
LM-based Question Answer Generation

2024.11.08

Data Mining & Quality Analytics Lab.

추창욱

발표자 소개



❖ 추창욱 (Changwook Chu)

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- M.S. Student (2023.09 ~ Present)

❖ Research Interest

- Diffusion Models for Time Series
- Large Language Models

❖ Contact

- chaliechu117@korea.ac.kr

Contents

❖ Introduction

- Importance of Question Answer Generation

❖ Question Answer Generation

- An Empirical Comparison of LM-based Question and Answer Generation Methods (2023, ACL)
- Cooperative Self-training of Machine Reading Comprehension (2022, NAACL)
- LIQUID: A Framework for List Question Answering Dataset Generation (2023, AAAI)

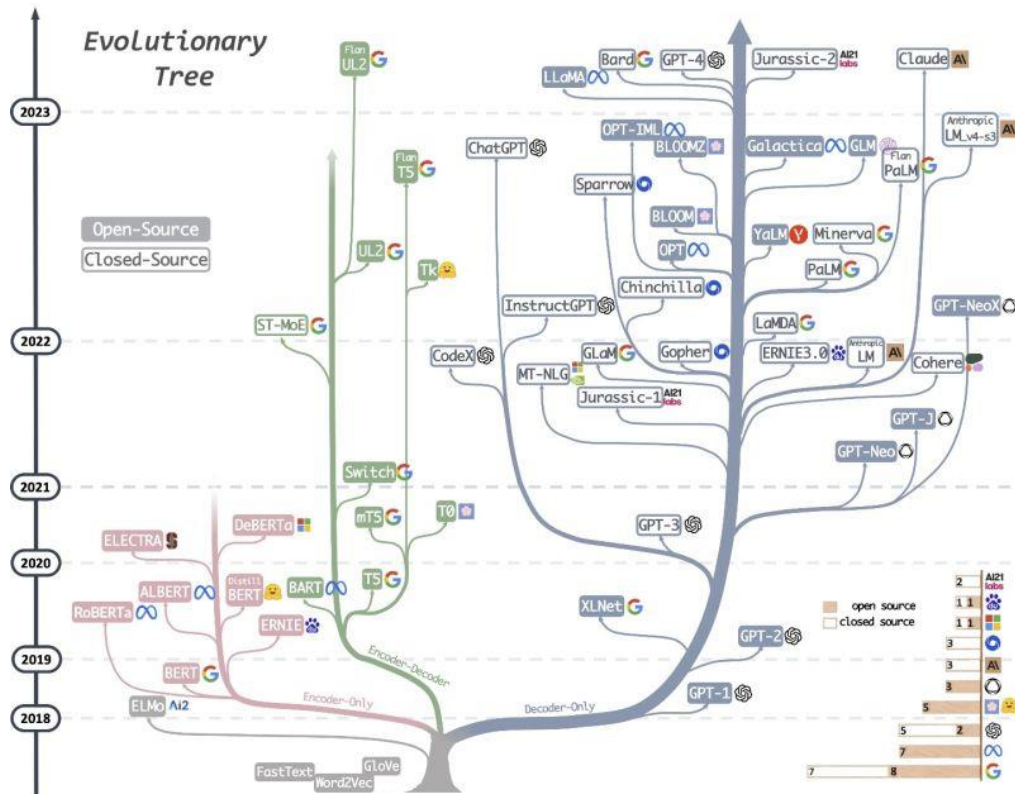
❖ Conclusion

Introduction

Importance of Question Answer Generation

❖ 거대 언어 모델 (Large Language Models)

- ChatGPT 이후에도 LLM은 꾸준히 발전
- 거대 언어 모델은 정보 집약적인 답변이 가능하므로 업무 효율화에 대한 혁신을 일으키고 있음



Introduction

Importance of Question Answer Generation

❖ 질의응답 (Question Answering, QA)

- 인간의 언어 (자연어)로 된 질문을 모델을 통해 기계가 이해하고 대답하는 작업
- 도메인에 따른 분류 – Open domain QA, Closed domain QA
- 방법론에 따른 분류 – Extraction based QA, Generation based QA

Open Domain QA

- 분야에 관계 없이 다양한 주제의 QA 구현
- SQuAD, Wiki, TriviaQA 등 방대한 데이터셋 사용

Extraction based QA

- 질문과 문서가 주어졌을 때 정답을 문서에서 추출
- 전문적인 정보 전달이 필요할 때 사용

Closed Domain QA

- 특정 분야를 대상으로 QA 구현
- 기술, 의료, 금융 등 한정된 데이터셋 사용

Generation based QA

- 질문이 주어졌을 때 정답을 생성
- 완성된 긴 문장 혹은 문단 형태의 답변을 제공

Introduction

Importance of Question Answer Generation

❖ 질의응답 (Question Answering, QA)

- 인간의 언어 (자연어)로 된 질문을 모델을 통해 기계가 이해하고 대답하는 작업
- 도메인에 따른 분류 – Open domain QA, Closed domain QA
- 방법론에 따른 분류 – Extraction based QA, Generation based QA

Open Domain QA

- 분야에 **특정 도메인에 대한 챗봇을 만드는 연구들이 활발히 진행!**에서 추출
- SQuAD, Wiki, TriviaQA 등 방대한 데이터셋 사용

Extraction based QA

- 전문적인 정보 전달이 필요할 때 사용

Closed Domain QA

- 특정 분야를 대상으로 QA 구현
- 기술, 의료, 금융 등 한정된 데이터셋 사용



Generation based QA

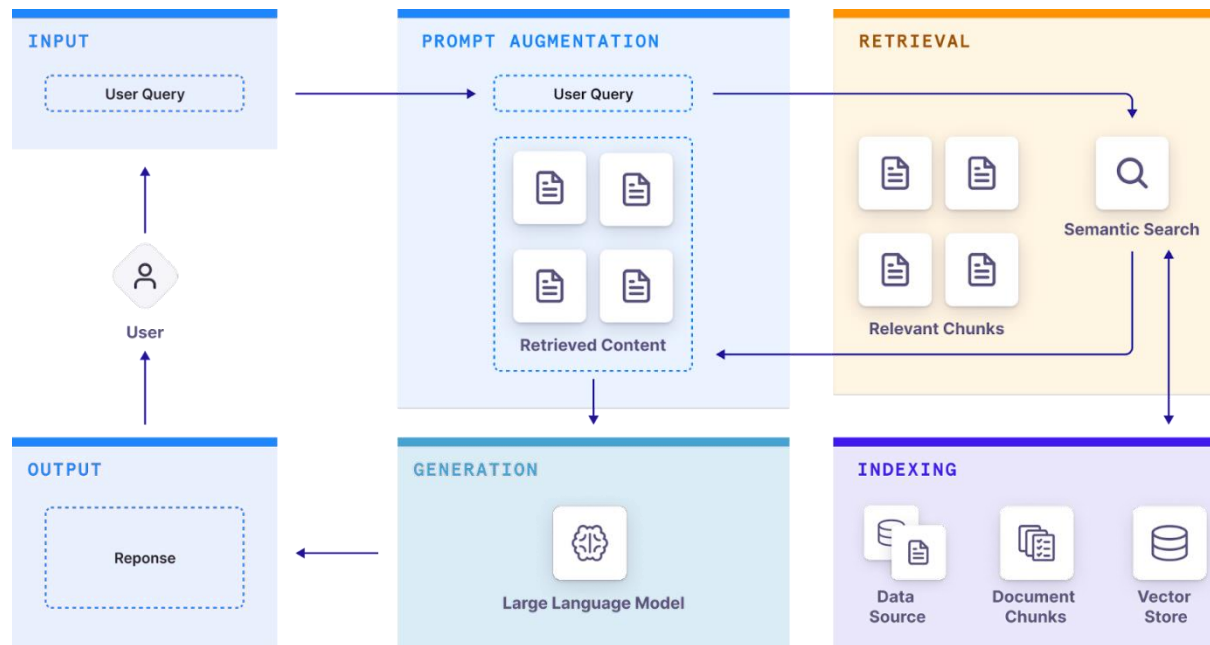
- 질문이 주어졌을 때 정답을 생성
- 완성된 긴 문장 혹은 문단 형태의 답변을 제공

Introduction

Importance of Question Answer Generation

❖ 회수 모델 (Retrieval model)

- 사전 학습된 언어모델은 일반적으로 사용되는 언어 데이터셋으로 학습
- **특정 도메인**에 대한 질의-응답 Task에서 낮은 성능을 보임
- 회수 모델을 활용한 LLM 연구들이 활발히 진행








[Retrieval augmented generation frameworks]

Introduction

Importance of Question Answer Generation

❖ 연구실 관련 세미나

<p>종료</p> <h3>What is LLM and ChatGPT?</h3> <p> 2023. 07. 28 Data Mining & Quality Analytics Lab.</p> <p>What is LLM and ChatGPT?</p> <p>발표자:  채고은</p> <p>📅 2023년 7월 28일 🕒 오후 12시 ~ 📺 온라인 비디오 시청 (YouTube)</p> <p>세미나 정보 보기 →</p>	<p>종료</p> <h3>Training Techniques and Research Trends of LLM</h3> <p> 2023. 08. 04 Data Mining & Quality Analytics Lab. 김현지</p> <p>Training Techniques and Research Trends</p> <p>발표자:  김현지</p> <p>📅 2023년 8월 4일 🕒 오전 12시 ~ 📺 온라인 비디오 시청 (YouTube)</p> <p>세미나 정보 보기 →</p>	<p>종료</p> <h3>Retriever for Language Models</h3> <p>2024.06.07 고려대학교 산업경영공학과 Data Mining & Quality Analytics Lab. 이정민</p> <p>Retriever for Language Models</p> <p>발표자:  이정민</p> <p>📅 2024년 6월 7일 🕒 오전 12시 ~ 📺 온라인 비디오 시청 (YouTube)</p> <p>세미나 정보 보기 →</p>
--	--	---

[거대 언어 모델 관련 세미나]

[회수 모델 관련 세미나]

Introduction

Importance of Question Answer Generation

❖ 특정 도메인 데이터셋

- 특정 도메인에 대한 **관련 문서들은 풍부함**
- 특정 도메인에 대한 **질의-응답 데이터셋** 혹은 **관련 문서-질의-응답 데이터셋이 부족**

TechQA Dataset

기술 지원 도메인을 위한 QA 데이터셋

- Training set: 450 질의-응답 쌍
- Development set: 160 질의-응답 쌍
- Technotes(관련 문서): 801,998건
- 데이터셋은 적고 관련 문서만 풍부함

Question:

Title:
Netcool/Impact 7.1.0: The StateChange value being used by the OMNIBusEventReader is too high

Body:

The value being used is a date and time in the future and as such is preventing the EventReader from capturing the current events.

Answer:

The simplest solution is to manually reset the EventReader StateChange value via the GUI. Stop the EventReader, open it for edit, click the "Clear State" button, exit the editor and restart the EventReader.

Technote

MSQA Dataset

산업 분야에 특화된 QA 데이터셋

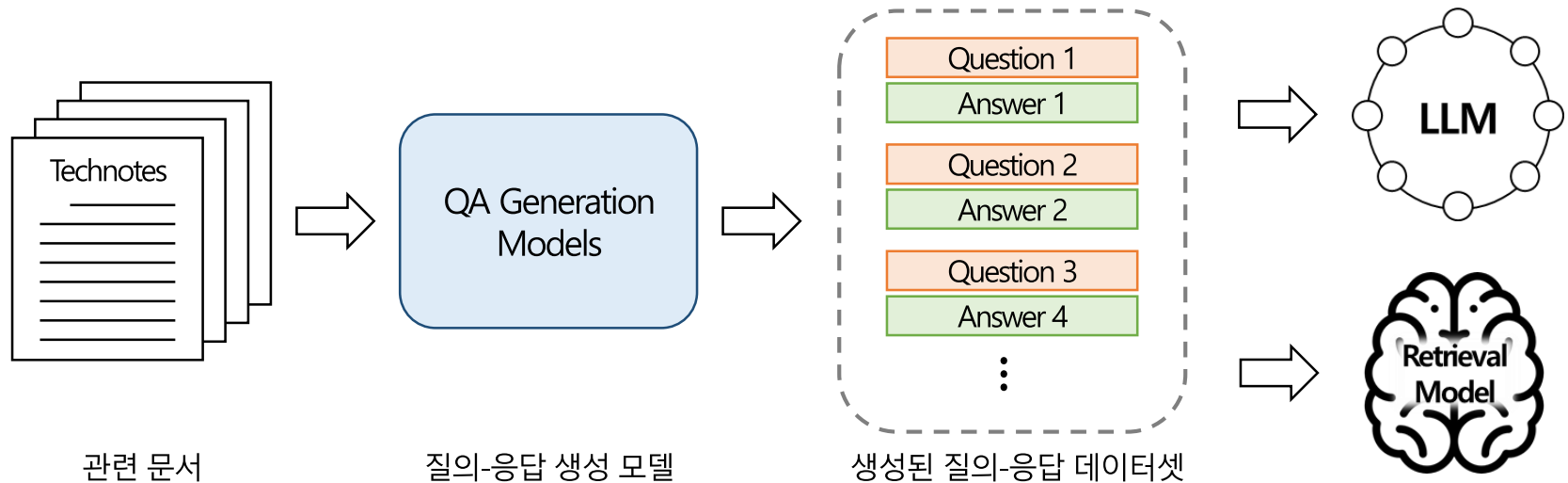
- Data set: 32,000 질의-응답 쌍
- Technotes(관련 문서): 184,655건 이상
- 관련 문서-질의-응답 데이터셋이 없음

Introduction

Importance of Question Answer Generation

❖ 질의-응답 데이터셋 생성 (Question Answer Generation, QAG)

- 언어모델을 활용한 데이터셋 생성을 통해 데이터셋을 증강하여 활용
- 사람이 직접 데이터셋을 생성하지 않아 비용을 줄일 수 있음 (human-labeled data)
- 생성된 데이터셋을 활용하여 특정 도메인에 대한 언어모델 혹은 회수 모델을 학습 가능



Question Answer Generation

An Empirical Comparison of LM-based Question and Answer Generation Methods (2023, ACL)

❖ QAG 방법론 정리

- Question Answer Generation은 Question Generation보다 훨씬 복잡한 Task
- 2023년 ACL에 기재된 QAG model들의 구조에 대해 잘 정리된 논문

An Empirical Comparison of LM-based Question and Answer Generation Methods

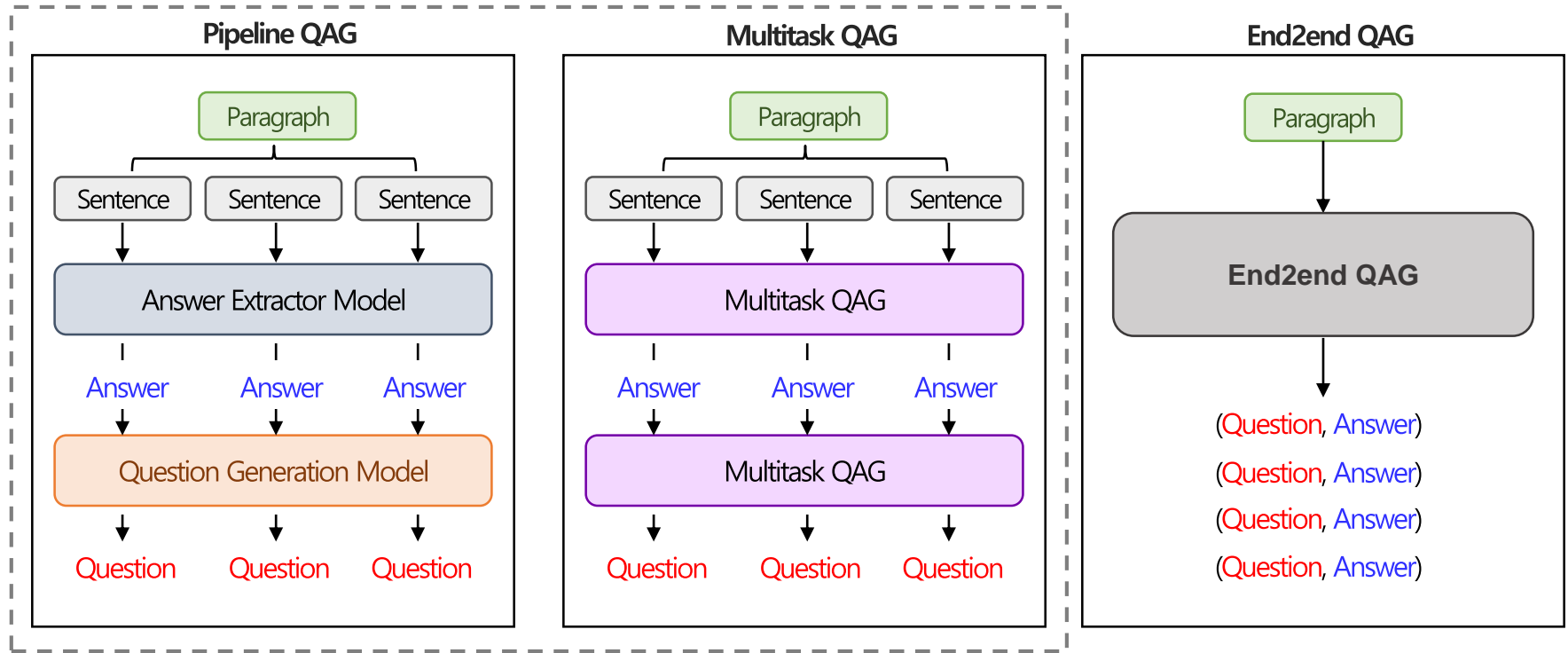
Asahi Ushio and **Fernando Alva-Manchego** and **Jose Camacho-Collados**
Cardiff NLP, School of Computer Science and Informatics, Cardiff University, UK
{UshioA,AlvaManchegoF,CamachoColladosJ}@cardiff.ac.uk

Question Answer Generation

An Empirical Comparison of LM-based Question and Answer Generation Methods (2023, ACL)

❖ QAG 방법론 정리

- 주어진 관련 문서에 대해 질의-응답을 생성하기 위해 언어 모델을 미세조정 하는 관점
- 아래와 같이 3가지 접근법



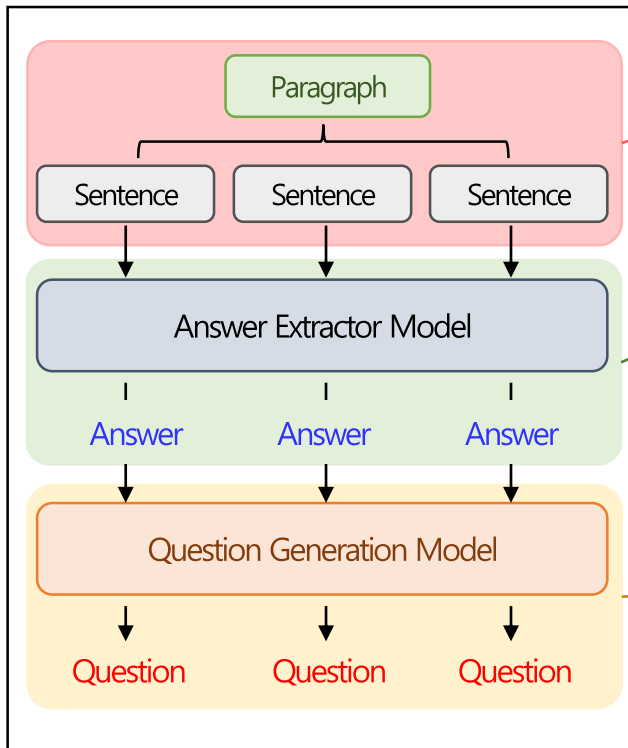
Question Answer Generation

An Empirical Comparison of LM-based Question and Answer Generation Methods (2023, ACL)

❖ Pipeline 질의-응답 생성 방법론

- 관련 문서, 문장, 답, 질문이 있는 어떤 형태의 데이터셋에 대해서 학습 가능
- 두 모델은 독립적으로 학습됨

Pipeline QAG



1. 하나의 문단을 여러 개의 문장 단위로 쪼개는 과정

2. Sub task 1: Answer Extraction

- Answer Extraction Model (P_{ae})는 관련 문서 (c)의 문장 (s)이 주어졌을 때 답변 후보 (\tilde{a})를 생성
- $\tilde{a} = \arg \max_a P_{ae}(a|c, s)$

3. Sub task 2: Question Generation

- Question Generation Model (P_{qg})는 관련 문서 (c)와 답변 후보 (\tilde{a})가 주어졌을 때 질문 (\tilde{q}) 생성
- $\tilde{q} = \arg \max_q P_{qg}(q|c, s, a)$

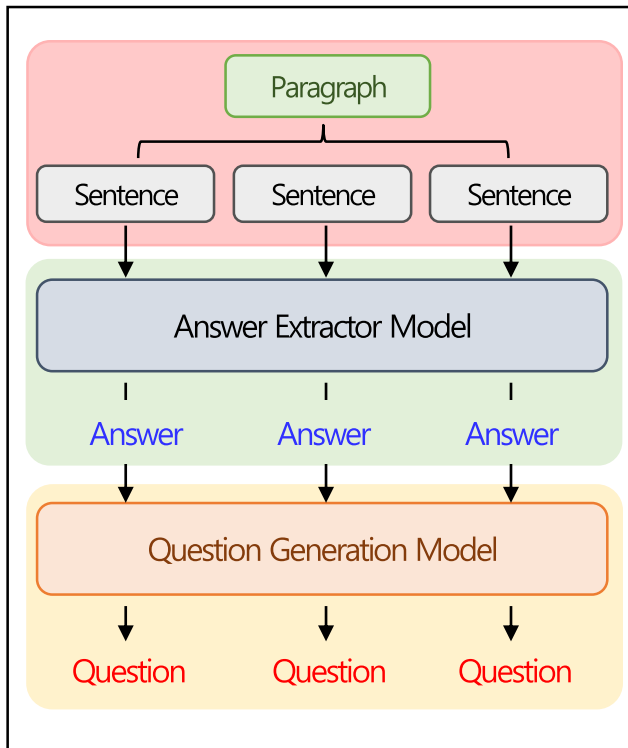
Question Answer Generation

An Empirical Comparison of LM-based Question and Answer Generation Methods (2023, ACL)

❖ Pipeline 질의-응답 생성 방법론

- 관련 문서, 문장, 답, 질문이 있는 어떤 형태의 데이터셋에 대해서 학습 가능
- 두 모델은 독립적으로 학습됨

Pipeline QAG



Maximize the conditional log-likelihood of

Sub task 1: Answer Extraction

$$\tilde{a} = \arg \max_a P_{ae}(a|c, s)$$

Sub task 2: Question Generation

$$\tilde{a} = \arg \max_q P_{qg}(a|c, s, a)$$

Answer Extractor Model Input form

$$= [c_1, \dots, \langle hl \rangle, s_1, \dots, s_{|s|}, \langle hl \rangle, \dots, c_{|c|}]$$

s_i and c_i : i th token of s and c

$|\cdot|$: token number of text

$\langle hl \rangle$: highlighted token

Question Generation Model Input form

$$= [c_1, \dots, \langle hl \rangle, a_1, \dots, a_{|s|}, \langle hl \rangle, \dots, c_{|c|}]$$

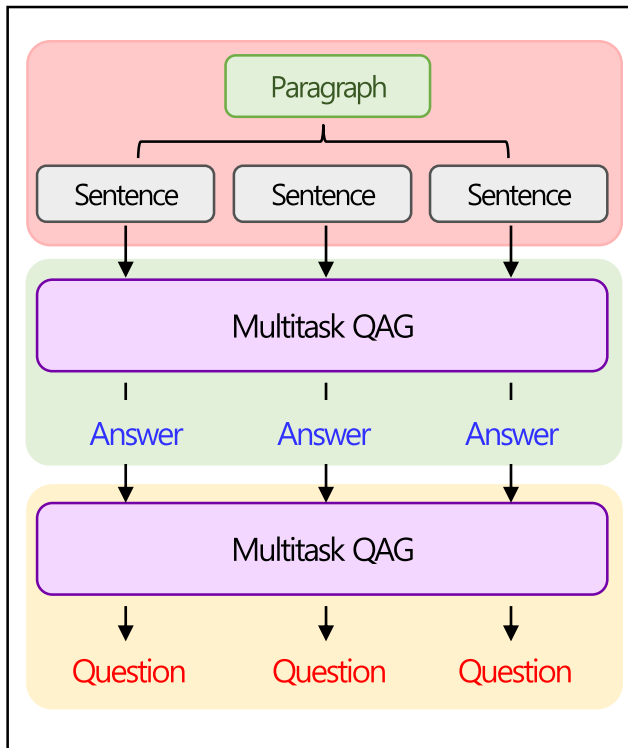
Question Answer Generation

An Empirical Comparison of LM-based Question and Answer Generation Methods (2023, ACL)

❖ Multitask 질의-응답 생성 방법론

- Pipeline 질의-응답 생성 방법론과 같이 2가지 하위 작업으로 모델을 학습
- 두 모델을 독립적으로 학습하는 것이 아닌, 하나의 모델을 다중 작업 학습 방식으로 미세조정

Multitask QAG



1. 훈련 데이터 준비

- AE와 QG 작업에 사용할 데이터 샘플을 모두 혼합

2. 태스크 구분을 위한 prefix 추가

- 각 작업을 구분하기 위해 입력 텍스트 앞에 태스크를 나타내는 **prefix** 추가
- AE 작업: "extract answer", QG 작업: "generate question"

3. 훈련 과정

- 모델이 각 반복마다 무작위로 데이터 샘플링 후 학습
- 각 입력에 있는 **prefix**를 통해 수행해야 할 작업 구분

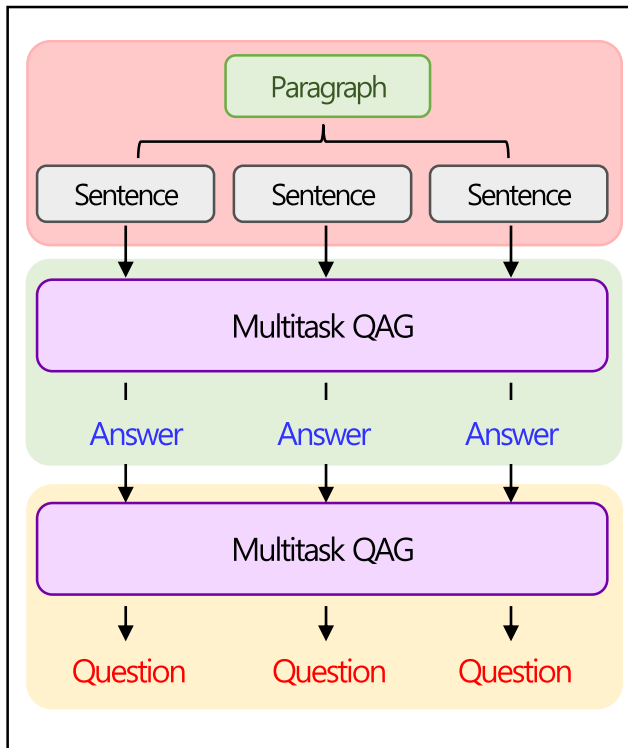
Question Answer Generation

An Empirical Comparison of LM-based Question and Answer Generation Methods (2023, ACL)

❖ Multitask 질의-응답 생성 방법론

- Pipeline 질의-응답 생성 방법론과 같이 2가지 하위 작업으로 모델을 학습
- 두 모델을 독립적으로 학습하는 것이 아닌, 하나의 모델을 다중 작업 학습 방식으로 미세조정

Multitask QAG



1. 데이터

- 문장: "The capital of France is Paris"

2. 태스크 구분을 위한 prefix 추가

- AE 작업: "extract answer: The capital of France is Paris"
- QG 작업: "generate question: Paris"

3. 훈련 과정

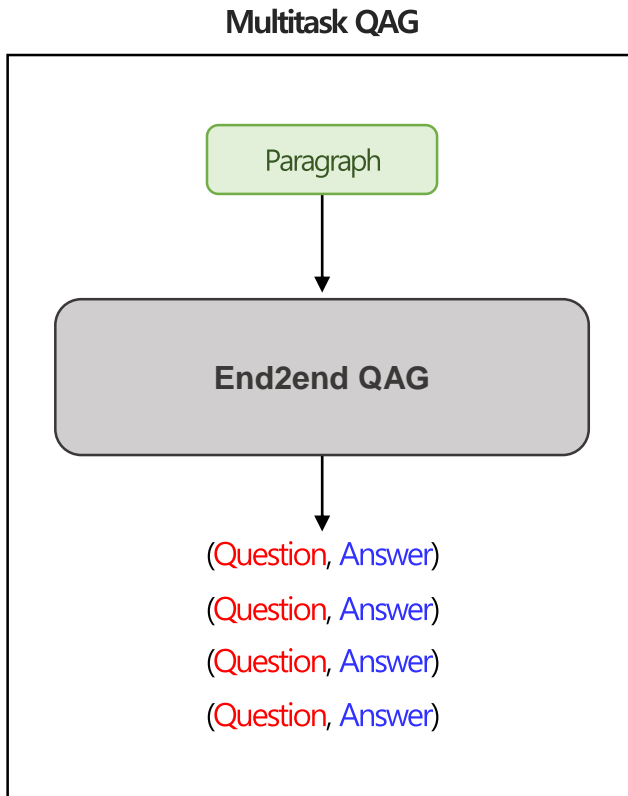
- AE 작업: "Paris"라는 답변을 추출하도록 훈련
- QG 작업: "What is the capital of France?"를 생성하도록 훈련

Question Answer Generation

An Empirical Comparison of LM-based Question and Answer Generation Methods (2023, ACL)

❖ End2end 질의-응답 생성 방법론

- 생성 과정을 두가지 하위 작업으로 나누지 않음
- 질문과 답을 flatten하여 하나의 y 로 취급 → 관련 문서 (c)가 인풋으로 들어가면 y 를 생성



Maximize the conditional log-likelihood of

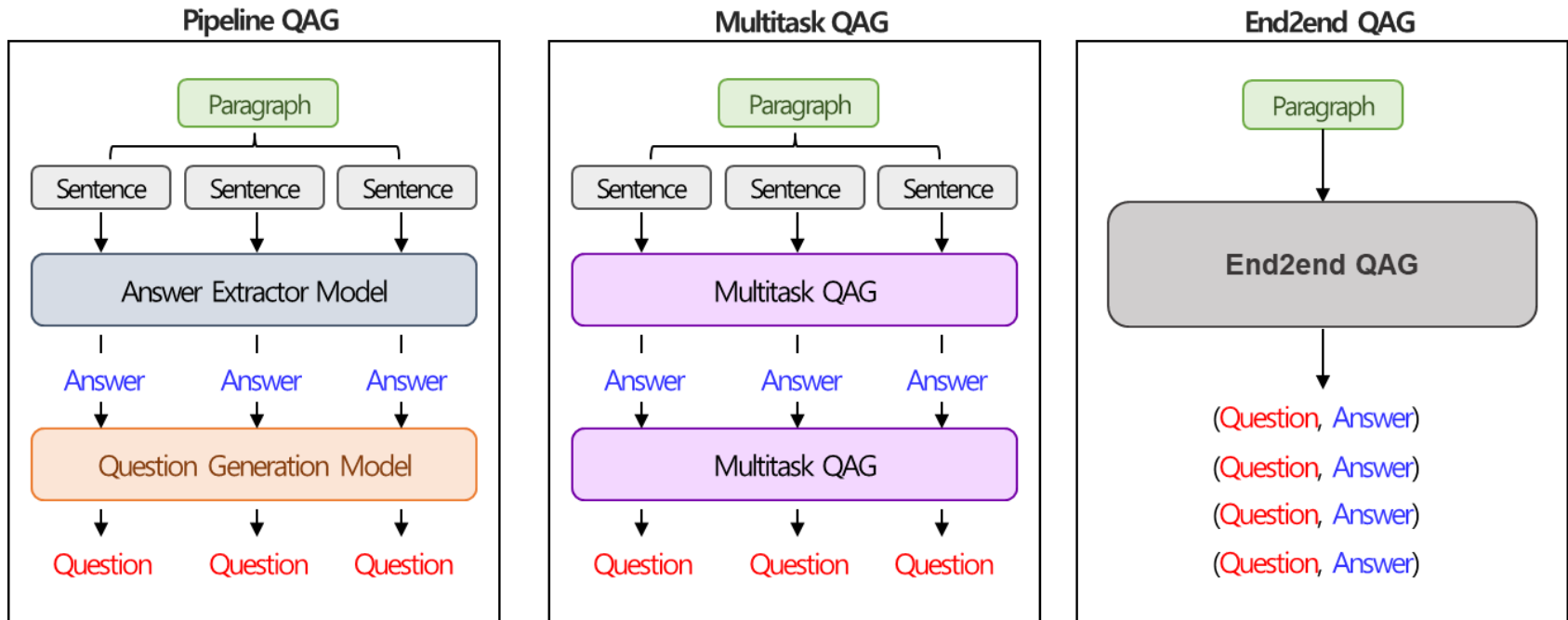
$$\tilde{y} = \arg \max_y P_{qag}(y|c)$$

Question Answer Generation

An Empirical Comparison of LM-based Question and Answer Generation Methods (2023, ACL)

❖ 세가지 질의-응답 생성 방법론 비교

- 세가지 방법론 모두 SQuAD 데이터셋을 이용하여 학습
- 언어 모델은 BART 모델과 T5 모델을 활용

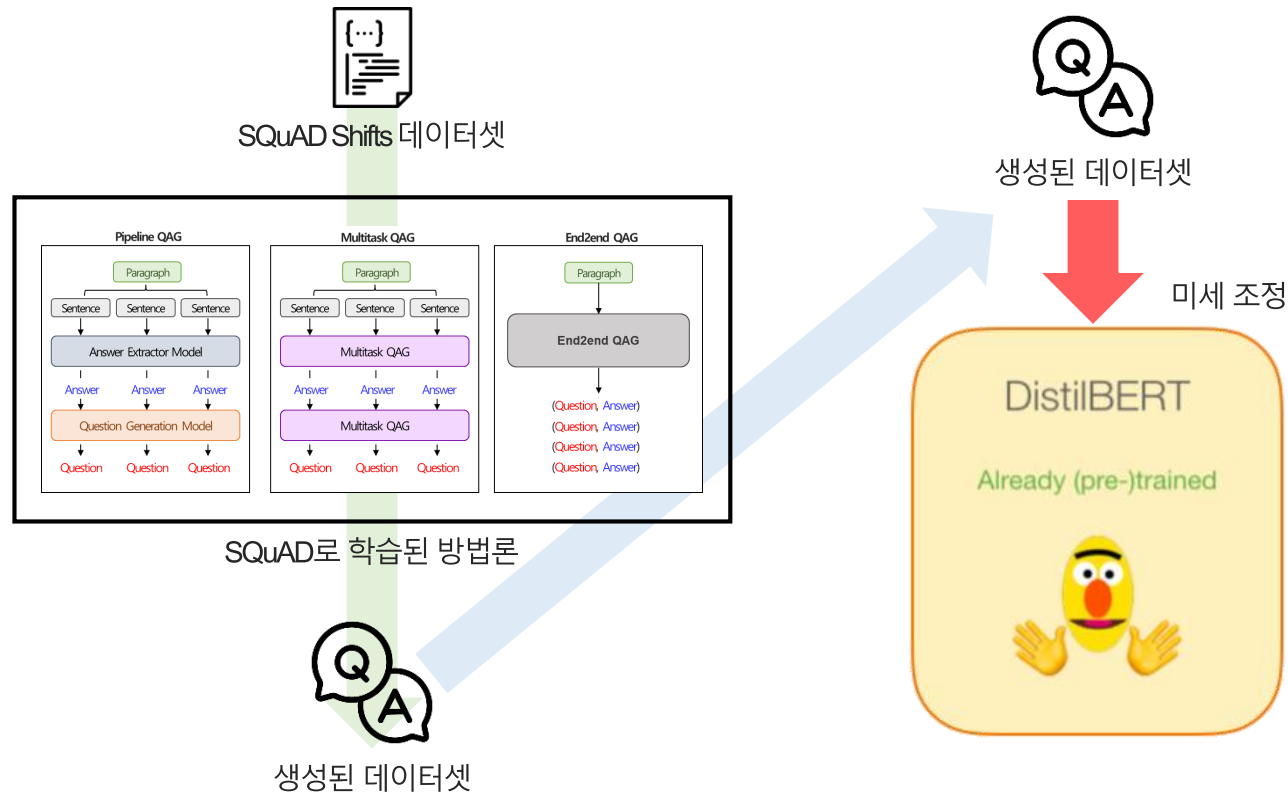


Question Answer Generation

An Empirical Comparison of LM-based Question and Answer Generation Methods (2023, ACL)

❖ 세가지 질의-응답 생성 방법론 비교

- 다양한 도메인의 데이터가 혼합되어 있는 SQuADShifts를 이용하여 질의-응답 데이터 생성
- 생성된 데이터를 활용하여 Question Answering 모델 (DistilBERT) 미세조정 후 성능 평가



Question Answer Generation

An Empirical Comparison of LM-based Question and Answer Generation Methods (2023, ACL)

❖ 세가지 질의-응답 생성 방법론 비교

- $BART_{LARGE}$ 를 사용한 multitask 방식과 $T5_{LARGE}$ end2end 방식이 F1과 exact match 기준 가장 우수
- $T5_{LARGE}$ 모델은 end2end 방식이 일관된 성능을 보이지만, $BART_{LARGE}$ 는 성능이 저조함
- 어떤 특정 구조가 가장 좋다는 결론을 내리기 어려움

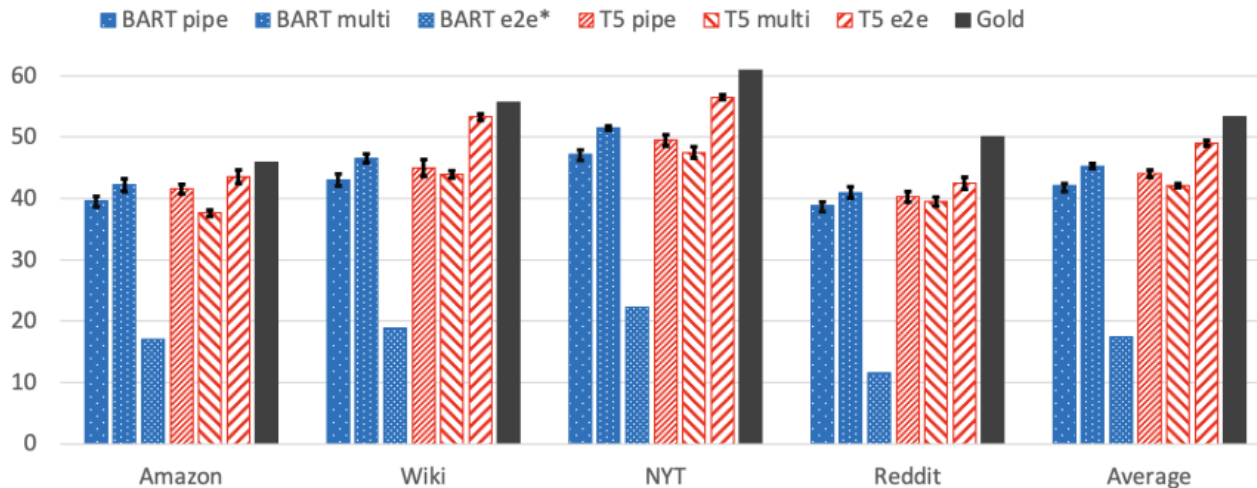


Figure 2: Downsampled (equal-sized) SQuADShifts QA evaluation results (F_1 score with 95% confidence interval) for $T5_{LARGE}$ multitask/pipeline/end2end and $BART_{LARGE}$ pipeline, compared with the original result of each model and the gold QA dataset.

Question Answer Generation

An Empirical Comparison of LM-based Question and Answer Generation Methods (2023, ACL)

❖ 세가지 질의-응답 생성 방법론 비교

- **훈련 비용** (cost): Pipeline \approx Multitask $>$ End2end
- **메모리 요구량** (memory): Pipeline $>$ Multitask \approx End2end
- **생성된 질의-응답 쌍** (Generated QA): Pipeline \approx Multitask $>$ End2end

Approach	Size (training / validation)
Gold QA	3,141 / 1,571
BART _{LARGE} (pipeline)	11,900 / 8,192
BART _{LARGE} (multitask)	11,752 / 8,103
BART _{LARGE} (end2end)	2,012 / 1,399
T5 _{LARGE} (pipeline)	12,239 / 8,417
T5 _{LARGE} (multitask)	12,148 / 8,357
T5 _{LARGE} (end2end)	6,555 / 4,550

Table 3: Average number of question-answer pairs generated for SQuADShifts QA evaluation by each model over all the domains.

[각 도메인별 생성된 질의-응답 쌍 평균 개수]

	Cost	Memory	Generated QA
Pipeline	9.2x	2x	2.7x
Multitask	9.2x	x	2.7x
End2end	x	x	x

Table 4: Comparison among the three proposed QAG approaches in terms of training cost, memory requirements, and generated question-answer pairs, using end2end as a reference. The comparison is performed for T5_{LARGE} with the data used for the main experiments (§ 4.1). Generated QA are averaged across the four SQuADShifts domains.

[비용, 메모리, 생성 개수 비교]

Question Answer Generation

❖ Cooperative Self-training of Machine Reading Comprehension

- 2022년 NAACL에 기재된 Pipeline QAG 방법론을 활용한 QA 모델 자가 학습 방법을 제안한 논문

Cooperative Self-training of Machine Reading Comprehension

Hongyin Luo¹ Shang-Wen Li² Mingye Gao³ Seunghak Yu^{2*} James Glass¹

¹MIT CSAIL, ²Amazon AI ³MIT MTL

{hyluo, mingye, glass}@mit.edu, {shangwel, yuseungh}@amazon.com



Knowledge Base

Tang Dynasty ... Chengdu became nationally known as a supplier of armies and the home of Du Fu, who is sometimes called China's greatest poet.



AER Agent

a supplier of armies and the home of Du Fu

What was Sichuan known for in the ancient world before 957?



QG Agent



QAE Agent

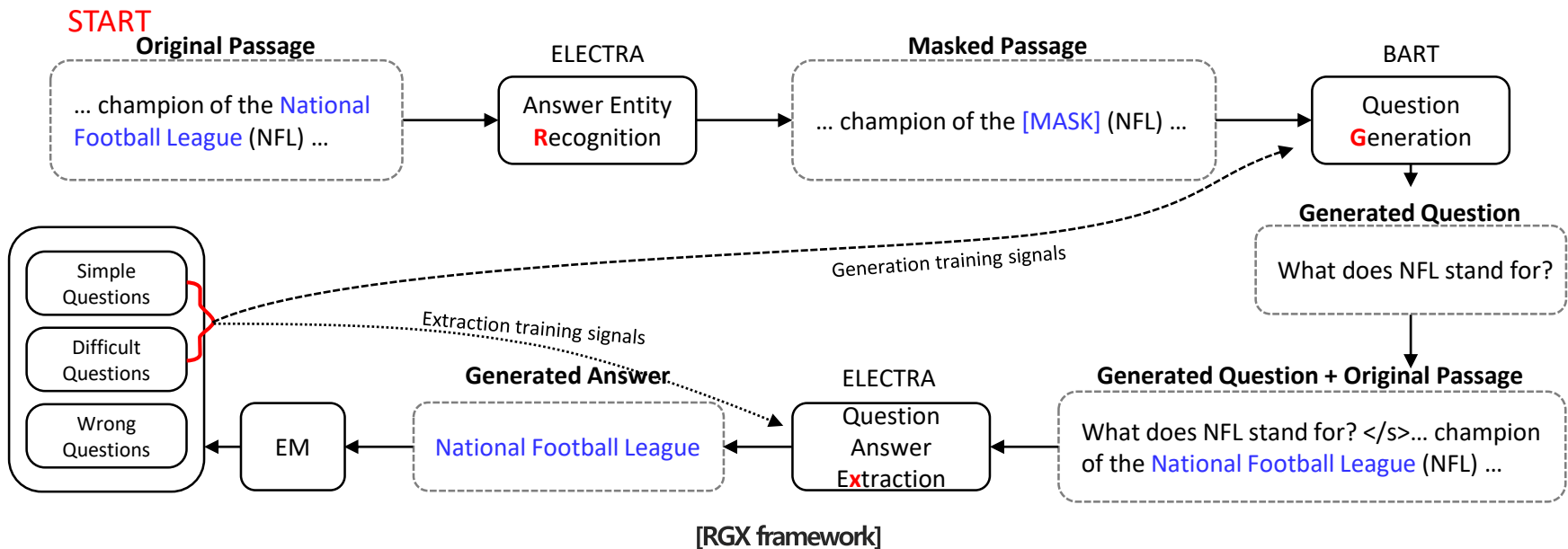
A supplier of armies

Question Answer Generation

Cooperative Self-training of Machine Reading Comprehension (2022, NAACL)

❖ RGX Framework 구조

- Pipeline QAG 형식의 RGX Framework 제안
 - ✓ Answer entity **R**ecognizer: Passage에서 답변이 될 수 있는 Entity를 식별
 - ✓ Question **G**enerator: Answer Entity 부분을 Masking 한 Passage를 기반으로 질문 생성
 - ✓ Question-answer **eX**tractor: 생성된 질문과 Original Passage로 답변 생성

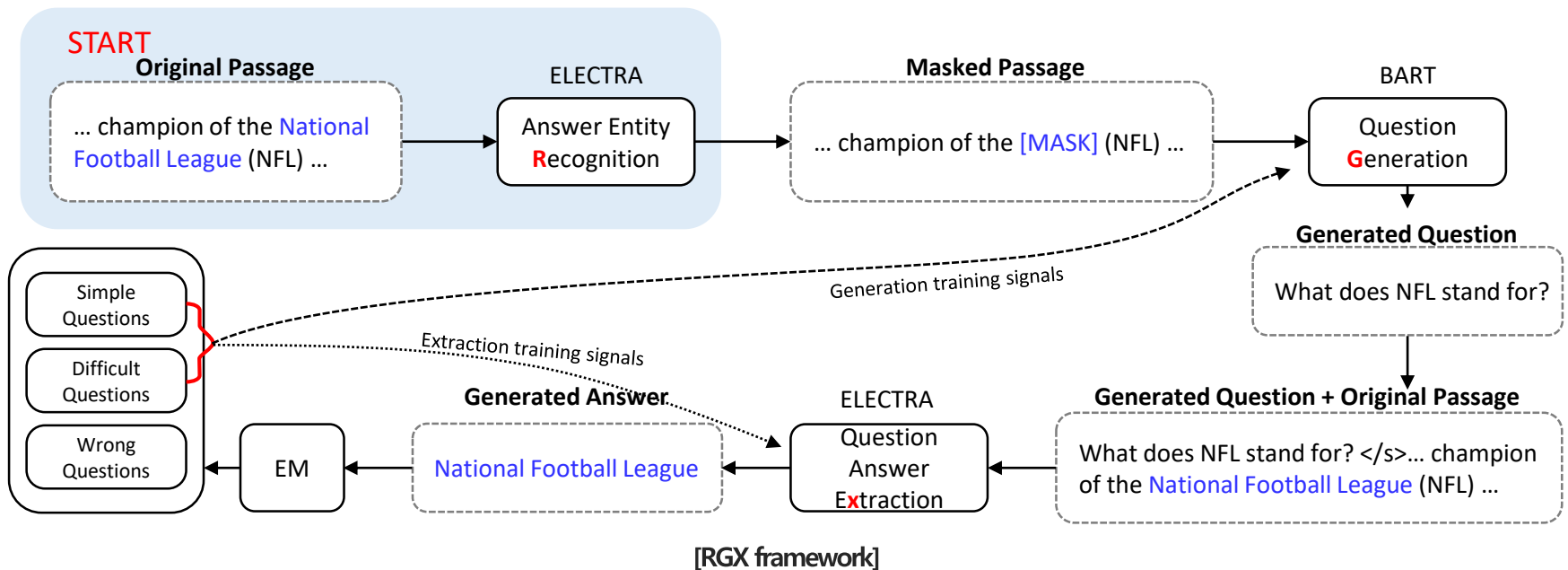


Question Answer Generation

Cooperative Self-training of Machine Reading Comprehension (2022, NAACL)

❖ RGX Framework 구조

- Answer Entity Recognizer: Passage에서 답변이 될 수 있는 Entity를 식별
 - ✓ 다양한 답변 객체 인식 모델에 대한 실험 진행
 - ✓ 어떤 AER 모델을 사용하느냐에 따라 모델의 성능에 큰 영향을 미침
 - ✓ 본 논문에서는 Supervised ELECTRA-Large 사용

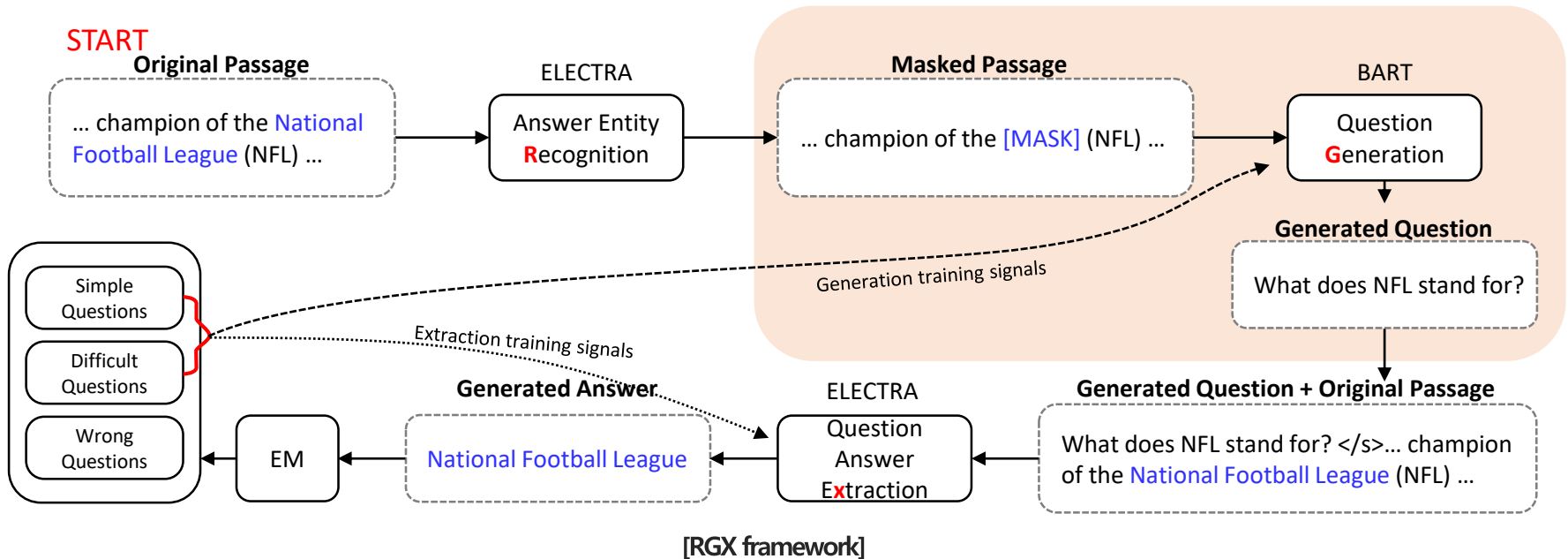


Question Answer Generation

Cooperative Self-training of Machine Reading Comprehension (2022, NAACL)

❖ RGX Framework 구조

- Question Generator: Answer Entity 부분을 Masking 한 Passage를 기반으로 질문 생성
 - ✓ 답변 객체 (e)를 포함한 문장 (p)에서 답변 객체를 [MASK] 토큰으로 대체하여 마스킹 된 문장(p^*)을 만듦
 - ✓ p^* 와 e의 concatenation이 들어갔을 때 질문 (q)을 생성하는 Question Generator 구축, $q = Q([p^*, e])$
 - ✓ 본 논문에서는 BART sequence-to-sequence 모델을 사용

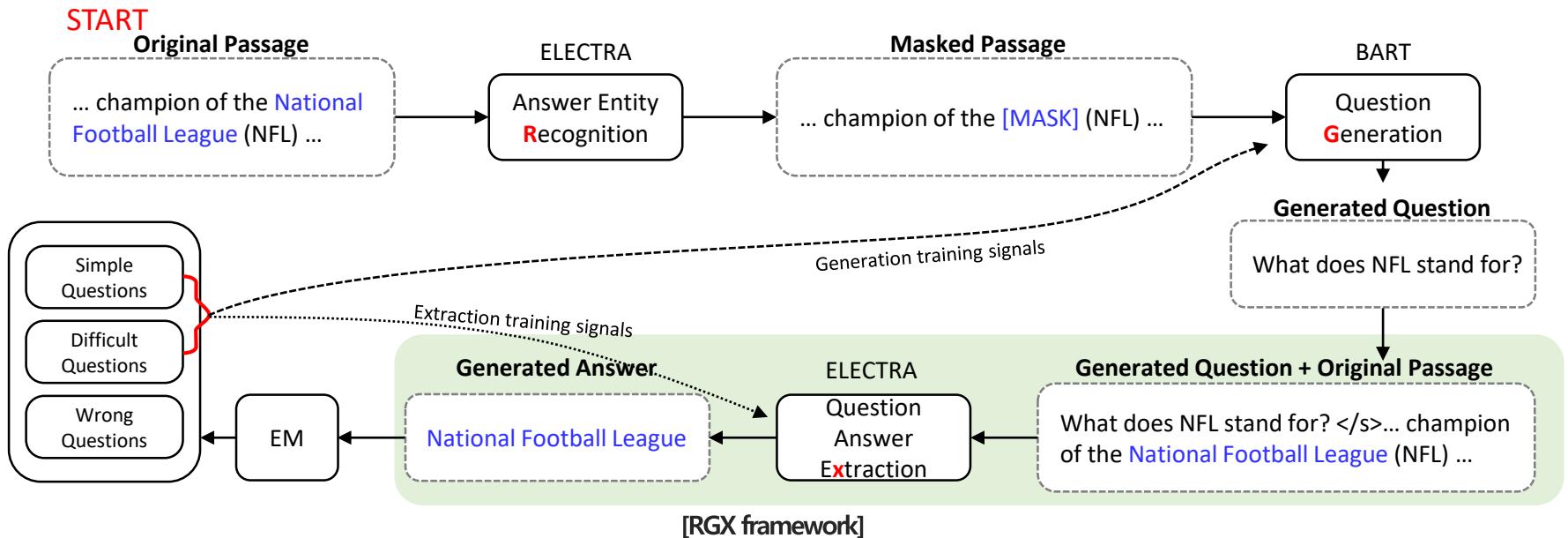


Question Answer Generation

Cooperative Self-training of Machine Reading Comprehension (2022, NAACL)

❖ RGX Framework 구조

- Question-answer extractor: 생성된 질문과 Original Passage로 답변 생성
 - ✓ 생성된 질문 (q)와 Original Passage (p)를 입력으로 받음
 - ✓ 답변의 시작 위치 (l_{st})와 끝 위치 (l_{ed})를 예측
 - ✓ AER 모델은 단순히 엔티티만 고려하기 때문에 답변을 보다 세밀하게 추출하는 Question Answering 모델 사용
 - ✓ 본 논문에서는 ELETRA 모델을 사용

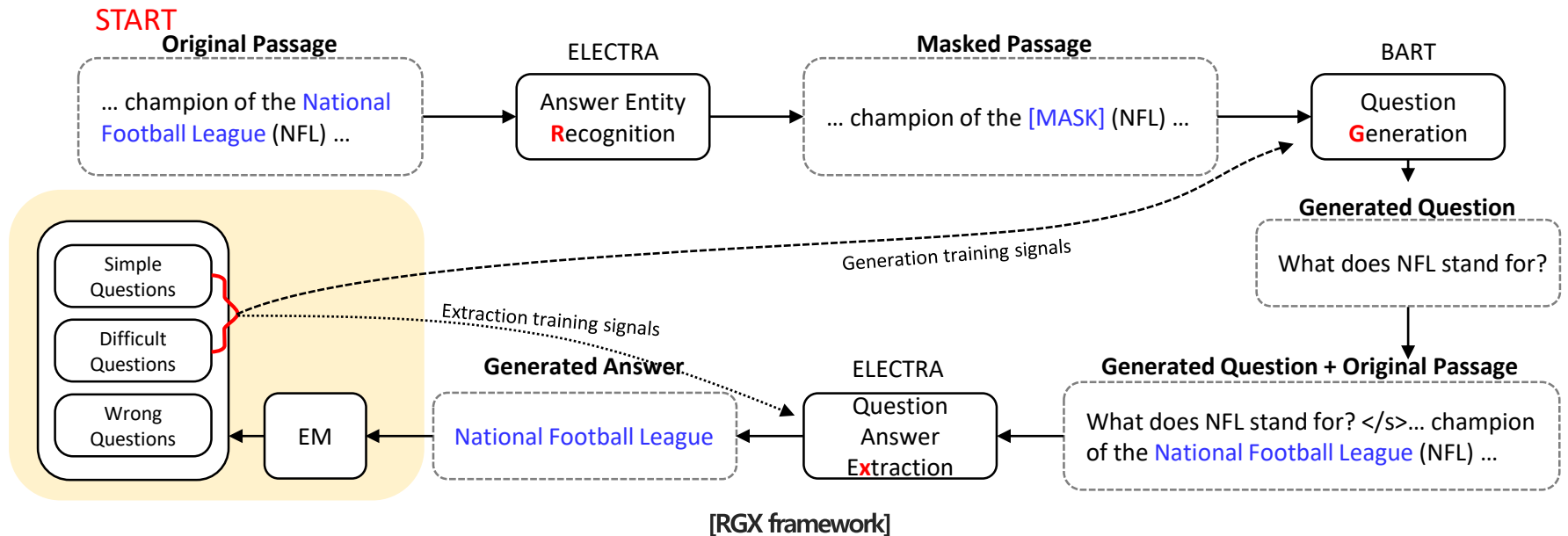


Question Answer Generation

Cooperative Self-training of Machine Reading Comprehension (2022, NAACL)

❖ RGX Framework 구조

- 모든 생성된 QA쌍이 학습에 유용하지 않기 때문에 모델이 학습하기 적합한 QA쌍을 선택
- Expectation-maximization (EM)을 통해 손실 값에 따른 질문을 그룹화
 - ✓ Low-loss 질문: QAE 모델이 쉽게 답변할 수 있는 간단한 질문들
 - ✓ Medium-loss 질문: QAE가 어느 정도 어려움을 겪는 도전적인 질문들
 - ✓ High-loss 질문: 잡음이 포함된 질문으로 문법적 오류 혹은 틀린 답변을 요구하는 질문들

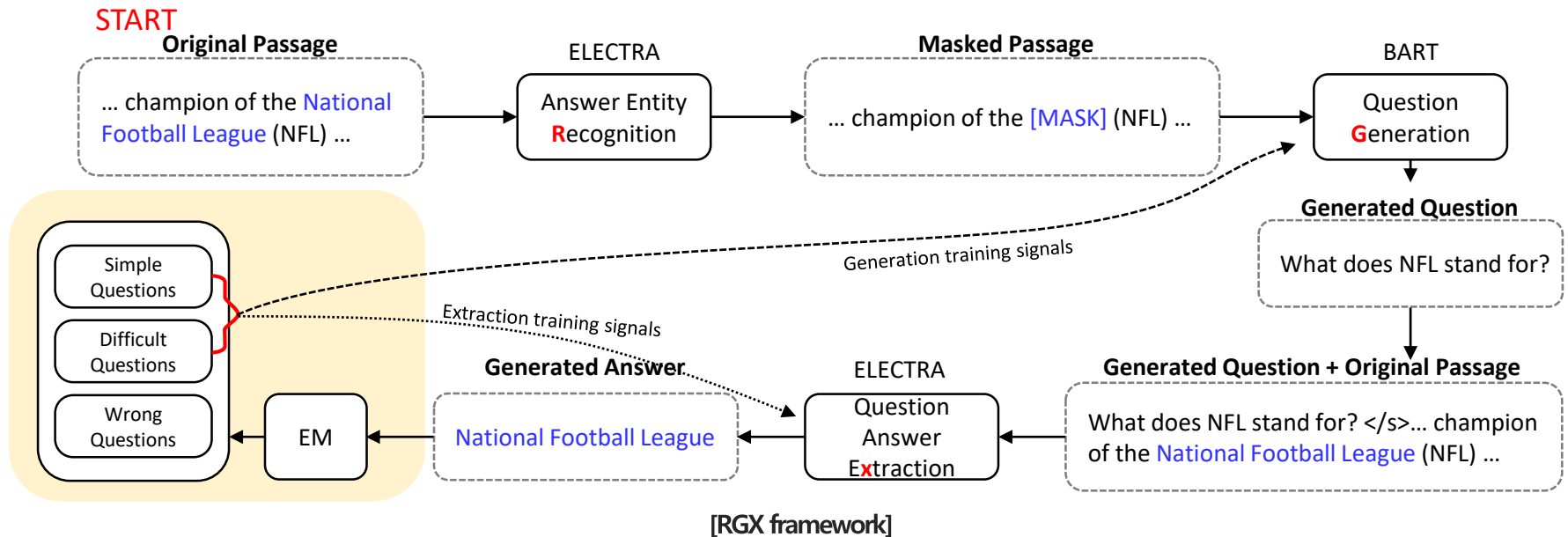


Question Answer Generation

Cooperative Self-training of Machine Reading Comprehension (2022, NAACL)

❖ RGX Framework 구조

- 모든 생성된 QA쌍이 학습에 유용하지 않기 때문에 모델이 학습하기 적합한 QA쌍을 선택
- Expectation-maximization (EM)을 통해 손실 값에 따른 질문을 그룹화
 - ✓ Low-loss 질문: QAE 모델이 쉽게 답변할 수 있는 간단한 질문들
 - ✓ Medium-loss 질문: QAE가 어느 정도 어려움을 겪는 도전적인 질문들
 - ✓ ~~High loss 질문: 잡음이 포함된 질문으로 문법적 오류 혹은 틀린 답변을 요구하는 질문들~~



Question Answer Generation

Cooperative Self-training of Machine Reading Comprehension (2022, NAACL)

❖ RGX Framework 결과

- ELECTRA: Natural Question과 SQuAD 데이터셋으로 사전 학습된 QA 모델
- QAGen2S: RGX와 같이 자가 학습을 통해 QA를 수행하는 모델
- SynQA: RGX와 QAGen2S와 같이 자가 학습을 통해 QA를 수행하는 모델
- 다양한 도메인에서 RGX 모델의 성능이 가장 우수한 모습을 보임

Model <i>Domain</i>	BioASQ <i>Bio</i>		TextbookQA <i>Book</i>		RACE <i>Exam</i>		RelExt. <i>Wiki</i>		DuoRC <i>Movie</i>		DROP <i>Wiki</i>		Avg	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Source Domain: NaturalQuestions _{wiki} , Method: Extraction														
ELECTRA	41.9	59.0	31.9	41.5	32.4	43.4	67.7	81.8	40.0	48.5	39.3	51.1	42.2	54.2
QAGen2S	43.2	64.1	39.9	51.7	33.7	45.5	71.6	84.4	43.8	53.2	24.2	37.1	42.7	56.0
RGX (Ours)	50.3	70.1	49.9	60.9	40.3	52.4	76.1	87.2	47.8	58.4	27.6	42.1	48.7	61.9
Source Domain: SQuAD _{wiki} (SQuAD+AQA+Wiki for SynQA), Method: Extraction														
ELECTRA	58.7	73.1	43.0	53.6	38.3	52.5	79.0	88.4	53.1	64.2	48.3	60.8	53.4	65.4
QAGen2S	56.8	71.7	48.0	56.5	43.4	54.9	73.4	84.8	53.3	64.6	42.2	54.5	52.8	64.5
SynQA	55.1	68.7	41.4	50.2	40.2	54.2	78.9	88.6	51.7	62.1	64.9	73.0	55.3	66.1
RGX (Ours)	60.3	74.8	51.2	61.2	44.9	58.7	79.2	88.6	57.4	66.2	47.6	60.9	56.8	68.4

[Out-of-domain 상황에서의 결과]

Question Answer Generation

Cooperative Self-training of Machine Reading Comprehension (2022, NAACL)

❖ RGX Framework 결과

- 실제 RGX를 통해 생성된 QA 데이터셋

The National History Museum of Montevideo is located in the historical residence of General Fructuoso Rivera. It exhibits artifacts related to the history of Uruguay. In a process begun in 1998, the National Museum of Natural History (1837) and the National Museum of Anthropology (1981), merged **in 2001**, becoming the National Museum of Natural History and Anthropology. In July 2009, the two institutions again became independent. The Historical Museum has annexed eight historical houses in the city, five of which are located in the Ciudad Vieja. One of them, on the same block with the main building, is the historic residence of Antonio Montero, which houses the Museo Romantico.

When was the national history museum of montevideo founded?

In the 1920s, John Maynard Keynes prompted a division between microeconomics and macroeconomics. Under Keynesian economics macroeconomic trends can overwhelm economic choices made by individuals. Governments should promote aggregate demand for goods as a means to encourage economic expansion. Following World War II, Milton Friedman created the concept of monetarism. Monetarism focuses on using the **supply and demand of money** as a method for controlling economic activity. In the 1970s, monetarism has adapted into supply-side economics which advocates reducing taxes as a means to increase the amount of money available for economic expansion.

Monarism focuses on the relationship between the?

Starting in 2006, Apple's industrial design shifted to favor aluminum, which was used in the construction of the first MacBook Pro. Glass was added in 2008 with the introduction of the unibody MacBook Pro. These materials are billed as environmentally friendly. The iMac, MacBook Pro, MacBook Air, and Mac Mini lines currently all use aluminum enclosures, and are now made of a single unibody. Chief designer **Jonathan Ive** continues to guide products towards a minimalist and simple feel, including eliminating of replaceable batteries in notebooks. Multi-touch gestures from the iPhone's interface have been applied to the Mac line in the form of touch pads on notebooks and the Magic Mouse and Magic Trackpad for desktops.

Who is the designer of the macbook pro?

The city's total area is 468.9 square miles (1,214 km²). 164.1 sq mi (425 km²) of this is water and 304.8 sq mi (789 km²) is land. The highest point in the city is Todt Hill **on Staten Island**, which, at 409.8 feet (124.9 m) above sea level, is the highest point on the Eastern Seaboard south of Maine. The summit of the ridge is mostly covered in woodlands as part of the Staten Island Greenbelt.

Where is the highest point in new york city?

Question Answer Generation

❖ LIQUID: A Framework for List Question Answering Dataset Generation

- 2023년 AAAI에 기재된 list 형식의 QA Dataset을 만들기 위한 Pipeline QAG 방법론을 제안한 논문

LIQUID: A Framework for List Question Answering Dataset Generation

Seongyun Lee,^{*1} Hyunjae Kim,^{*1} Jaewoo Kang^{1,2}

¹Korea University, ²AIGEN Sciences
{sy-lee, hyunjae-kim, kangj}@korea.ac.kr

Abstract

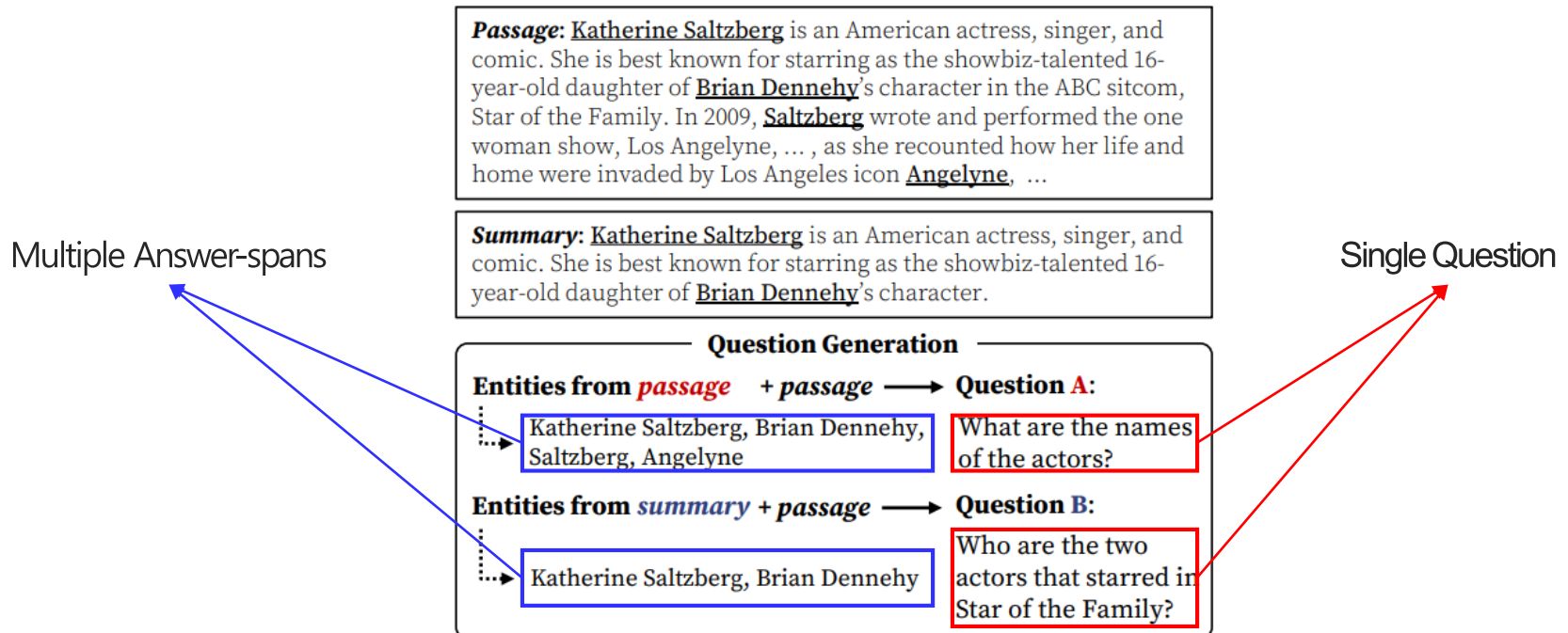
Question answering (QA) models often rely on large-scale training datasets, which necessitates the development of a data generation framework to reduce the cost of manual annotations. Although several recent studies have aimed to generate synthetic questions with single-span answers, no study has been conducted on the creation of list questions with multiple, non-contiguous spans as answers. To address this gap, we propose LIQUID, an automated framework for generating list QA datasets from unlabeled corpora. We first convert a passage from Wikipedia or PubMed into a summary and extract named entities from the summarized text as candidate answers. This allows us to select answers that are semantically correlated in context and is, therefore, suitable for constructing list questions. We then create questions using an off-the-shelf question generator with the extracted entities and original passage. Finally, iterative filtering and answer expansion are performed to ensure the accuracy and completeness of the answers. Using our synthetic data, we significantly improve the performance of the previous best list QA models by exact-match F1 scores of 5.0 on MultiSpanQA, 1.9 on Quoref, and 2.8 averaged across three BioASQ benchmarks.

Question Answer Generation

LIQUID: A Framework for List Question Answering Dataset Generation (2023, AAAI)

❖ LIQUID Framework 목적

- Passage를 이용하여 list 형식의 Question Answer dataset를 생성
- List Question Answer: 여러 답변을 요구하는 질문
- Pipeline QAG 형식

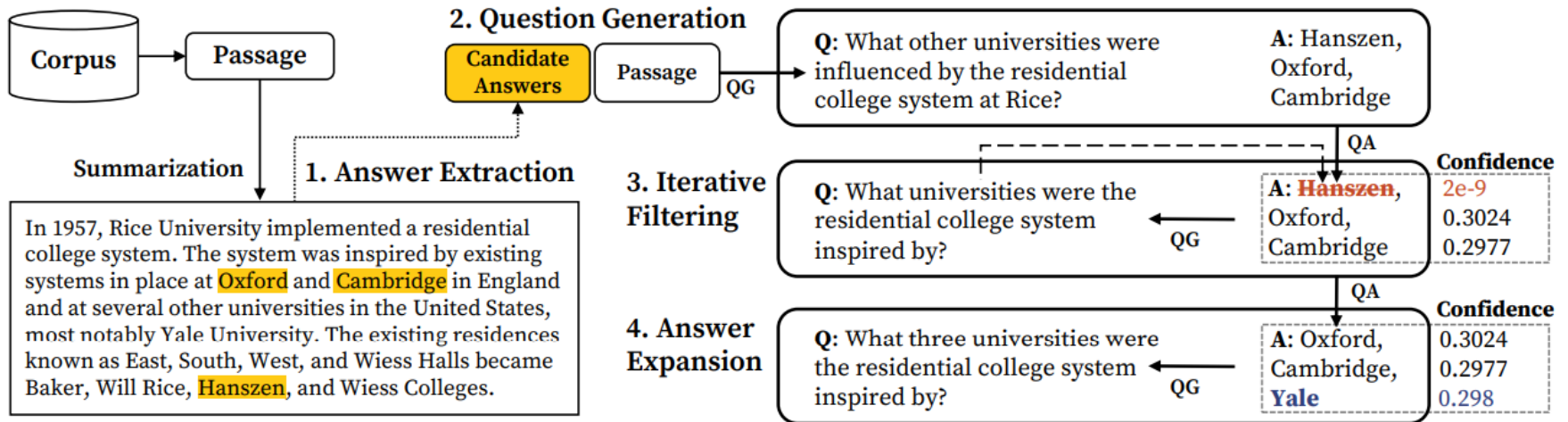


Question Answer Generation

LIQUID: A Framework for List Question Answering Dataset Generation (2023, AAAI)

❖ LIQUID Framework 구조

- 일반적인 pipeline QAG 방법론들의 구성을 가지고 있음
- Answer extraction: 다수의 답변 객체 추출
- Question generation: 질문 생성
- Iterative filtering & Answer Expansion: 생성된 데이터의 품질을 높이기 위한 방법



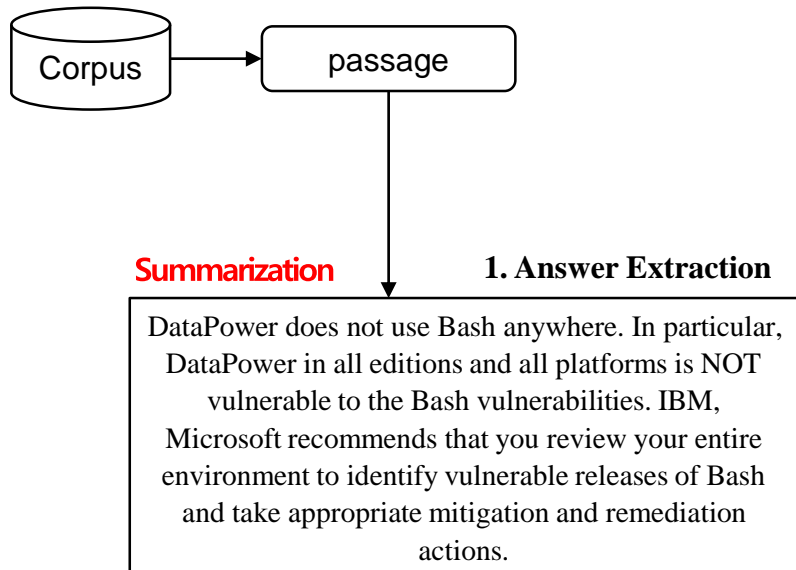
[LIQUID framework]

Question Answer Generation

LIQUID: A Framework for List Question Answering Dataset Generation (2023, AAAI)

❖ LIQUID Framework 구조

- 주어진 corpus의 passage (c)를 summarization하여 summarized passage(\bar{c})를 만들
- List QA의 답변 객체들은 서로 연관이 있어야 하기 때문
- CNN/Daily Mail 데이터셋으로 학습된 $BART_{base}$ 모델 사용



Summarization

1. Input text의 길이를 줄여 모델 학습 비용을 줄여 줄 수 있음
2. 필요 없는 정보들을 줄여 텍스트의 노이즈를 줄여줄 수 있음

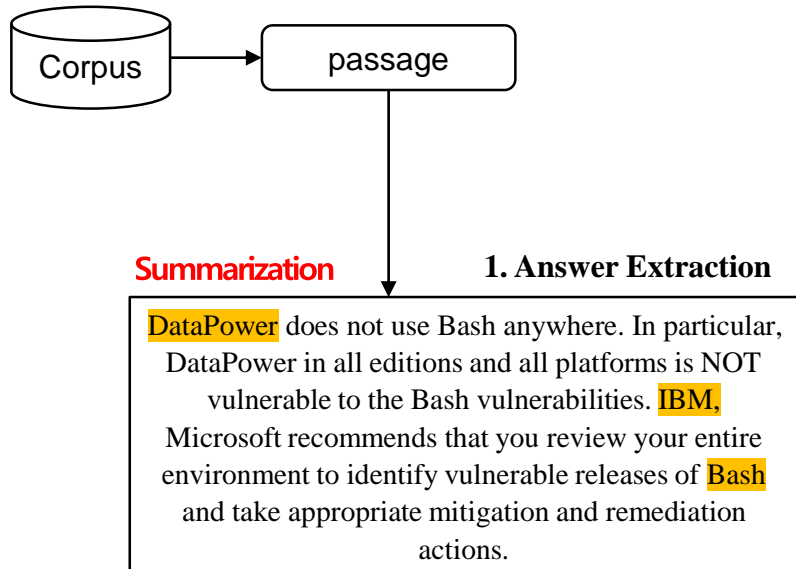
[LIQUID framework]

Question Answer Generation

LIQUID: A Framework for List Question Answering Dataset Generation (2023, AAAI)

❖ LIQUID Framework 구조

- Summarized passage (c)에서 NER 모델을 통해 답변 후보들을 선정
- 동일한 질문에 대한 답변 후보들은 동일한 entity 유형을 가질 가능성이 높음
 - ✓ 동일한 entity 유형을 가진 후보들을 선택
- 일반적인 데이터에는 spaCy NER tagger, biomedical 데이터에는 BERN2 모델을 활용



Answer extraction

1. A_1, \dots, A_L : 하나의 c 마다 L 개의 답변 후보 그룹을 구축
2. L : 미리 정의된 entity 유형 수(ex. object, name,...)
3. $A_l = \{a_1, \dots, a_{lm}\}$: l -번째 엔티티 집합을 의미

[LIQUID framework]

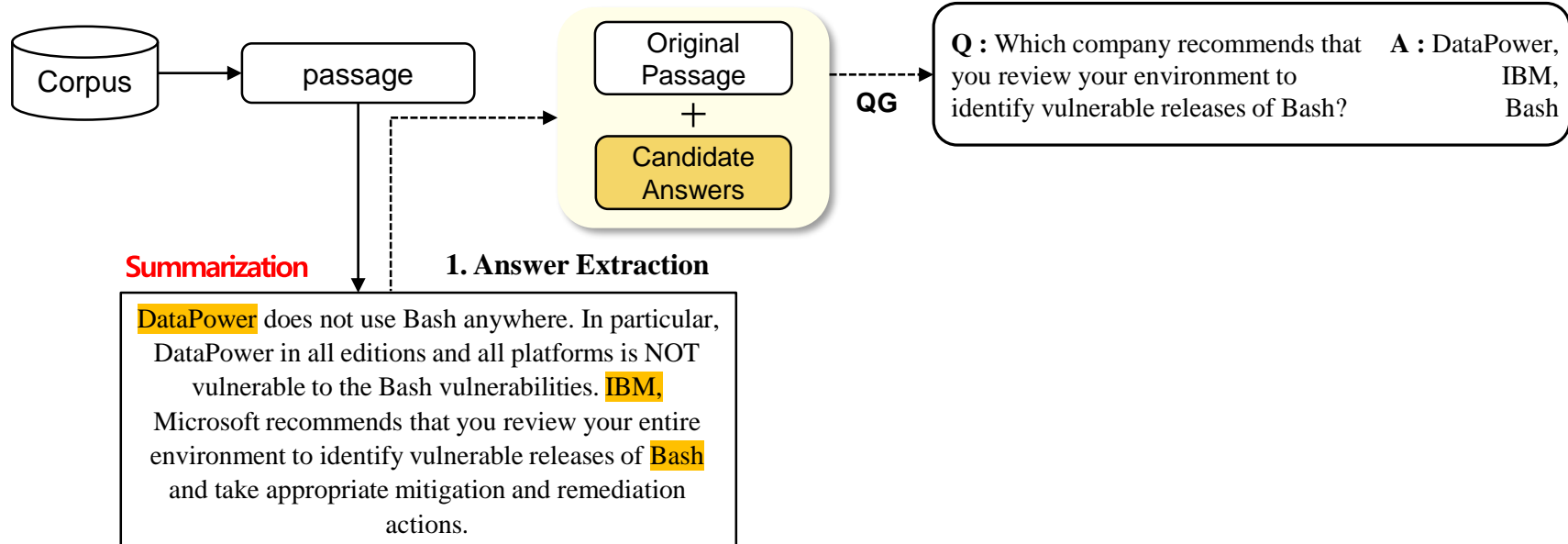
Question Answer Generation

LIQUID: A Framework for List Question Answering Dataset Generation (2023, AAAI)

❖ LIQUID Framework 구조

- 답변 후보 그룹과 Original Passage (c)를 인풋으로 질문 생성
- 기존의 QG 모델들은 1개의 답변 후보와 1개의 문장을 입력으로 받고 질문을 생성
- 위의 셋팅을 따라가기 위해 답변 후보들을 concatenate하여 하나의 답변 후보 형태를 만듦
- SQuAD 데이터셋로 학습된 $T5_{base}$ 모델을 활용

2. Question Generation



[LIQUID framework]

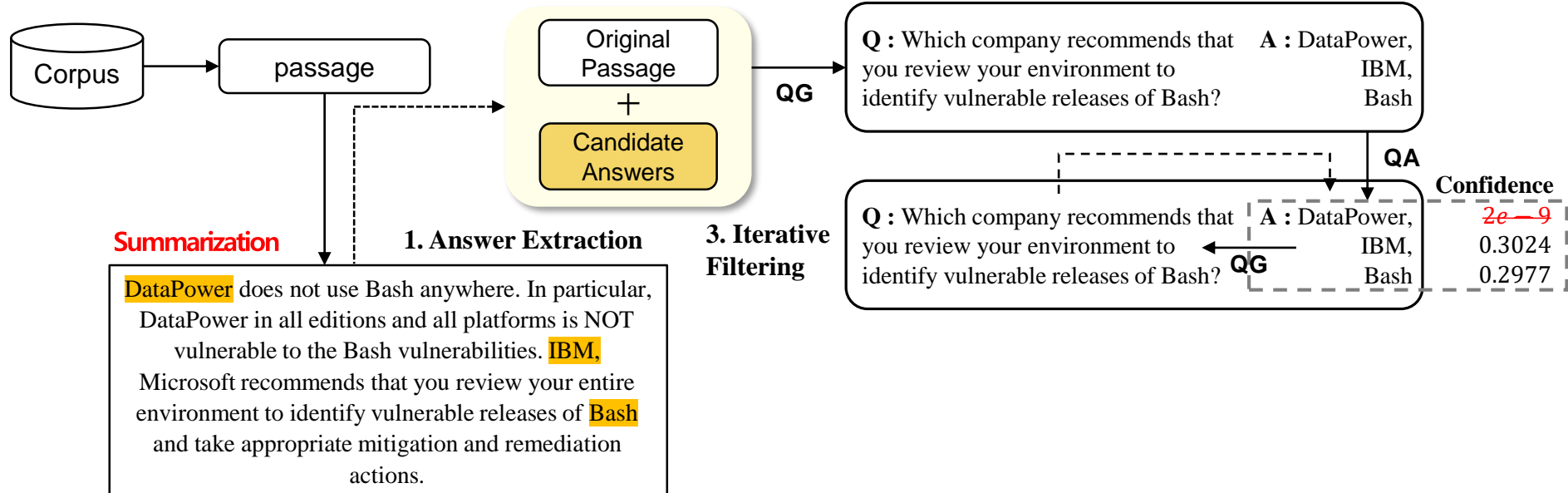
Question Answer Generation

LIQUID: A Framework for List Question Answering Dataset Generation (2023, AAAI)

❖ LIQUID Framework 구조

- QA 모델을 활용하여 답변 후보 그룹 중 부적절한 답변 후보를 필터링
- QA 모델을 사용하여 각 답변 후보에 대한 confidence score 측정
- 일정 threshold (τ)보다 작은 confidence score를 가진 답변 후보는 제외
- SQuAD 데이터셋으로 학습된 $RoBERTa_{base}$ 모델을 활용

2. Question Generation



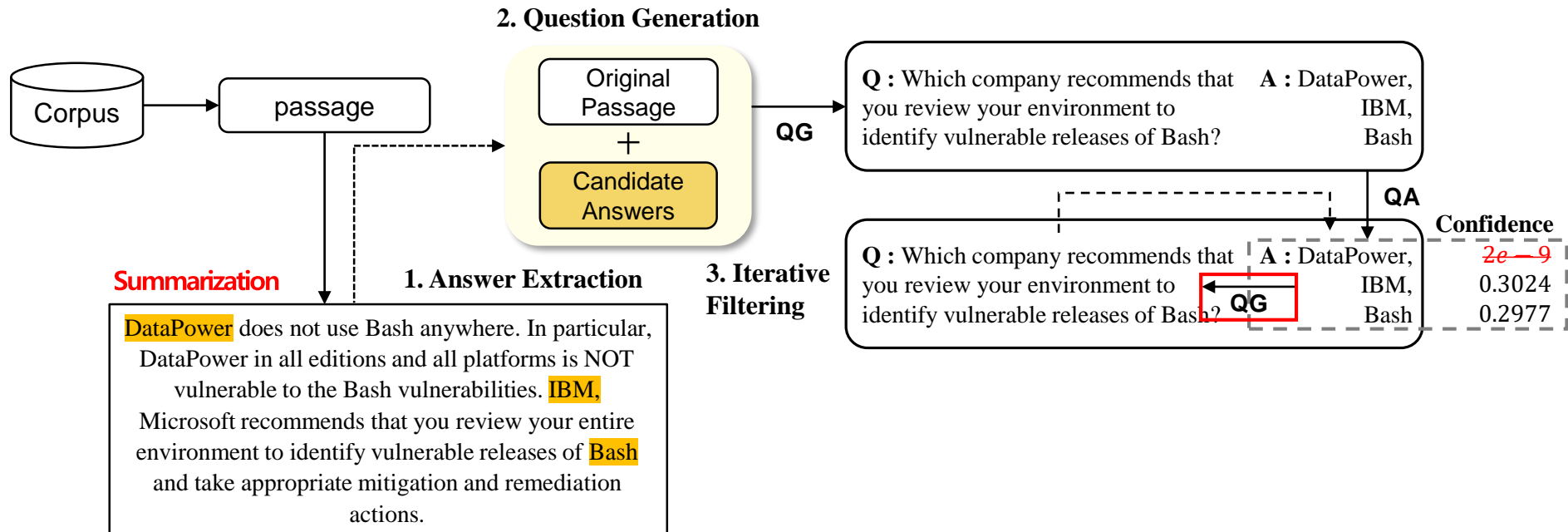
[LIQUID framework]

Question Answer Generation

LIQUID: A Framework for List Question Answering Dataset Generation (2023, AAAI)

❖ LIQUID Framework 구조

- 필터링 후의 답변 후보들과 생성된 질문의 align 맞지 않을 수 있기 때문에 질문을 다시 생성
- Iteration filtering은 이전 답변 후보들과 필터링 후보가 같거나 하이퍼 파라미터 T 만큼 반복 후 멈춤
- Original passage에서 답변 후보의 start와 end 포지션을 함께 저장
 - ✓ Summarized passage에서 뽑은 답변 후보를 사용했기 때문



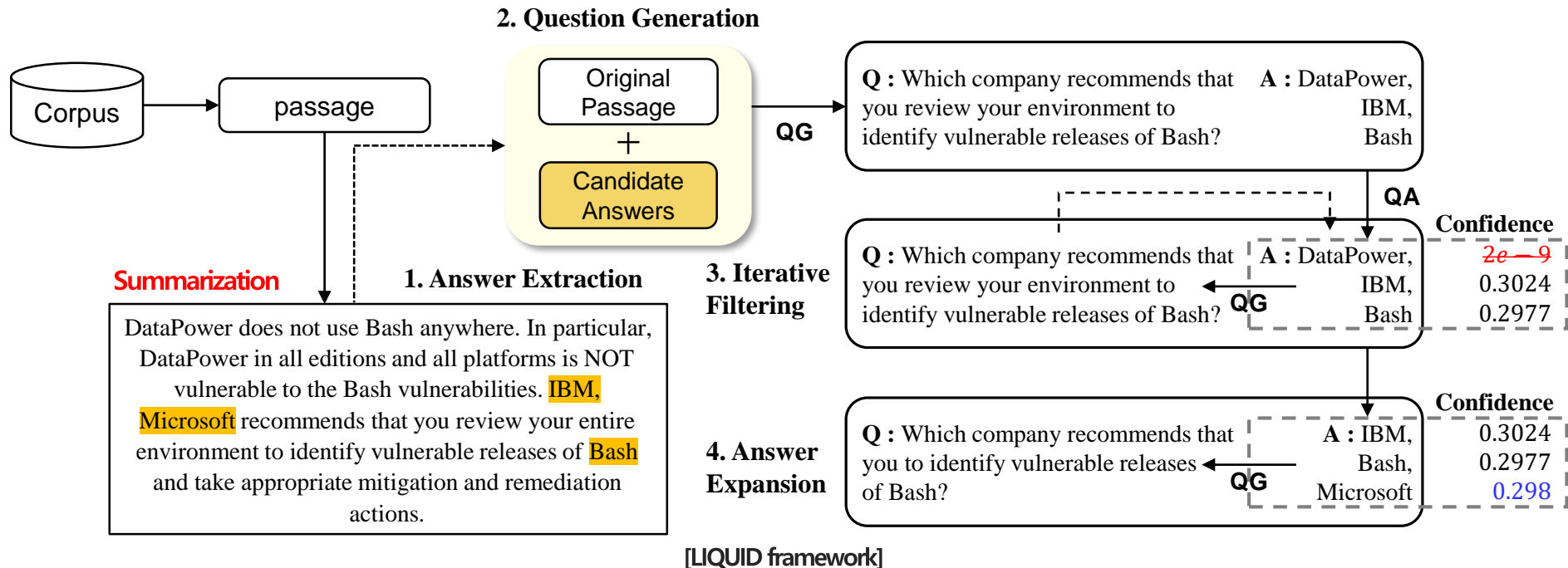
[LIQUID framework]

Question Answer Generation

LIQUID: A Framework for List Question Answering Dataset Generation (2023, AAAI)

❖ LIQUID Framework 구조

- 새로운 질문에 대해 초기 NER 모델이 놓친 답변 후보가 있을 가능성이 있음
- 필터링 된 답변 후보들 중 가장 낮은 신뢰도보다 높은 신뢰도를 갖는 답변 후보 식별
- 최종 답변 후보들을 이용하여 다시 한번 질문을 재생성 후 하나의 QA쌍을 구축



Question Answer Generation

LIQUID: A Framework for List Question Answering Dataset Generation (2023, AAAI)

❖ LIQUID Framework 결과

- 오리지널 데이터셋만으로 학습한 모델과 생성된 데이터를 활용하여 학습한 모델 결과 비교
- 대부분의 실험 결과에서 생성된 데이터를 활용하여 학습한 모델의 성능이 우수함

Model	MultiSpanQA		Quoref	
	Exact F1 (P/R)	Partial F1 (P/R)	Exact F1 (P/R)	Partial F1 (P/R)
<i>Baselines: labeled only (\mathcal{D})</i>				
BERT _{base} + Single-span*	14.4 (16.2/13.0)	67.6 (60.3/76.8)	-	-
BERT _{base} + Tagger*	56.5 (52.5/61.1)	75.2 (75.9/74.5)	-	-
BERT _{base} + Tagger (multi-task)*	59.3 (58.1/60.5)	76.3 (79.6/73.2)	-	-
RoBERTa _{base} + Single-span	10.5 (14.4/8.3)	63.0 (60.0/66.3)	55.4 (65.2/48.0)	69.0 (76.7/62.6)
RoBERTa _{base} + Tagger	62.9 (63.0/62.9)	78.0 (82.5/73.9)	81.2 (73.8/90.1)	85.7 (80.1/92.2)
RoBERTa _{large} + Tagger	66.4 (62.3/71.2)	82.6 (82.1/83.0)	84.2 (76.1/94.2)	88.8 (82.6/96.0)
CorefRoBERTa _{large} + Tagger	64.0 (56.5/73.8)	81.7 (77.7/86.0)	86.5 (81.3/92.4)	89.7 (86.1/93.7)
<i>Our models: synthetic & labeled ($\tilde{\mathcal{D}} \rightarrow \mathcal{D}$)</i>				
RoBERTa _{base} + Single-span	19.4 (19.7/19.0)	71.0 (62.9/81.4)	60.7 (63.8/57.9)	74.3 (77.4/71.3)
RoBERTa _{base} + Tagger	67.4 (65.7/69.2)	81.2 (80.9/81.5)	85.7 (82.3/89.3)	89.1 (86.5/91.8)
RoBERTa _{large} + Tagger	71.4 (75.0/68.2)	80.9 (85.3/77.0)	86.7 (85.8/87.6)	90.2 (89.4/91.1)
CorefRoBERTa _{large} + Tagger	65.8 (64.0/67.8)	80.2 (79.8/80.5)	88.4 (84.8/92.2)	91.7 (89.1/94.4)

[일반 도메인 데이터셋에 대한 실험 결과]

Question Answer Generation

LIQUID: A Framework for List Question Answering Dataset Generation (2023, AAAI)

❖ LIQUID Framework 결과

- 오리지널 데이터셋만으로 학습한 모델과 생성된 데이터를 활용하여 학습한 모델 결과 비교
- 바이오 도메인에서는 특히 더 생성된 데이터를 사용했을 때의 효과가 확실하게 드러남

Model	BioASQ 7b		BioASQ 8b		BioASQ 9b	
	Exact F1 (P/R)	Partial F1 (P/R)	Exact F1 (P/R)	Partial F1 (P/R)	Exact F1 (P/R)	Partial F1 (P/R)
<i>Baselines: labeled only (\mathcal{D})</i>						
BioBERT _{base} + Single-span	42.1 (55.9/33.8)	60.2 (82.3/47.5)	34.4 (44.9/27.9)	53.5 (40.2/79.9)	56.1 (46.2/71.3)	73.8 (70.3/77.7)
BioBERT _{base} + Tagger	46.1 (39.7/55.1)	70.5 (68.5/72.6)	41.8 (33.5/55.5)	67.6 (64.0/71.5)	66.7 (60.1/74.9)	80.6 (76.4/85.2)
BioBERT _{large} + Tagger	49.5 (40.5/63.6)	74.6 (70.7/78.9)	45.0 (34.7/64.0)	72.2 (65.8/80.0)	68.2 (60.9/77.5)	81.4 (76.3/87.2)
<i>Our models: synthetic & labeled ($\tilde{\mathcal{D}} \rightarrow \mathcal{D}$)</i>						
BioBERT _{base} + Single-span	51.8 (49.0/55.0)	70.2 (69.7/70.7)	44.2 (41.4/47.5)	65.2 (65.4/65.0)	64.0 (58.0/71.4)	76.6 (72.6/81.1)
BioBERT _{base} + Tagger	49.0 (41.0/61.0)	73.1 (70.4/76.0)	44.2 (36.6/55.8)	69.4 (67.3/71.7)	71.5 (67.0/76.6)	83.2 (80.0/86.7)
BioBERT _{large} + Tagger	52.3 (44.5/63.5)	74.9 (71.9/78.1)	46.5 (38.5/58.8)	72.3 (68.9/76.1)	72.2 (67.3/77.8)	83.4 (80.4/86.7)

[바이오 도메인 데이터셋에 대한 실험 결과]

Question Answer Generation

LIQUID: A Framework for List Question Answering Dataset Generation (2023, AAAI)

❖ LIQUID Framework 결과

- 실제 LIQUID를 통해서 생성된 QA 데이터셋
- Bold는 답변 객체를 의미

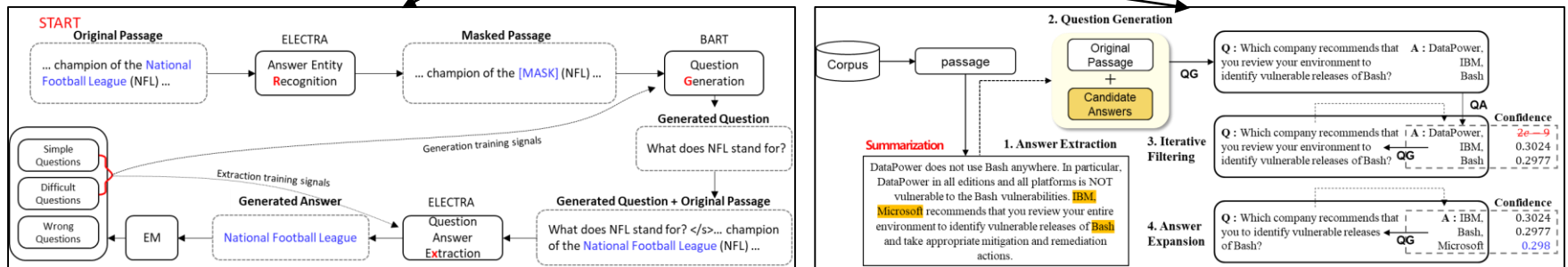
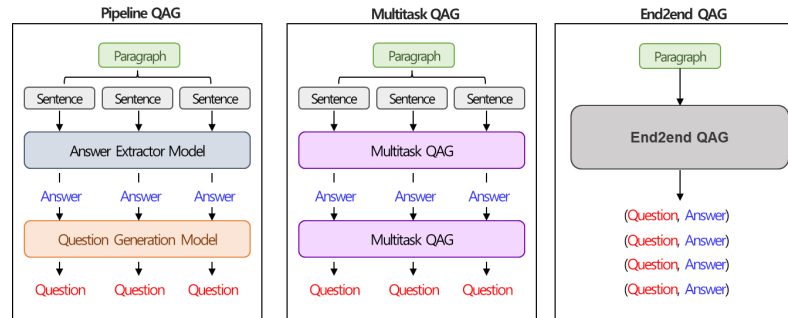
Question Type	Passage & Answer Spans	Question
Simple Questions	Ya Rab is a 2014 Bollywood movie directed by Hasnain Hyderabadwala starring Ajaz Khan, Arjumman Mughal, Raju Kher, Vikram Singh (actor), Imran Hasnee . . .	Who starred in Ya Rab?
Lexical Variation	. . . In June 2007, a Hackday event was hosted at Alexandra Palace by the BBC and Yahoo . . .	What <i>media companies</i> hosted a Hackday event in 2007?
Inter-sentence Reasoning	. . . SBOBET was the shirt sponsor of West Ham United . up until the end of 2012-2013 season. They were also the shirt sponsor of Cardiff City for 2010-2011 season . . .	What teams did SBOBET sponsor?
Number of Answers	. . . While working with her mother, Bundy's uncle offered to pay for her to attend any cookery school in the world. She was accepted into and attended Le Cordon Bleu and Le Notre in Paris, training at Fauchon Patisserie . . .	What <i>two</i> French cookery schools did Bundy attend?
Entailment	. . . Around the same time, Zhao Yun also came to Ye (present-day Handan, Hebei), Yuan Shao's headquarters, where he met Liu Bei again . . .	Who were the people who came to Ye?

Conclusion

LM-based Question Answer Generation

❖ Question Answer Generation

- QA 데이터셋을 만드는 프로세스는 단순해 보이지만 복잡한 작업
- 사용자가 원하는 QA 형태, 종류에 따라 다양한 방식의 생성 기법을 적용할 수 있음
- 본 세미나에서는 대표적인 QA 데이터셋 생성 기법들의 구조와 과정에 대해 설명



고맙습니다
