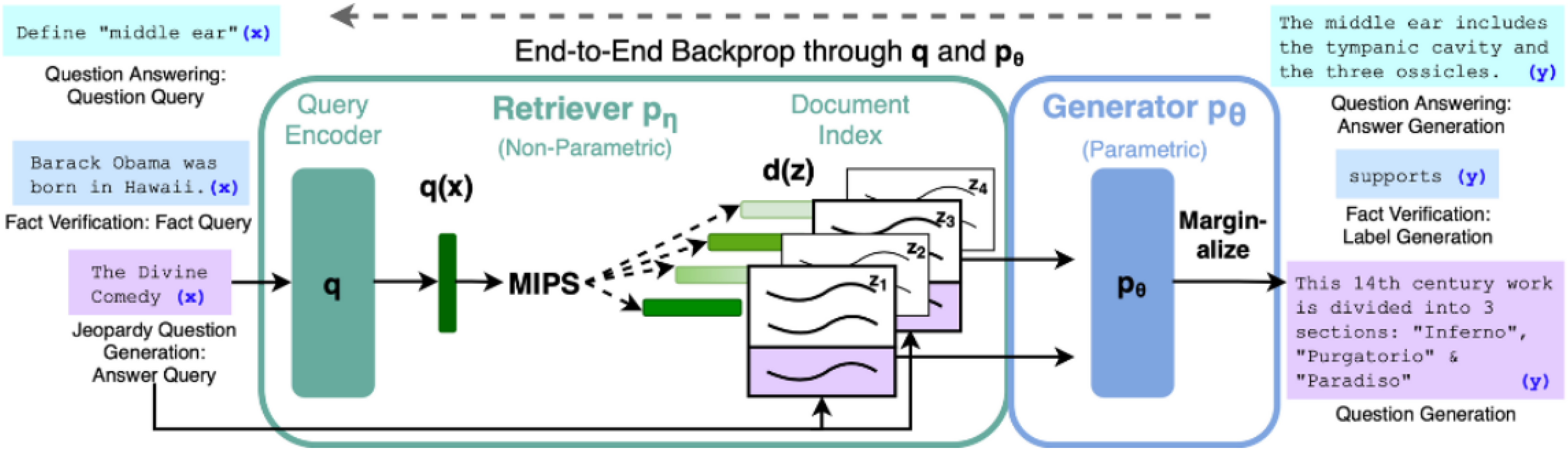


데이터 추가 시 반복해서 모델 저장 및 다시 로드 하는 방법으로 진행
대회 데이터인 KLUE/MRC 데이터는 KorQuad와 거의 동일한 데이터라
판단 및 마지막 순서로 진행

L_EM	L_F1	V_EM	V_F1
62.5	75.65	77.17	85.61

Adapt to Fact-verification

1. RAG
   a. Retrieval + Generator
   b. 입력이 Answer,Document set인 Question Generation도 가능
   c. claim generation으로 응용 가능성


Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
End-to-End Backprop through q and pθ
Query Encoder
Retriever pη
(Non-Parametric)
Document Index
q
q(x)
MIPS
d(z)
z4
z3
z2
z1
Generator pθ
(Parametric)
pθ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation

2. ColBERT
   a. token 단위 유사도 검색을 진행 기존 TF-IDF와 유사한 성능도 확보
   b. bi-encoder 구조로 추론 속도 또한 확보

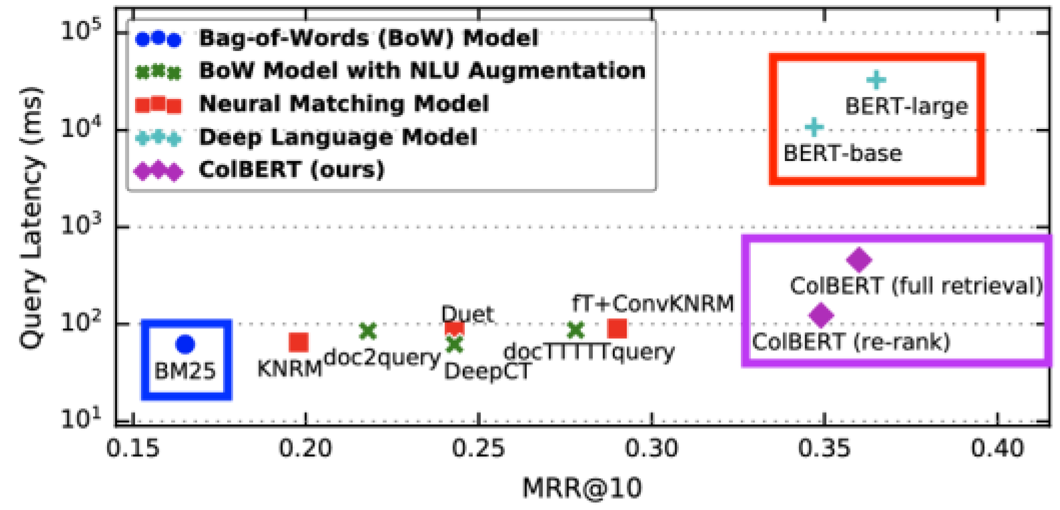
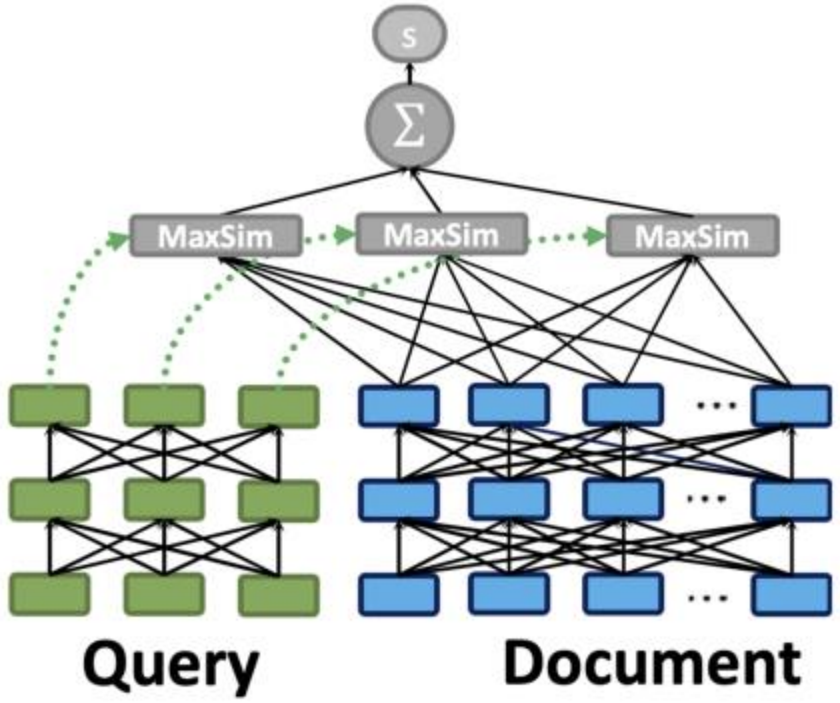


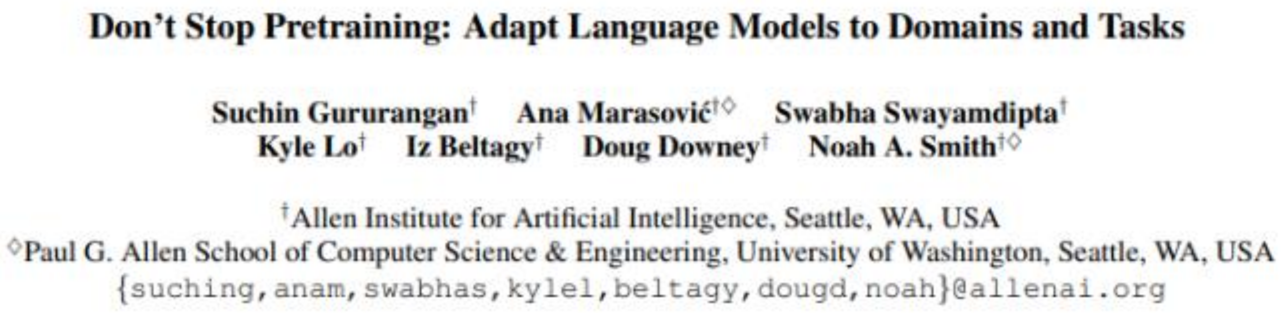
Figure 1: Effectiveness (MRR@10) versus Mean Query Latency (log-scale) for a number of representative ranking models on MS MARCO Ranking [24]. The figure also shows ColBERT. Neural re-rankers run on top of the official BM25 top-1000 results and use a Tesla V100 GPU. Methodology and detailed results are in §4.



(d) Late Interaction
(i.e., the proposed ColBERT)

Adapt to Fact-verification

3. DAPT(Domain adaptive Pretraining Task)
   a. LM을 wikipedia set(Document)에서 다시 학습
   b. 대용량 Corpus(wikipedia)를 기준으로 LM 재생성
   c. Domain이 화학, 법률 등으로 특이한 곳에 더 좋은 성능을 보임



Don't Stop Pretraining: Adapt Language Models to Domains and Tasks

Suchin Gururangan[†] Ana Marasović[†◇] Swabha Swayamdipta[†]
Kyle Lo[†] Iz Beltagy[†] Doug Downey[†] Noah A. Smith[†◇]

[†]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
{suching,anam,swabhas,kylel,beltagy,dougd,noah}@allenai.org

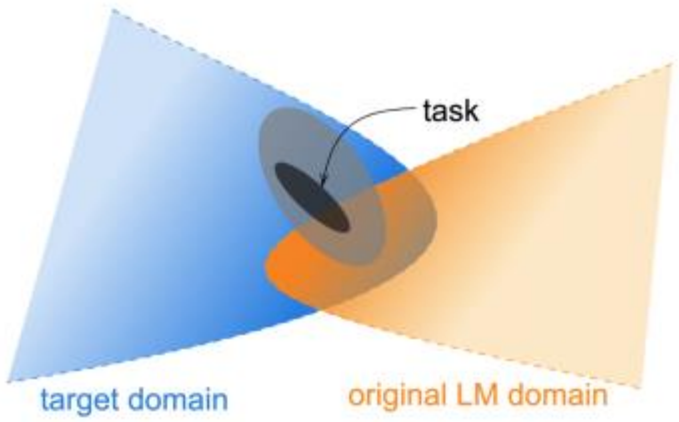
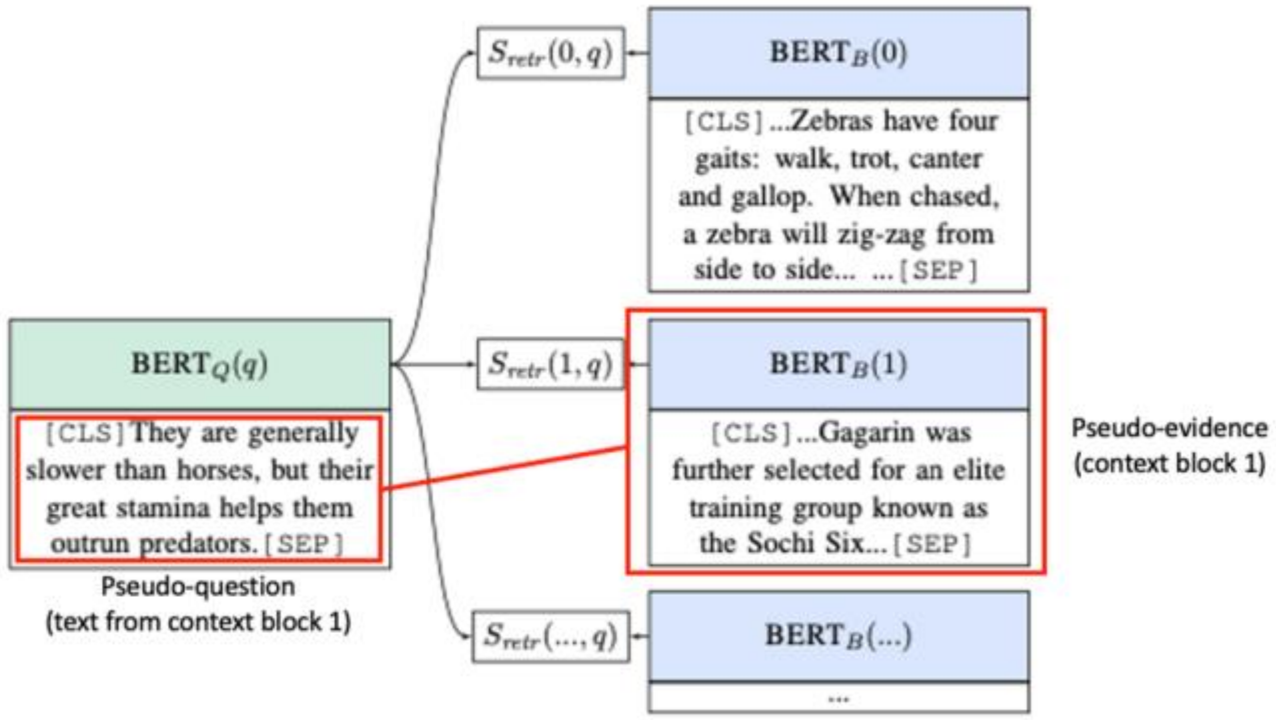


Figure 1: An illustration of data distributions. Task data is comprised of an observable task distribution, usually non-randomly sampled from a wider distribution (light grey ellipsis) within an even larger target domain, which is not necessarily one of the domains included in the original LM pretraining domain – though overlap is possible. We explore the benefits of continued pretraining on data from the task distribution and the domain distribution.

4. OrQA ICT training
   a. 문장 임베딩의 성능을 높이기 위한 방법론
   b. sudo-question을 만들어 Retrieval학습



[CLS] They are generally slower than horses, but their great stamina helps them outrun predators. [SEP]

[CLS] ...Zebras have four gaits: walk, trot, canter and gallop. When chased, a zebra will zig-zag from side to side... ...[SEP]

[CLS] ...Gagarin was further selected for an elite training group known as the Sochi Six... [SEP]

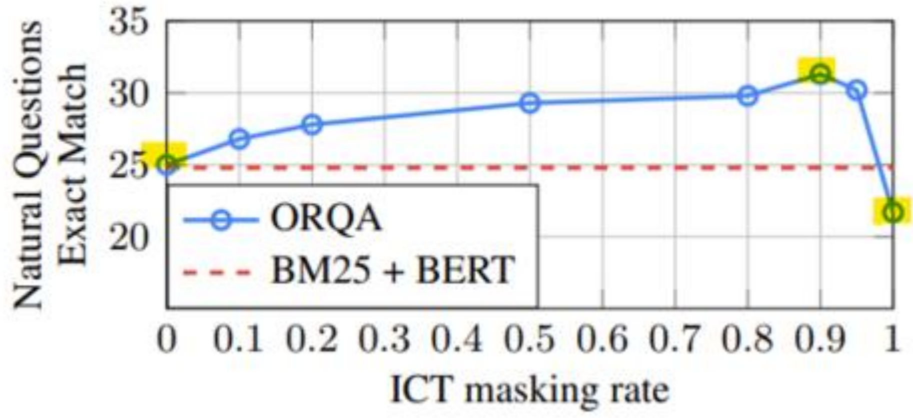


Figure 3: **Analysis**: Performance on our open version of the Natural Questions dev set with various masking rates for the ICT pre-training. Too much masking prevents the model from learning to exploit exact n-gram overlap. Too little masking makes language understanding unnecessary.

감사합니다