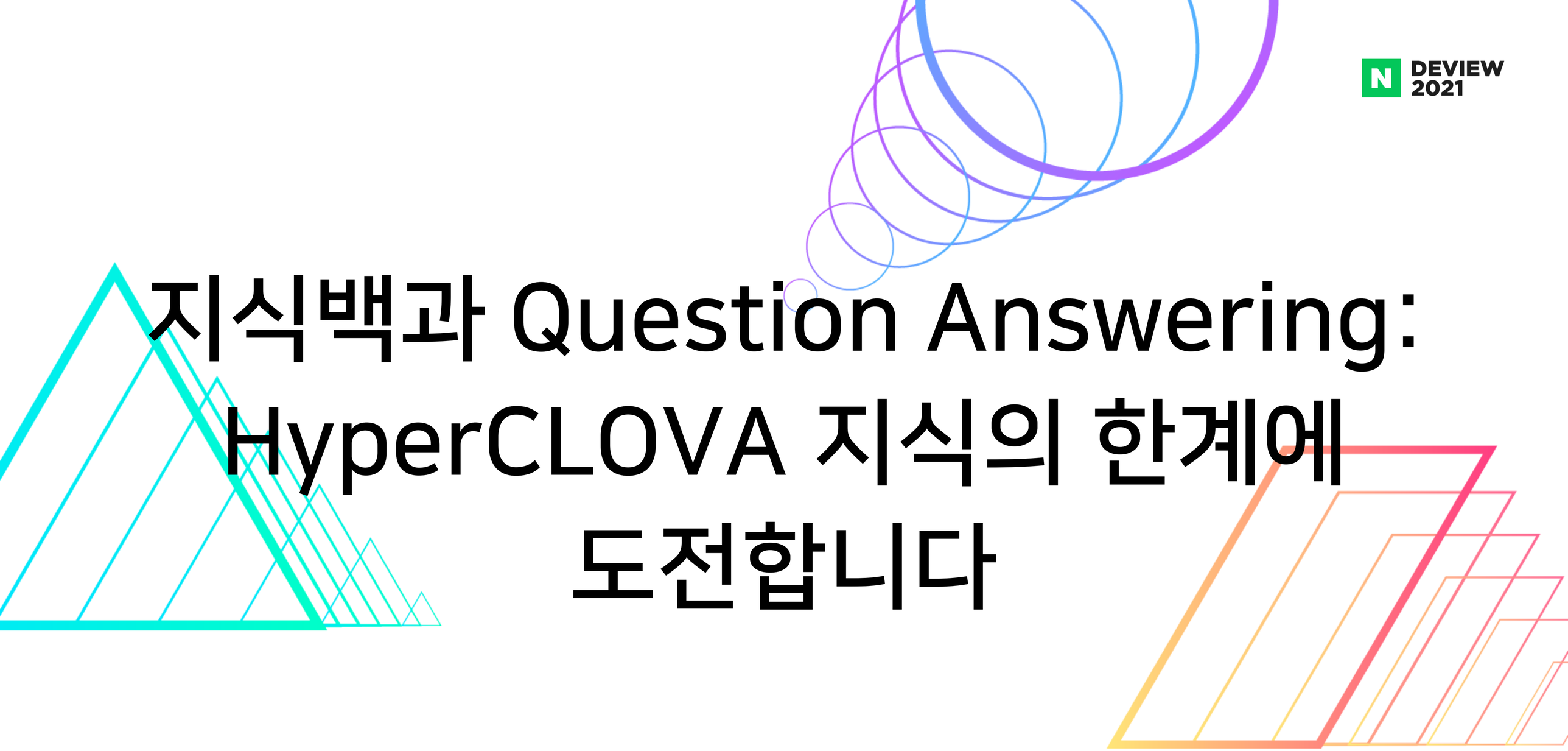


지식백과 Question Answering: HyperCLOVA 지식의 한계에 도전합니다

전동현 NAVER

CONTENTS

1. 지식백과 기반 Open-domain Question Answering

1.1 Open-domain Question Answering

DEVIEW 2021

Subject+Predicate question

- 한글 만든 사람이 누구인가요?

Multi-hop question

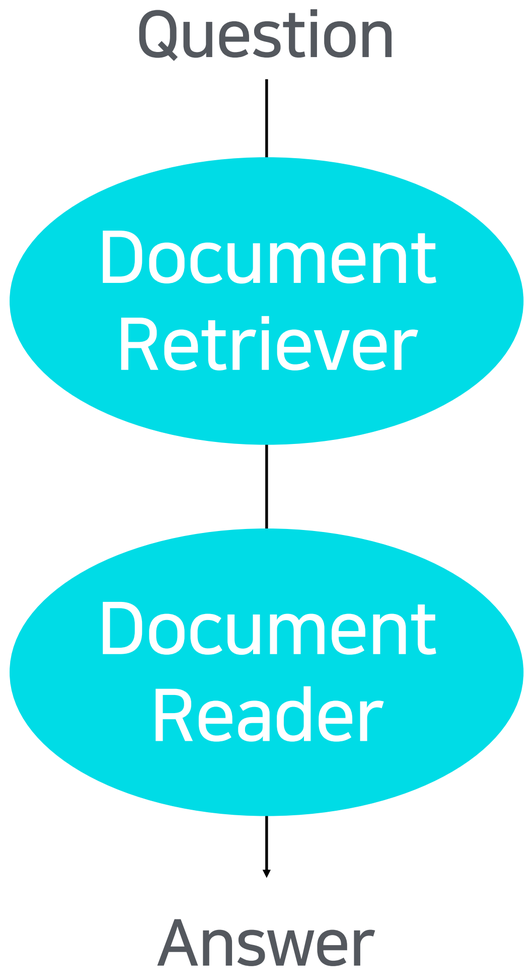
- 미국 대통령 중 그래미 상을 받은 사람은?

Boolean question

- 해파리는 뇌가 있어?

Long-form question

- 시서스가 다이어트에 왜 도움이 되니?



1.1 Open-domain Question Answering

QA의 어려움

- 검색과 대화의 사이 그 어딘가
- 문서 검색도 잘해야해
- 문서 이해도 잘해야해

품질은 아직도 50~70점

Model	top-k	NQ		TriviaQA		WebQ		*params*
		dev	test	dev	test	dev	test	
Parametric Models								
T5-*base* (Roberts *et al.*, 2020)	0	-	25.7	-	24.2	-	28.2	220M
T5-*large* (Roberts *et al.*, 2020)	0	-	27.3	-	28.5	-	29.5	770M
T5-*XXL* (Roberts *et al.*, 2020)	0	-	32.8	-	42.9	-	35.6	11B
GPT-3 (Brown *et al.*, 2020)	0	-	29.9	-	-	-	41.5	175B
Semi-Parametric Models								
BM25 + BERT (Lee *et al.*, 2019)	5	24.8	26.5	47.2	47.1	27.1	21.3	220M
ORQA (Lee *et al.*, 2019)	5	31.3	33.3	45.1	45.0	36.8	30.1	330M
REALM (Guu *et al.*, 2020)	5	38.2	40.4	-	-	-	40.7	330M
DPR (Karpukhin *et al.*, 2020)	25	-	41.5	-	56.8	-	34.6	330M
RECONSIDER (Iyer *et al.*, 2021)†	30	-	43.1	-	59.3	-	44.4	440M
RAG-Sequence (Lewis *et al.*, 2020b)†	50	44.0	44.5	55.8	56.8	44.9	45.2	626M
Individual Top-k (Sachan *et al.*, 2021)	-	-	45.9	-	56.3	-	-	440M
Joint Top-k (Sachan *et al.*, 2021)	50	-	49.2	-	64.8	-	-	440M
FiD (Izacard and Grave, 2021b)	100	-	48.2	-	65.0	-	-	440M
FiD-KD (Izacard and Grave, 2021a)	100	48.0	49.6	68.6	68.8	-	-	440M
Our Implementation								
T5-*base*	0	26.0	25.1	26.7	27.8	31.0	32.4	220M
FiD (MSS retriever, MSS reader)	50	38.5	40.1	60.0	59.8	39.1	40.2	440M
FiD (DPR retriever, T5 reader)	50	47.3	48.3	65.5	66.3	46.0	45.2	440M
EMDR^2 (MSS retriever, MSS reader)	50	**50.4**	**52.5**	**71.1**	**71.4**	**49.9**	**48.7**	440M

*End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering [Sachan et al. '21]

1.1 Open-domain Question Answering

DEVIEW 2021

HyperCLOVA의 내부 Knowledge로 답해보자!

GPT3는 In-context learning을 잘한다고..

Few-shot for HyperCLOVA QA

Task description → 설명: 질문에 대한 답변을 하세요.

Examples
- 질문: 조선의 제1대 왕은?
- 답변: 이성계
- 질문: 한글을 창제한 조선의 왕은?
- 답변: 세종
- ...

질문: 한글 만든 사람이 누구야?

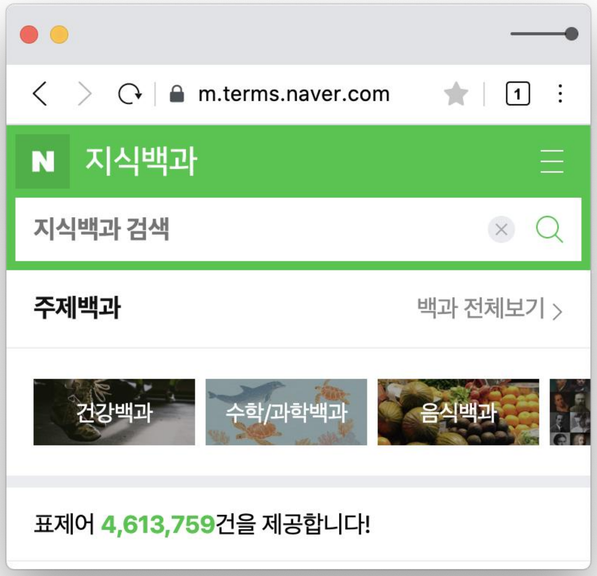
	HyperCLOVA-39B
#fewshot	HistoryQA-F1
0	30.9
1	42.6
5	50.1
10	50.2

...?

1.2 지식백과 문서 기반 Open-domain QA

DEVIEW 2021

Q: 한글 만든 사람이 누구인가요?



Retriever

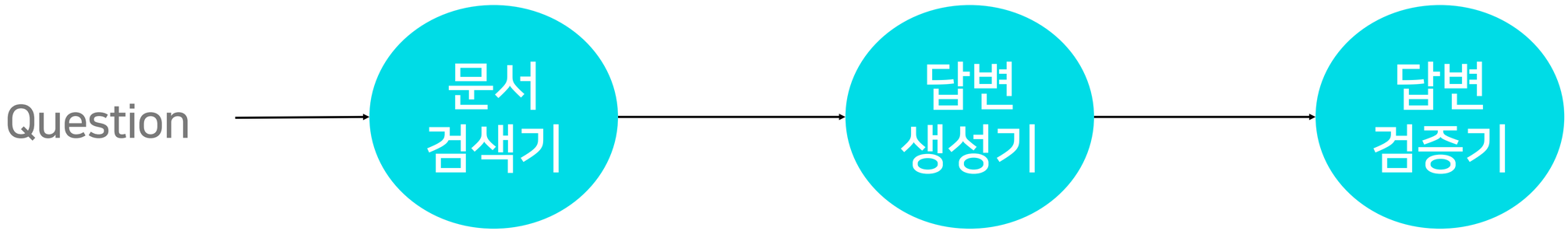
Top-k Documents

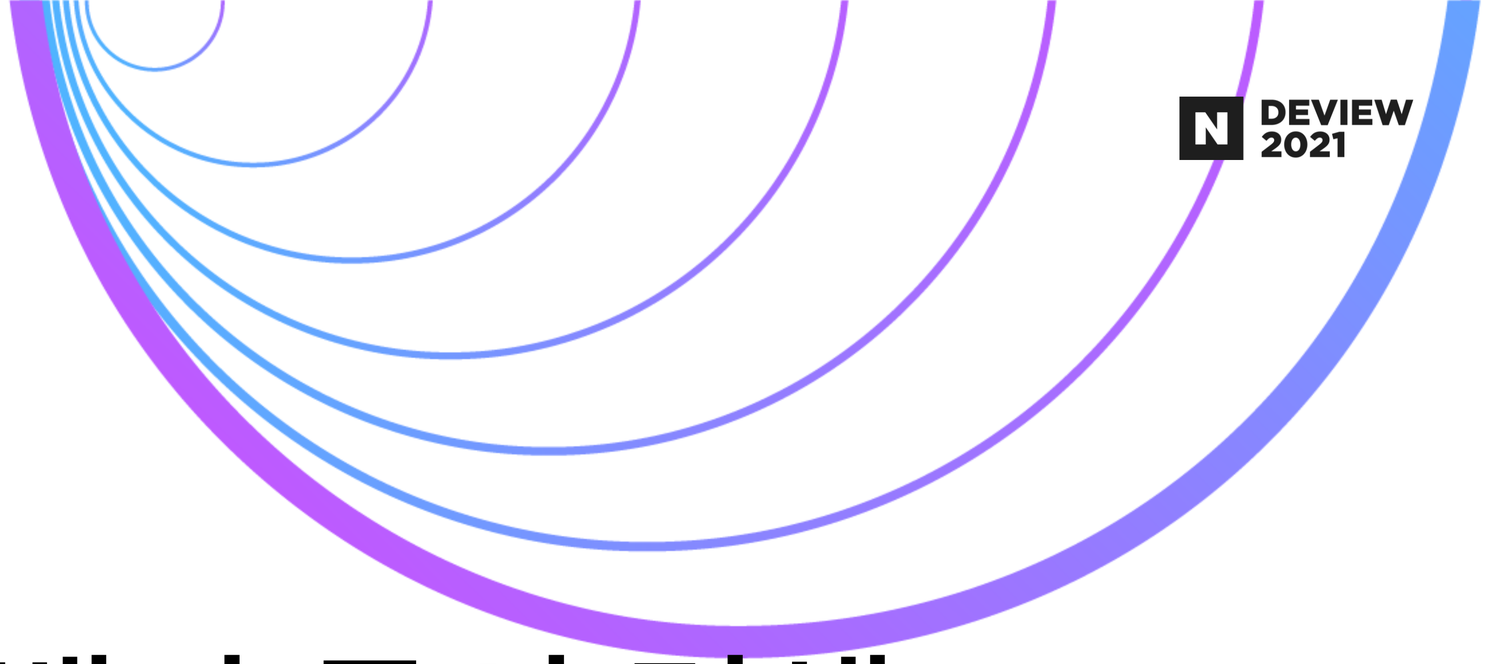
Reader

A: 세종대왕

네이버 지식백과 문서 약 480만 건 (고품질) vs 한국어 위키백과 약 55만건

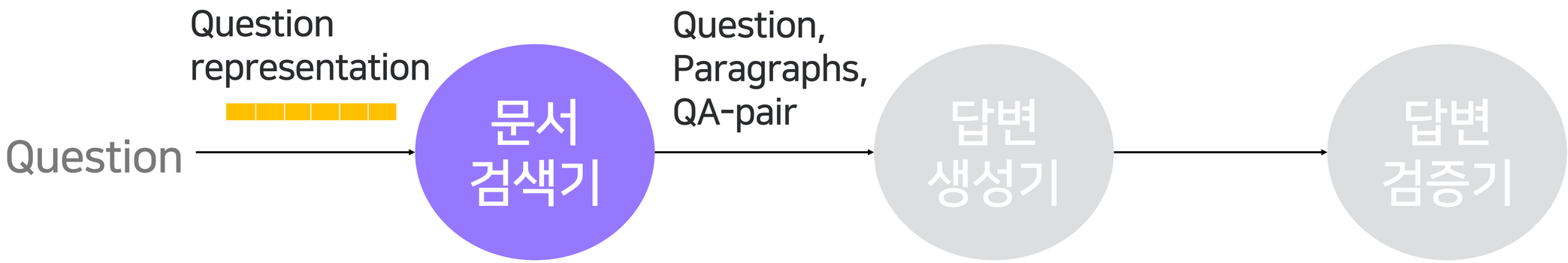
1.2 지식백과 문서 기반 Open-domain QA





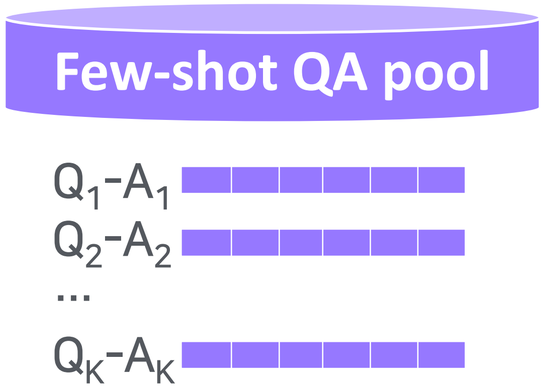
2.지식백과 문서 검색

2.1 지식백과 문서 검색



Paragraph-level Information Retrieval

Few-shot QA example Search



2.2 지식백과 문서 Paragraph-level IR

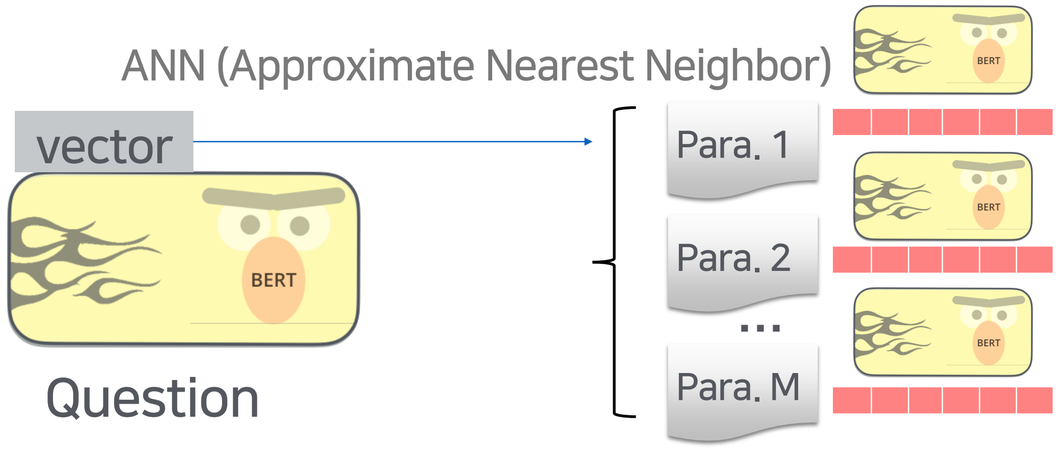
질문에 대해 텀매칭 기반 점수와 시맨틱 기반 점수를 사용해 검색된 문서들을 랭킹

1. Term-based search with BM25

텀매칭 기반의 relevance 점수 추출

2. Semantic search with BERT based Bi-Encoder

검색될 문단의 임베딩 벡터들은 BERT로 미리 색인
질문 벡터만 실시간으로 생성하여 유사도 계산



2.2 지식백과 문서 Paragraph-level IR

DEVIEW 2021

그런데 문서가 너무 길다!

→ 지식백과 문서를 적당한 paragraph 로 쪼개기

1) 250글자 길이로 청킹
2) 3문장 단위로 청킹
3) 100-word 청킹
4) 문단 단위 청킹. 문단이 긴 경우 Byte-level BPE 길이 기준으로 청킹

*Dense passage retrieval for open-domain question answering [Karpukhin et al. '20]

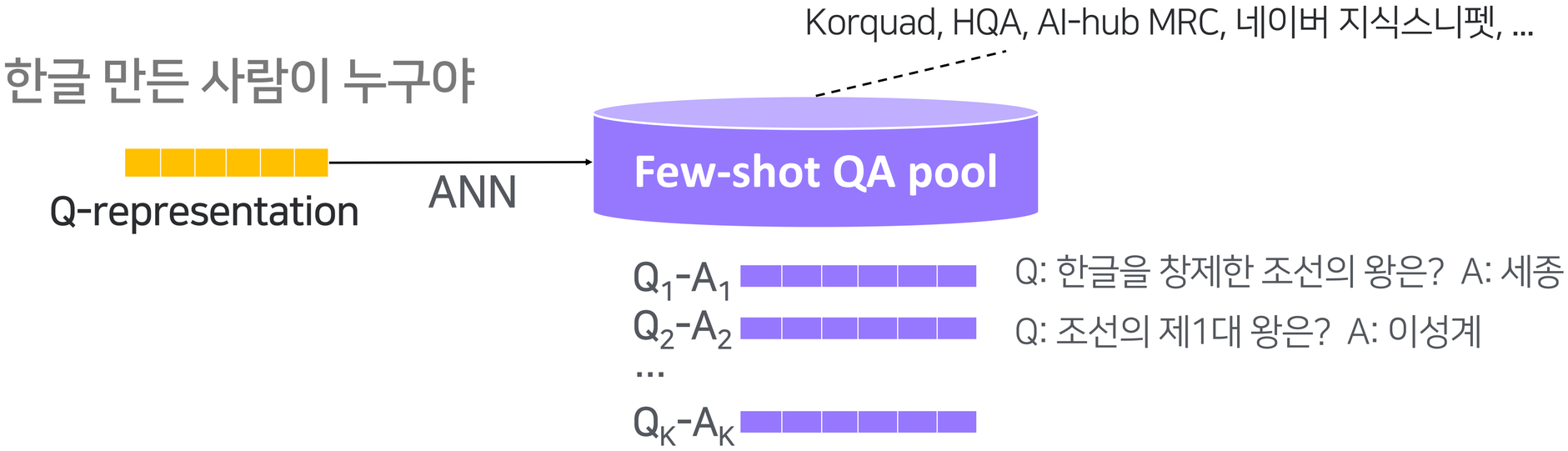
2.3 Few-shot QA example Search

DEVIEW 2021

HyperCLOVA는 관련된 컨텍스트의 Few-shot 줄수록 좋은 생성 만듦*

질문과 의미적으로 유사한 QA 쌍들을 few-shot 으로 사용

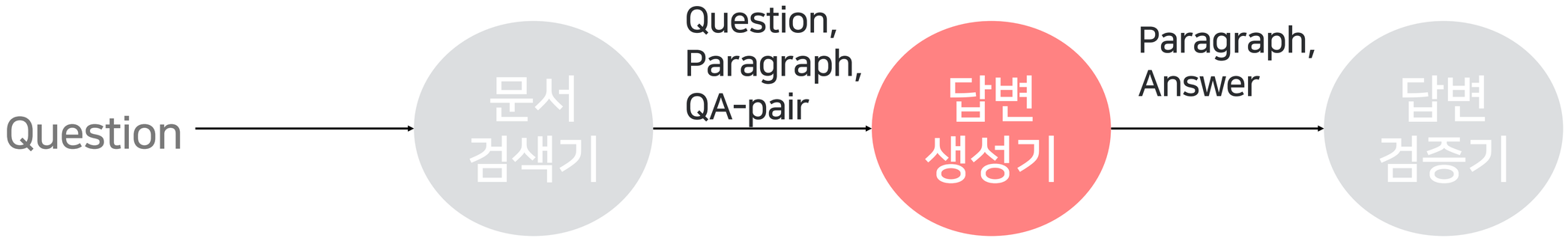
미리 학습된 Bi-encoder 모델로 학습셋의 Q 에 대해서 미리 벡터를 추출해 색인



*What Makes Good In-Context Examples for GPT-3? [Jiachang Liu et al. '21]

3.HyperCLOVA 기반 답변 생성

3.1 HyperCLOVA 기반 답변 생성 예시



HyperCLOVA

근거 문서1: 백성을 사랑하는 마음으로 만든 한글 한글은 조선의 4대 임금인 세종 대왕이 만들었어요. …
근거 문서2: 훈민정음 창제 만든 목적이 분명하고 만든 사람과 만든 시기가 분명한 글자는 한글이 세계적으로 유일하다. 한글 창제와 반포에 대해 당시에는 많은 반대가 있었지만 세종은 …

질문: 조선의 제1대 왕은?
답변: 이성계
질문: 한글을 창제한 조선의 왕은?
답변: 세종
…
질문: 한글 만든 사람이 누구야

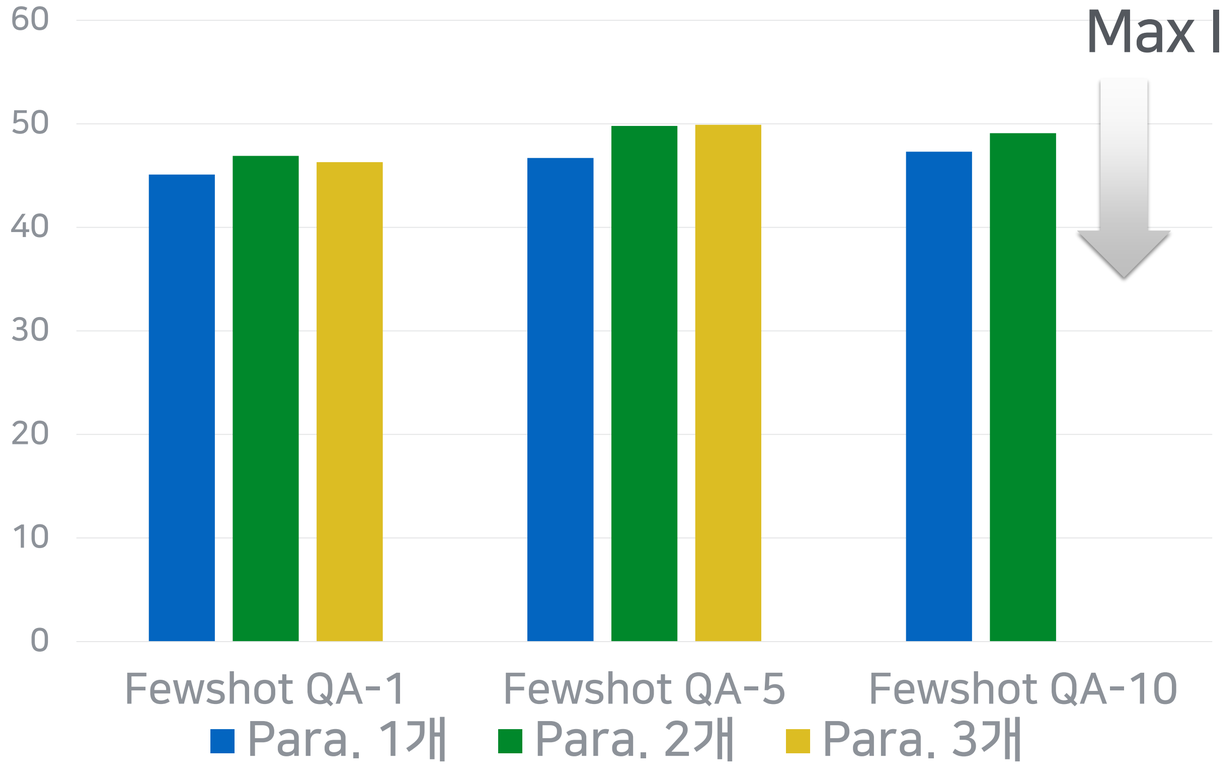
답변: 세종 대왕

3.2 HyperCLOVA만으로 얼마나 답변할까?

일단 정답 데이터가 구축되어 있는 HistoryQA로 HyperCLOVA 성능을 최대로 끌어올려보자

Fewshot QA와 문단는 얼마나 써야 적당한가?

문단3+QA10개 이상되면
Max Input length 초과



3.2 HyperCLOVA만으로 얼마나 답변할까?

하이퍼 파라미터는?

Temperature는 높을수록 randomness가 높아 정확도 떨어짐

정확하고 안정적인 답을 위해서는 Greedy Decoding으로

	13B (Temp. 0, Top-p 0.8)	13B Greedy	39B (Temp. 0, Top-p 0.8)	39B Greedy
#fewshot	F1	F1	F1	F1
Para.2+QA5	60.1	70.6	73.1	75.8

3.3 HyperCLOVA에 학습을 더하면?

DEVIEW 2021

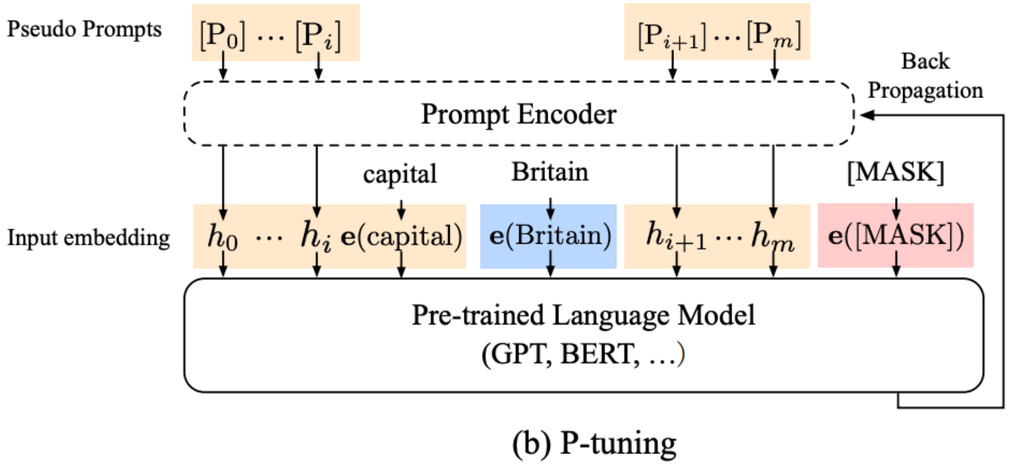
HyperCLOVA 39B Fine-tune

	39B	39B Fine-tune	39B Greedy	39B Fine-tune Greedy
#fewshot	EM'	EM'	EM'	EM'
Para.2+QA5	66.3	72.4	69.9	73.2

(정교한 성능 측정을 위해 띄어쓰기와 포함관계를 고려한 EM'으로 평가 메트릭 변경)

3.3 HyperCLOVA에 학습을 더하면?

P-Tuning : 작은 사이즈의 Prompt Encoder를 Embedding layer에 추가해 pretrain model의 weight는 제외하고 prompt 토큰만 continuous space에서 학습시킨 Light-weight Fine-tune 방식



(b) P-tuning

	39B	39B Fine-tune	39B Greedy	39B Fine-tune Greedy	39B P-tuning
#fewshot	EM′	EM′	EM′	EM′	EM′
Para.2+QA5	66.3	72.4	69.9	73.2	67.3

3.4.1 지식백과 QA 중간평가

클로바 스피커의 백과형 QA질의에 적용해보자!
평가셋 300건에 Human Evaluation 해보니...

평가태깅	설명	39B-HQA모델	
1	지식백과QA에서 응답 가능한 정답	153	51%
0.5	정답은 맞지만 '-이다', '반말', 주어 빠진표현	3	1%
0	오답 / 질의의도에 맞지 않은 답변 / 너무 난해한 답변	141	47%
-1	문법적으로 어색한 답변	3	1%
		300	100%

3.4.2 지식백과 QA 실패 케이스 분석

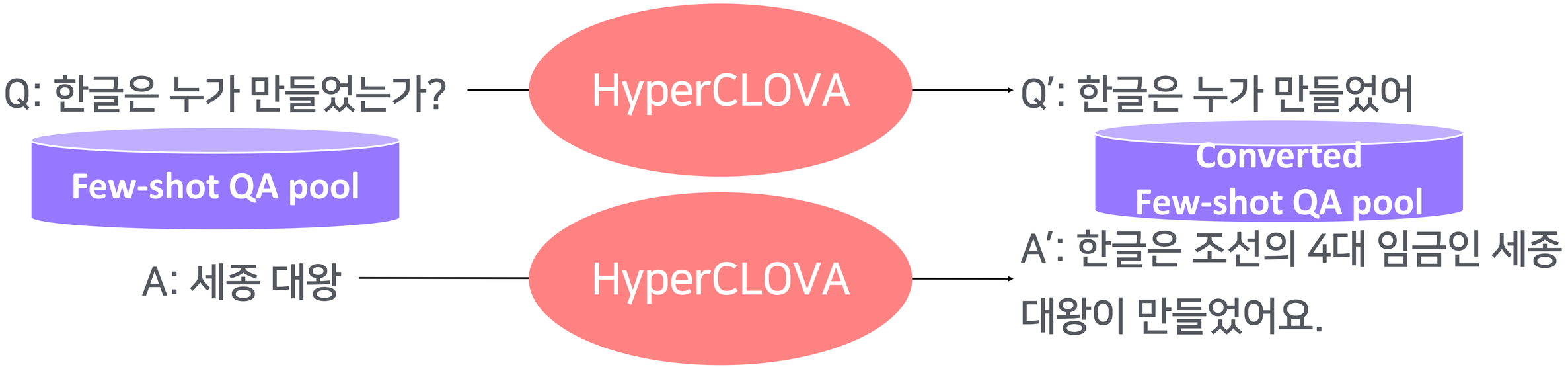
DEVIEW 2021

질문 어투에 따라서 민감!

Few-shot QA Pool 데이터의 어투 변환이 필요

반말 질문으로 변환

질문 내용이 포함된 자세한 답변으로 변환



3.4.2 지식백과 QA 실패 케이스 분석

기존의 HQA나 KorQuad 등의 데이터들은 작위적인 QA 형태가 다수..
단답형 질문 위주

- HQA Examples

Q: 몰도바의 역사 중 터키계 가가우즈인이 남부지역에서 독립을 선언하며 만든 것은 무엇인가?
A: 우지아 공화국

Q: 온조왕이 기원전 16년에 직접 정예병을 이끌고 출정하여 격파한 이민족은 누구인가?
A: 프랑스함대

백과형 QA 도메인에 맞는 데이터가 필요

3.5.1 지식백과 QA 학습 데이터 구축

지식백과 문서들을 기반해서 사용자가 물어볼만한 질의들로 다양한 주제의 학습 데이터 구축

과학, 국어, 동물, 사회, 상식/퀴즈, 세계사, 수학, 식물, 한국사

Query Count 높은 문서에 Click Query를 참고하여 QA를 구축

3.5.2 지식백과 QA 학습

- 클로바 지식백과용 QA 데이터 Fine-tune
- Prompt Engineering
- 하이퍼파라미터 튜닝
- 오프라인QA로 자동 생성한 데이터 추가 Fine-tune
- 문서 검색기 버전에 따른 Fine-tune 모델 앙상블

...

결과 20점 이상은 상승

3.6 어떻게 평가할 것인가

Human Evaluation 평가 기준

1) Task Success : 사실에 근거한 정답 여부

2) Make Sense : 대화의 맥락에 따라 일관적이고, 사실이나 논리에 기반한 대화를 나누는지

3) Specific : 해당 질문에 자세하게 답변하고 있는지

한 답변에 1명이 평가하는 것도 신뢰가 어렵다...결국은 다수결

질문: 목성은 지구의 몇 배야?

지식백과QA 답변: 목성은 태양계에서 가장 큰 행성으로 지구보다 11배 정도 커요.

해석에 따라 다름! 반지름 기준 11.2배, 부피 기준 1300배

3.6 어떻게 평가할 것인가

모델 실험할 때마다 사람이 평가하기에는 너무 리소스가 많이 들어감

모델 비교용 자동 평가 메트릭이 필요

- F1? Exact Match? ROUGE?

3.6 어떻게 평가할 것인가

자동 평가 메트릭들과 Human Evaluation 간의 상관도 추출

서술형 답변과 단답형 답변 각각 다른 메트릭을 사용

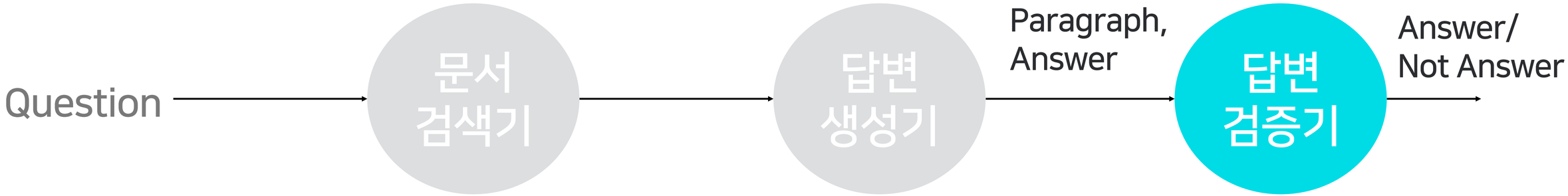
자동 평가 메트릭	Pearson Correlation	
	서술형	단답형
ROUGE-1	0.295	0.367
EM'(핵심단어 포함)	0.250	0.491
NLI(Entailment==1)	0.432	0.319
NLI(Entailment || Neutral ==1)	0.535	0.106
*RDASS-BERTMeanPooling	0.411	0.168
**BERTscore-f1	0.426	0.191

*Reference and Document Aware Semantic Evaluation Methods for Korean Language Summarization [Lee et al. '20]

**BERTScore: Evaluating Text Generation with BERT [Zhang et al. '19]

4. 답변 검증

4.1 답변 검증의 필요성



틀린 답변을 하는 경우가 아직 많아 Fact Check 필요

문서 검색의 품질이 부족

HyperCLOVA의 내부 knowledge의 부족, 잘못된 knowledge도 섞임

학습하더라도 모델이 너무 커서 Catastrophic forgetting 발생

4.2 답변 검증 방법 - 단답형

DEVIEW 2021

Answer span detection

Q: 한글 만든 사람이 누구야

HyperCLOVA

근거 문서1: 백성을 사랑하는 마음으로 만든 한글 한글은 조선의 4대 임금인 세종 대왕이 만들었어요. …

근거 문서2: 훈민정음 창제 만든 목적이 분명하고 만든 사람과 만든 시기가 분명한 글자는 한글이 세계적으로 유일하다. 한글 창제와 반포에 대해 당시에는 많은 반대가 있었지만 세종은 …

질문: 한글을 창제한 조선의 왕은?
답변: 한글을 창제한 조선의 왕은 세종입니다
…

Short answer : 세종 대왕
Answer: 조선의 4대 임금인 세종 대왕이 만들었어요

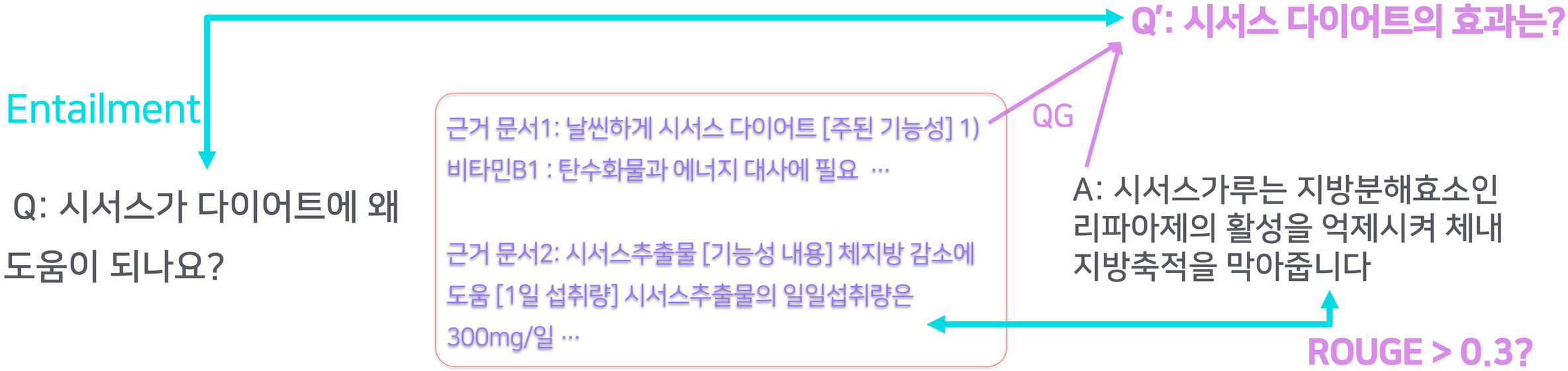
의미적으로 유사한 케이스 검증을 못함

근거 문서에 없는 HyperCLOVA의 내부 지식을 최대로 활용 못함

4.2 답변 검증 방법 - 서술형

ROUGE

Question Generation & Q-Q’ entailment (NLI)



QG 품질에 따라, 근거 문서에 따라, 의존 요소가 너무 많음..

ROUGE 점수가 높음에도 fact check가 잘 되지 않는 경우가 많음

4.2 답변 검증 방법 - Perplexity

DEVIEW 2021

HyperCLOVA의 PPL이 낮은 문장이라면 Fact 신뢰도 높은 문장?

기본 Perplexity

오아시스가 결성된 도시는 영국 런던입니다. PPL : 1121.2, False

공소효력의 범위에 대한 형사소송법의 조문은 대한민국 형사소송법 248조입니다. PPL : 169.1, True

Evidence-Conditioned Perplexity

Evidence: 오아시스는 1991년 영국 맨체스터에서 결성된 록 밴드이다 오아시스가 결성된 도시는 영국 런던입니다.	PPL : 266.4, False
Evidence: 대한민국 형사소송법 제 248조는 공소효력의 범위에 대한 형사소송법의 조문이다. 공소효력의 범위에 대한 형사소송법의 조문은 대한민국 형사소송법 248조입니다.	PPL : 9.6 , True

ROUGE와 비슷하게 Threshold기준 잡기가 힘듬..

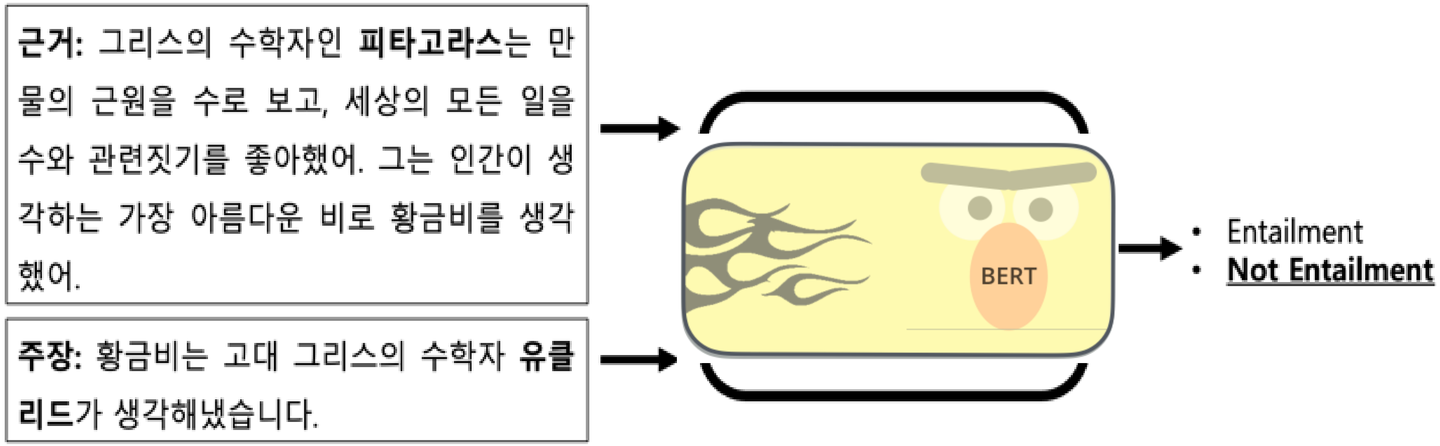
4.2 답변 검증 방법

DEVIEW 2021

단답형, 서술형 모두에서 의미적으로 맞는지 확인할 수 있는 방법?

FEVER* 같은 Fact check용 Claim-Evidence NLI를 만들자

기존 NLI는 한 문장 간의 관계를 계산



*FEVER: a Large-scale Dataset for Fact Extraction and VERification [Thorne et al. '18]

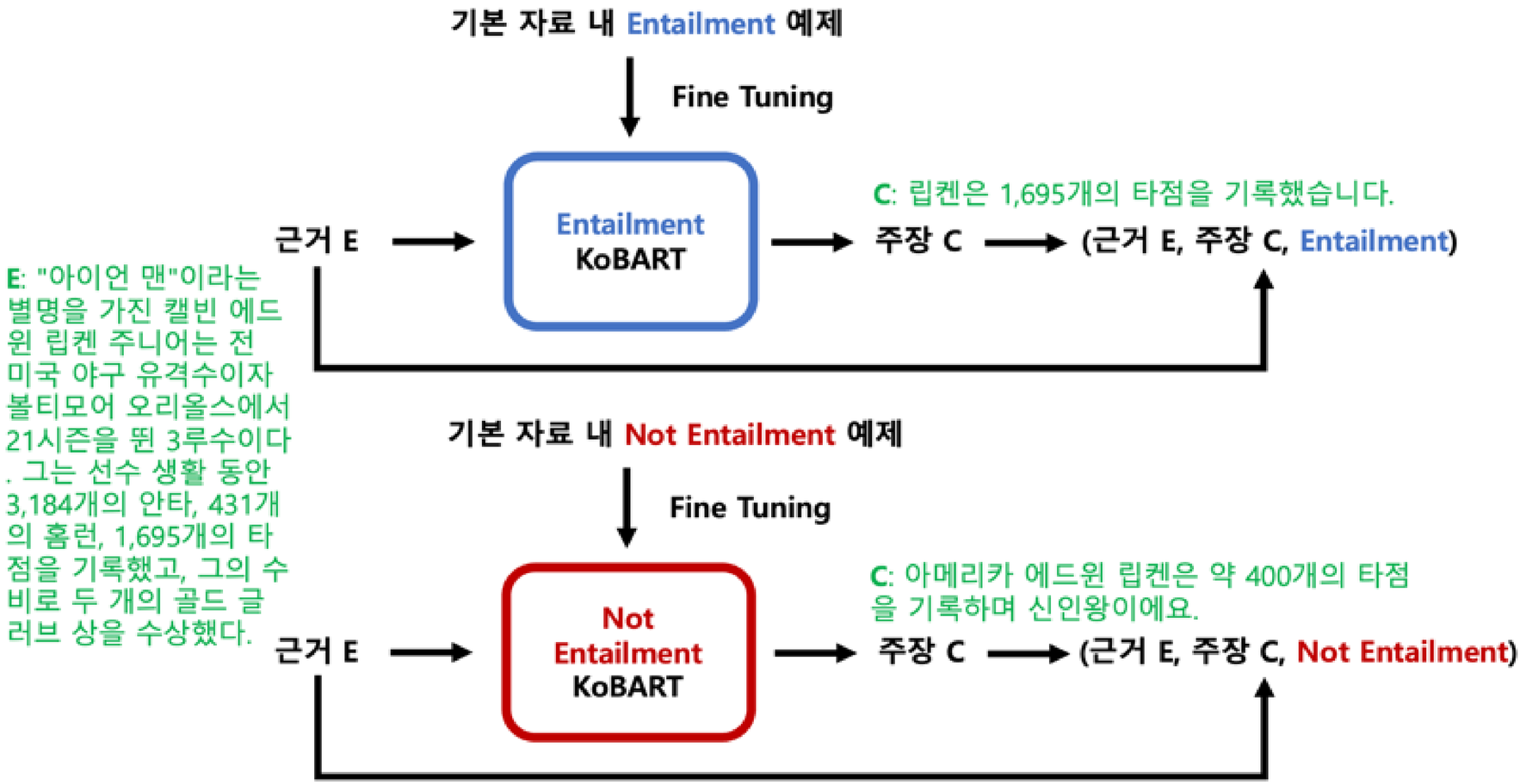
4.3 답변 검증 데이터 구축

Seed data는 지식백과 QA용 학습데이터의 Q, A, Evidence Paragraph

모델 평가 과정에서 태깅된 Entailment (정답), Not Entailment (오답) 데이터가 존재

4.3 답변 검증 Data Augmentaion

1. BART로 Evidence에서 Claim 생성 (E2C)

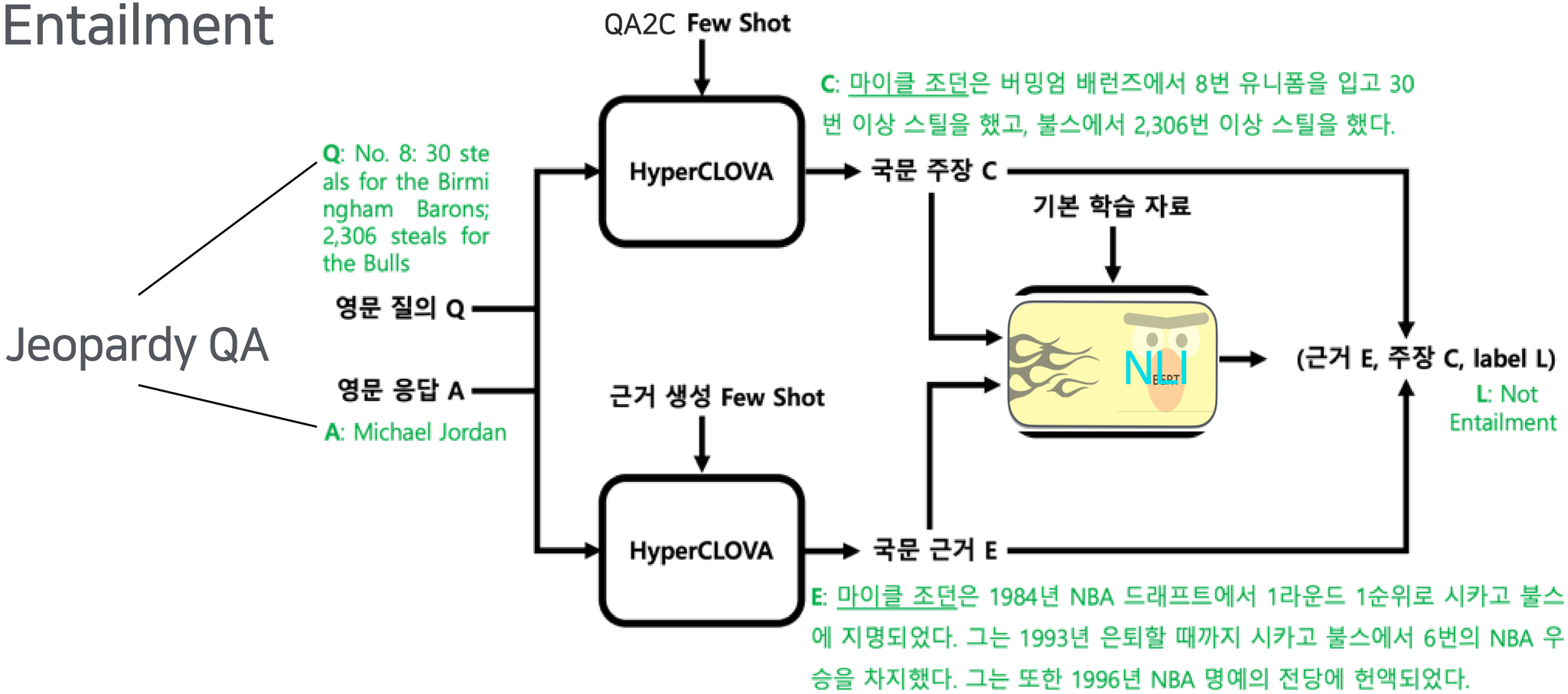


*자연어 생성 모델을 이용한 준지도 학습 기반 한국어 사실 확인 자료 구축 [Jeong et al. '21]

4.3 답변 검증 Data Augmentaion

2. HyperCLOVA로 Claim 및 Evidence 생성 (C2E)

- Entailment

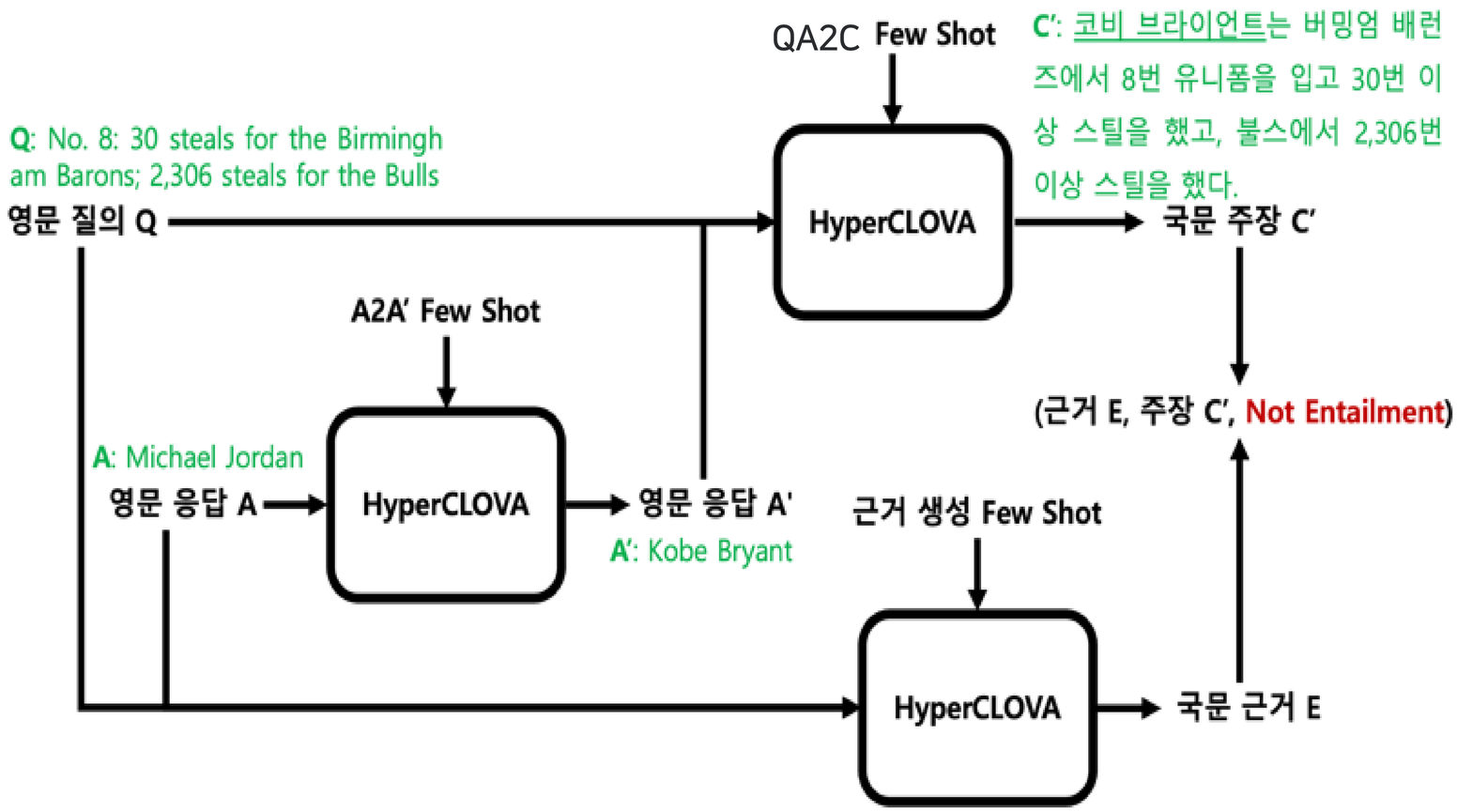


*자연어 생성 모델을 이용한 준지도 학습 기반 한국어 사실 확인 자료 구축 [Jeong et al. '21]

4.3 답변 검증 Data Augmentaion

2. HyperCLOVA로 Claim 및 Evidence 생성 (C2E)

- Not Entailment



4.4 답변 검증 모델 개발

	Seed data	Seed data + E2C	Seed data + C2E	Seed data + FEVER 번역	TOTAL
Accuracy	0.715	0.739	0.739	0.730	0.744
Recall	0.676	0.728	0.739	0.724	0.749
Precision	0.890	0.879	0.868	0.868	0.866
F1	0.769	0.796	0.798	0.790	0.803

*KGAT, **GEAR, ***MLA 등 Evidence 기반의 다양한 Fact verification 모델을 연구중

*Fine-grained Fact Verification with Kernel Graph Attention Network [Liu et al. ‘21]

**GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification [Zhou et al. ‘19]

***A Multi-Level Attention Model for Evidence-Based Fact Checking [Kruengkrai et al. ‘21]

5. 오프라인 QA 데이터 생성

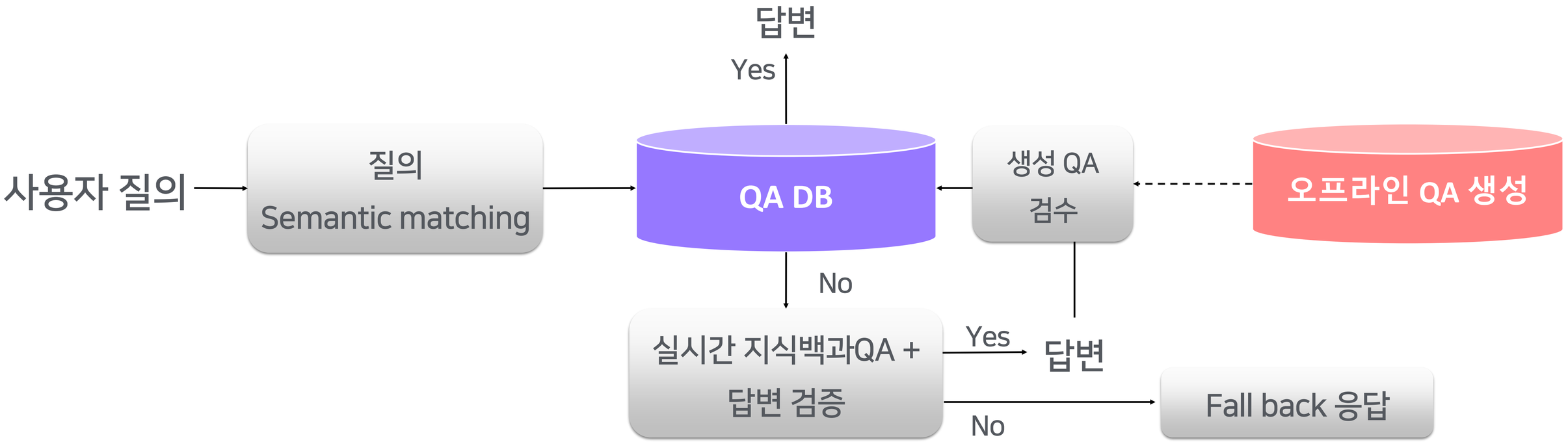
5.1 오프라인 QA 데이터 생성 필요성

DEVIEW 2021

자동으로 QA셋을 만들어 DB를 최대한 구축

DB와 질의간 Semantic Matching + 실시간 QA 통합

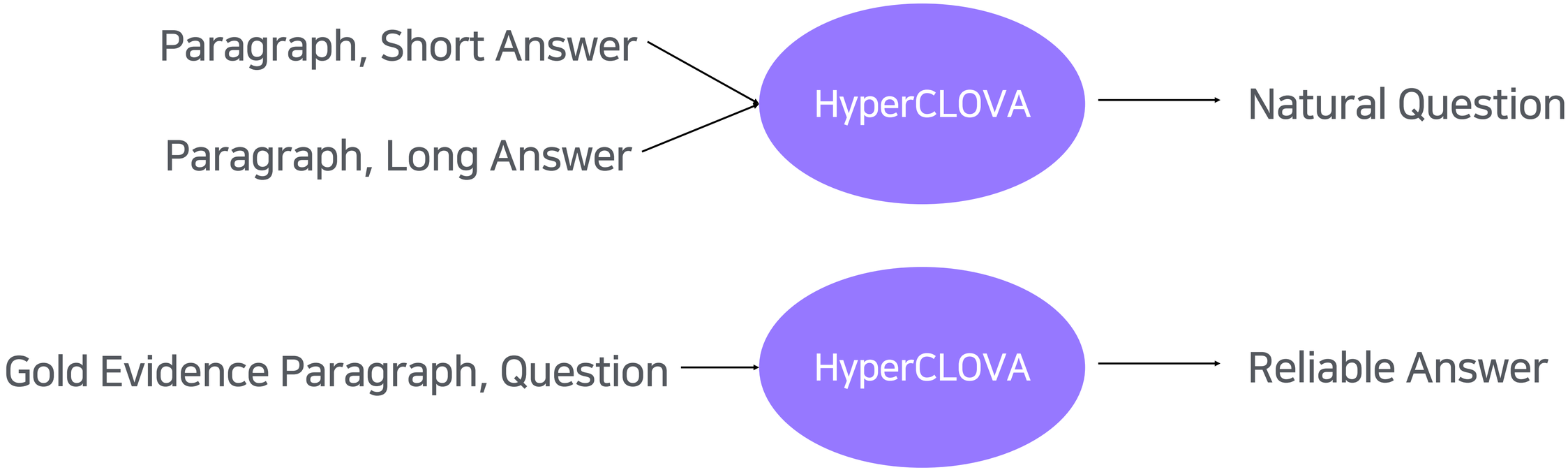
- 사용자의 질의을 미리 예측해서 HyperCLOVA의 latency도 줄여줌



5.2 문서기반 오프라인 QA 생성

Question Generation 모델

Gold Evidence based Answer Generation 모델

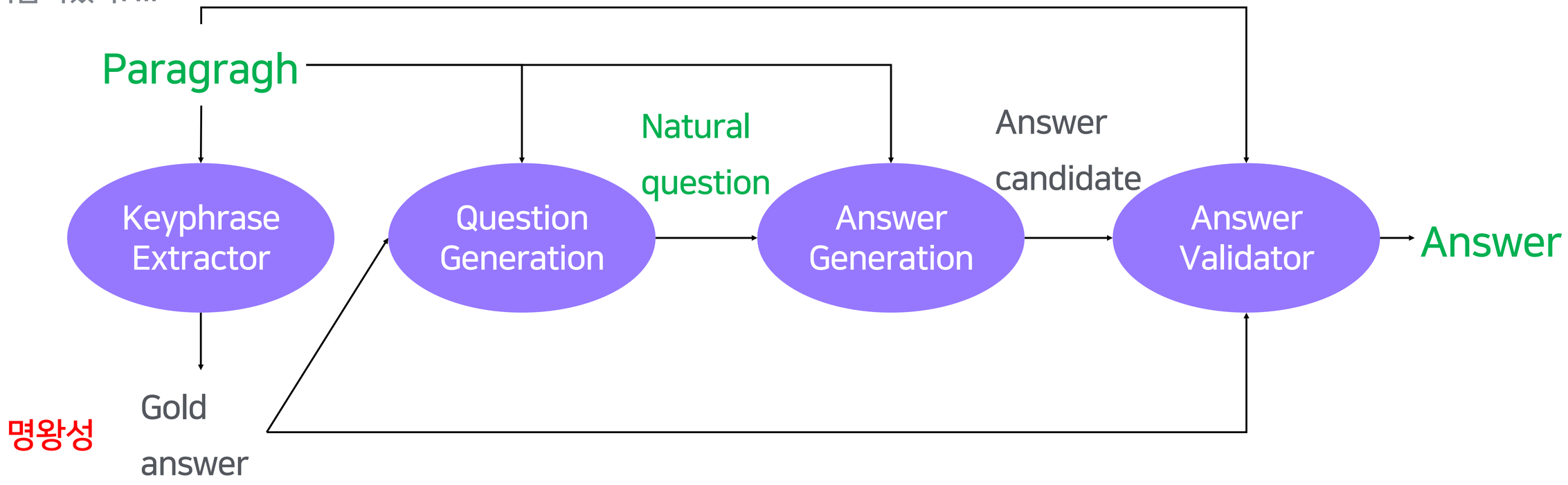


5.2 문서기반 오프라인 QA 생성

시보그 연구팀의
맥밀란이 94번 원소를
명왕성으로부터 유래한
플루토늄으로
이름지었다. ...

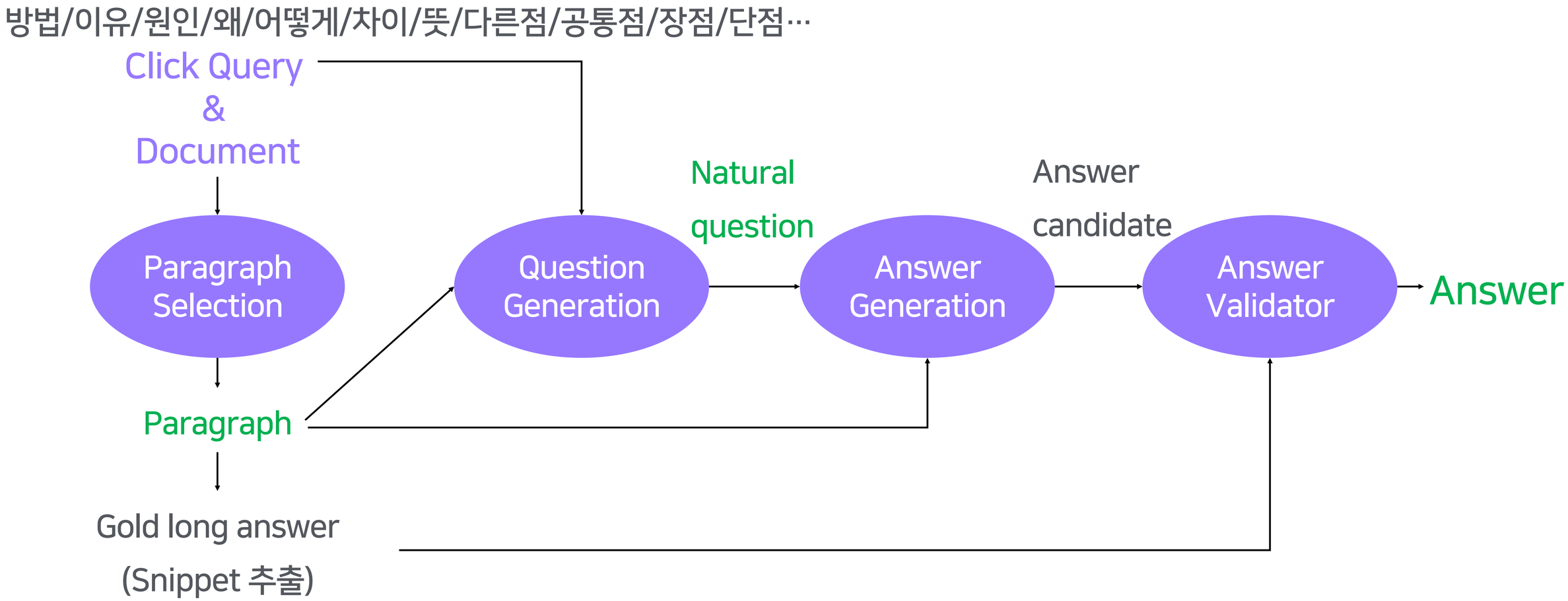
맥밀란의 플루토늄은
어디에서 유래했어?

플루토늄은 명왕성에서
유래된 원소에요

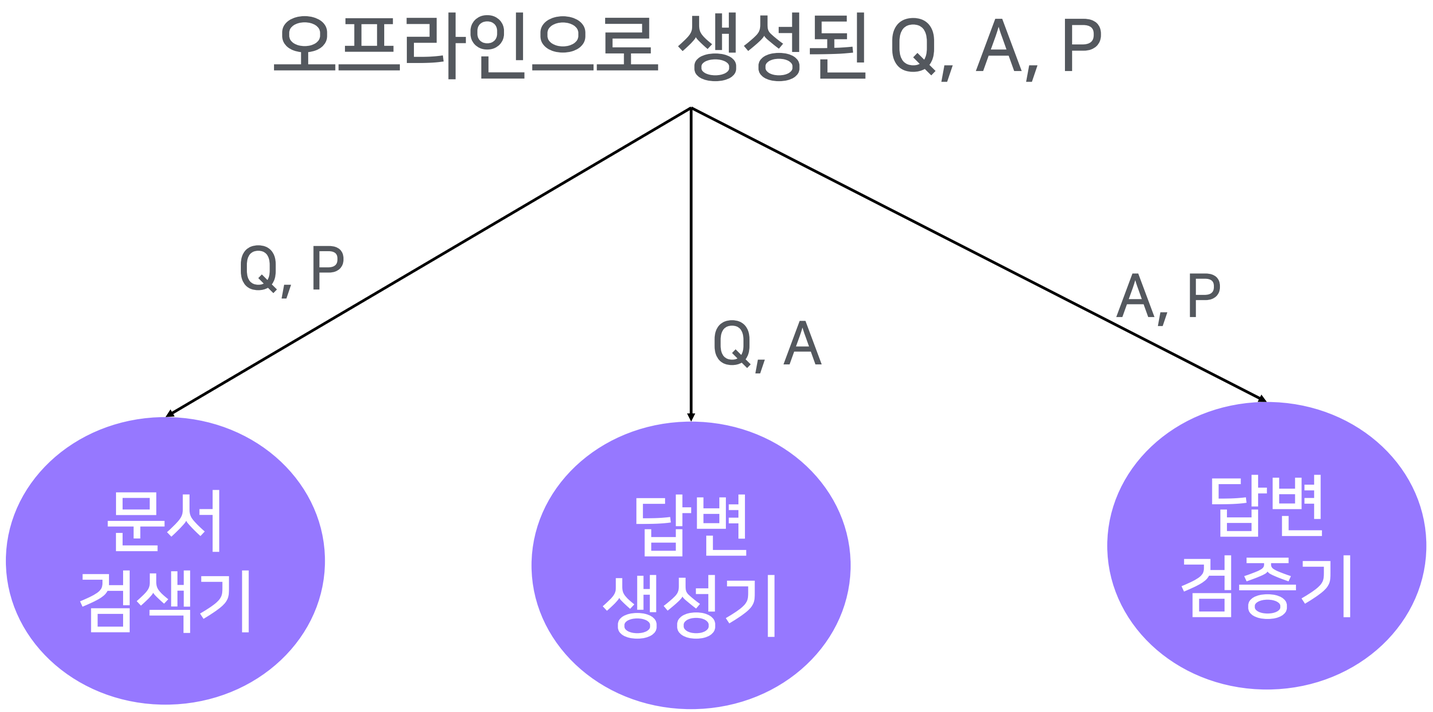


5.3 질의기반 오프라인 QA 생성

서술형 QA



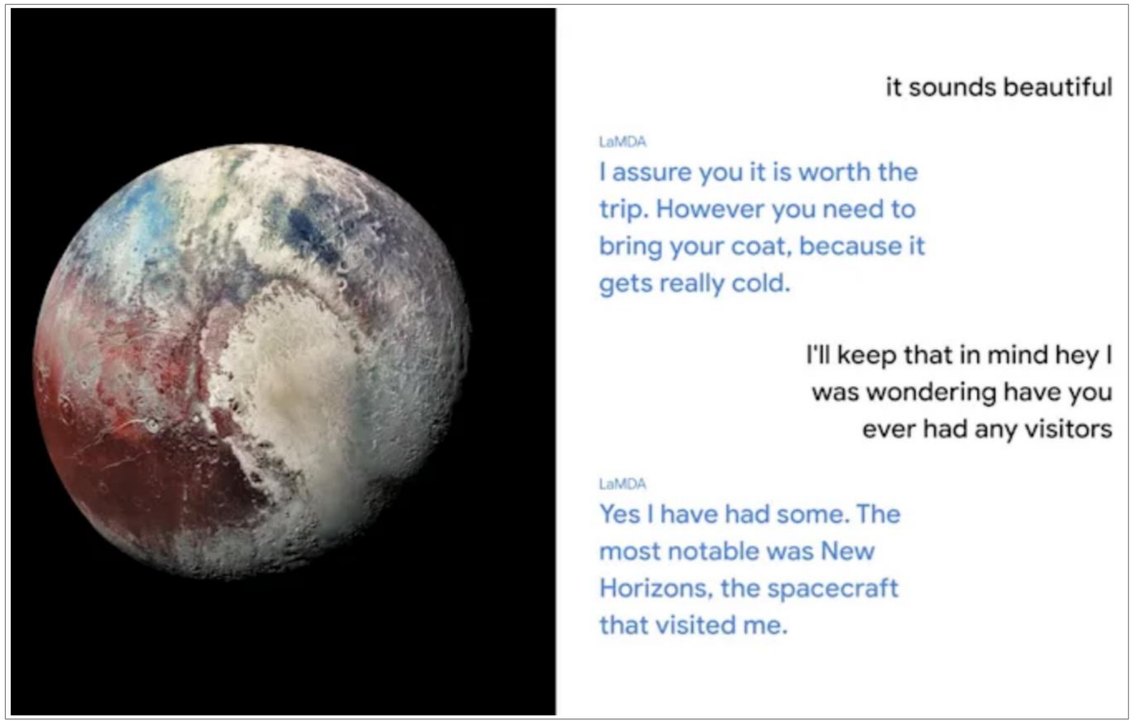
5.4 학습용으로도 재사용



6. Future Woks

6.1 서비스 적용 플랜

구글 LaMDA 같은 도메인별 QA
(구글 LaMDA를 뛰어넘는 QA 만들 사람을 모집중)



https://blog.google/technology/ai/lamda/

Thank You