Classic Question Answering

건국대학교 컴퓨터공학부 /
KAIST 전산학부 (겸직)

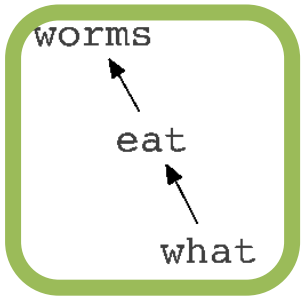
김학수

What is Question Answering?

One of the oldest NLP tasks (punched card systems in 1961)

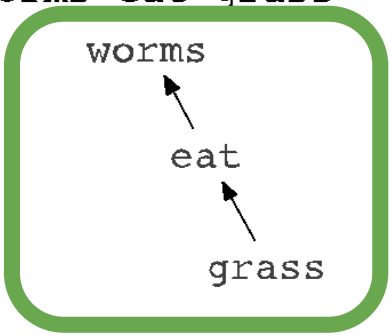
Question:

What do worms eat?

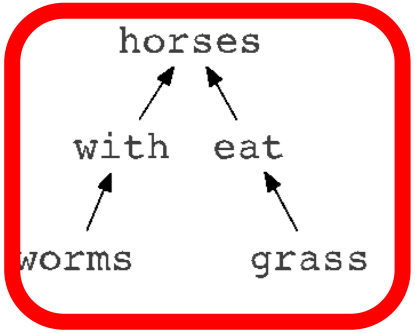


Potential Answers:

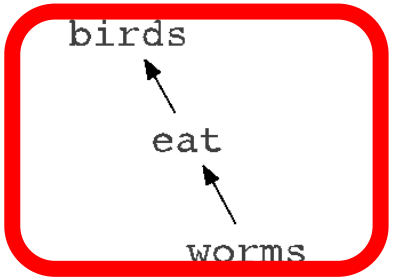
Worms eat grass



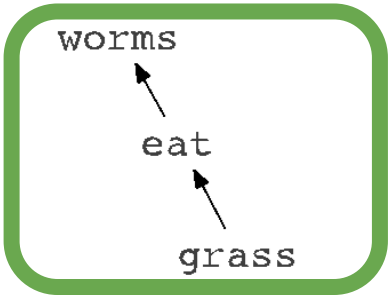
Horses with worms eat grass



Birds eat worms



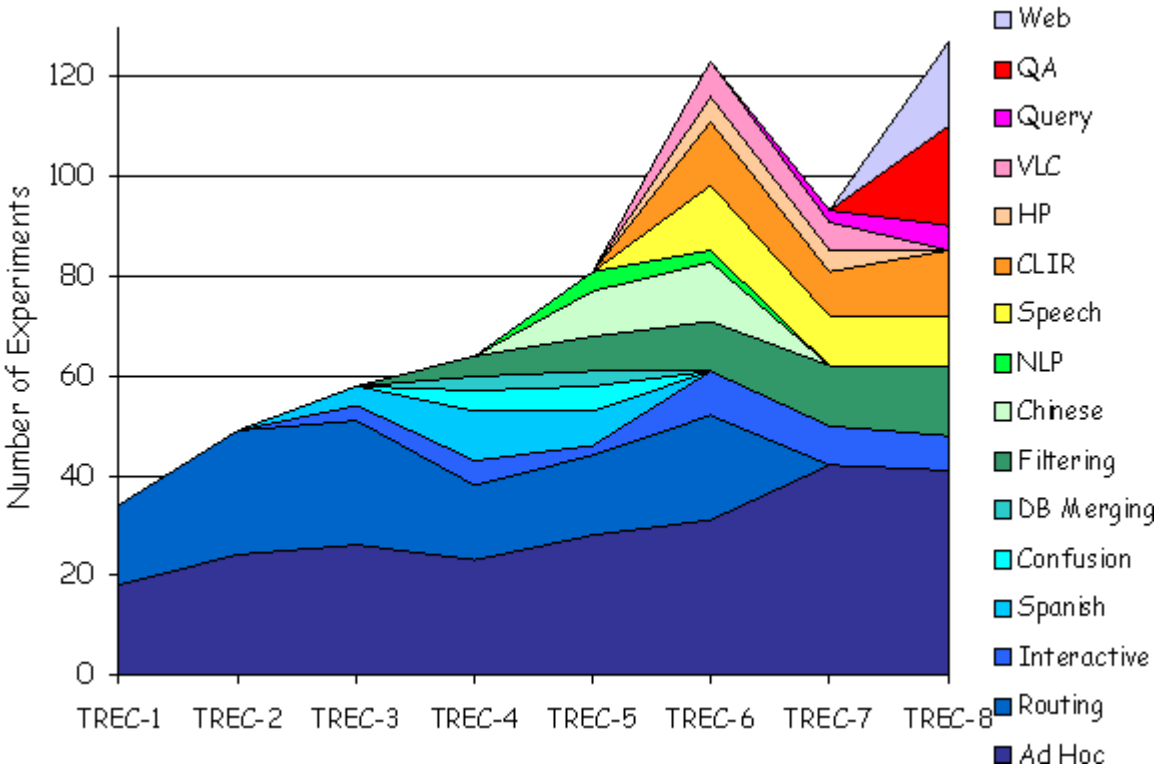
Grass is eaten by worms



Simmons, Klein, McConlogue. 1964. Indexing and Dependency Logic for Answering English Questions. American Documentation 15:30, 196-204

Question Answering in TREC

- TREC(Text Retrieval Conference) QA Track
  - Annual competition from 1992

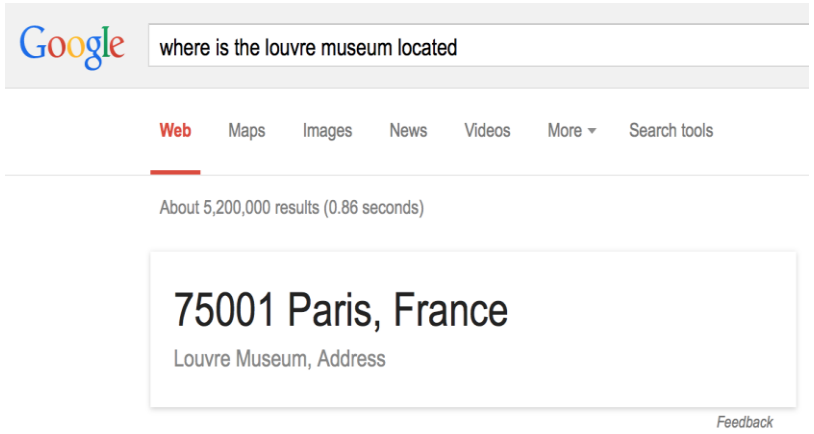


QA Track in TREC

	1999	2001	2003	2005	2006
Tasks	50-Byte/250-Byte	Main/List/Context	Main/Passage	Main/Document/Relationship	Main/ciQA
Document Collection	TREC-8 Ad Hoc Collection (TREC disks 4,5; 528,000 documents; 2GB)	TIPSTER/TREC (979,000 documents; 3GB)	Corpus of English News (1,033,000 documents; 3GB)	Same as TREC 2003	Same as TREC 2003
# of Questions	198 factoids	Main: 500 List:25 Context:42	Main (Factoid: 413 List:37 Definition:50) Passage: 413	Main (Factoid: 362 List: 93 Other:75) Document: 50 Relationship: 25	Main (Factoid: 403 List:89 Other:75) ciQA: 25
Question Source	FAQ Finder Log, Assessors, Participants	MSNSearch and AskJeeves Logs	AOL and MSNSearch Logs	Same as TREC 2003	Same as TREC 2005
Correctness Judgments	"correct" if string contains right answers; unsupported string are correct	Main/Context: Correct/ Incorrect/ Unsupported (Lenient: unsupported=correct; Strict: unsupported=incorrect) List: Correctness/Distinctness	Main (Factoid/list: Incorrect/Unsupported/Inexact/Correct) Definition: "Information nuggets" created and marked by assessors Passage: Incorrect/Unsupported/Correct	Main (Factoid/list: Incorrect/Unsupported/Inexact/Correct Other: same as TREC 2003 definition task) Document: relevant/not relevant Relationship: same as other task	Main (Factoid/list: Incorrect/Unsupported/Inexact/locally Correct/globally Correct Other: same as TREC 2005) ciQA: same as other task
Evaluation Measures	MRR	Main/Context: MRR List: Average Accuracy	Main: FinalScore= 0.5*FactoidScore+0.25*ListScore+0.25*DefScore Passage: Accuracy	Main: FinalScore =0.5*Factoid+0.25*List+0.25*Other Document: R-Prec, MAP Relationship:F(β=3)	Main: FinalScore =1/3*Factoid+1/3*List+1/3*Other ciQA: Pyramid F-Score
Best Main Task Results	50-Byte: 0.66(MRR) 250-Byte: 0.646(MRR)	0.68(MRR)	Final: 0.559 (Factoid:0.7, list:0.396, Def:0.442)	Final: 0.534 (Factoid:0.713, list:0.468,Other:0.248)	Final: 0.394 (Factoid:0.578,list:0.433 ,Other:0.250)

Paradigms for Question Answering

- IR-based approaches
  - TREC; IBM Watson; Google



- Knowledge-based and Hybrid approaches
  - IBM Watson; Apple Siri; Wolfram Alpha;
  - True Knowledge Evi

IR-based Factoid QA

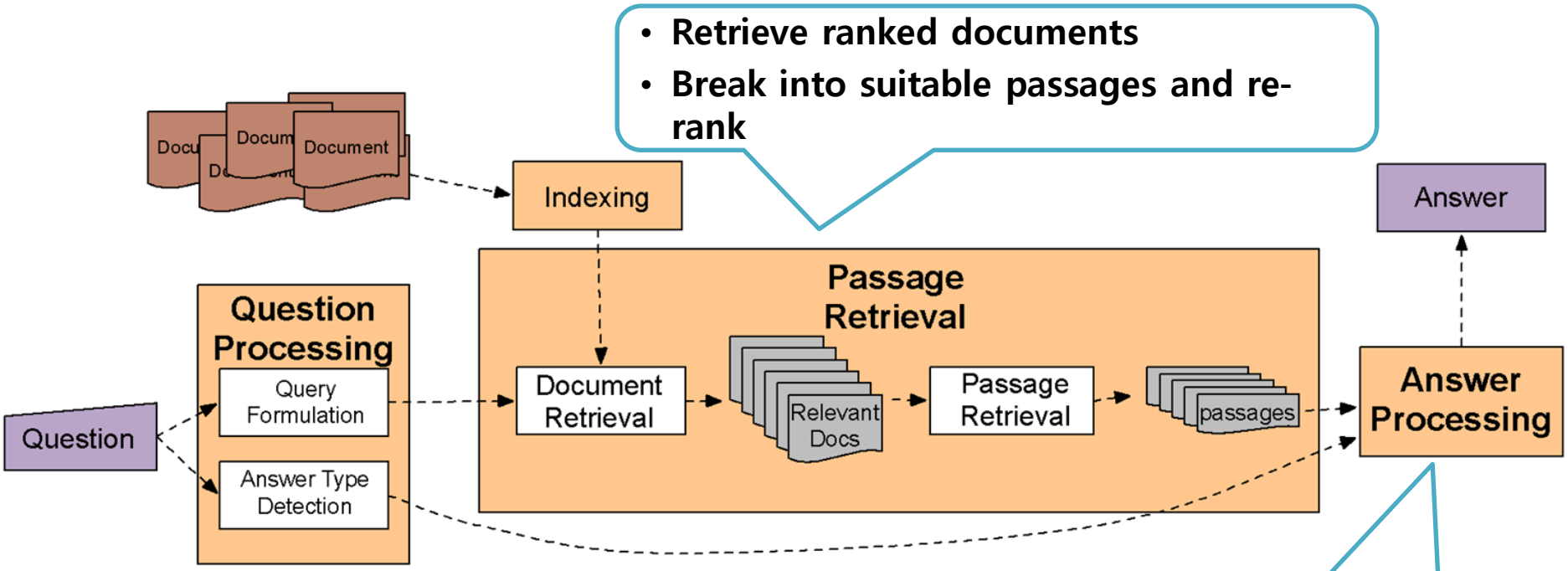


Figure by Jurafsky

Question Processing

- Answer Type Detection
  - Decide the **named entity type** (person, place) of the answer
- Query Formulation
  - Choose **query keywords** for the IR system
- Question Type classification
  - Is this a definition question, a math question, a list question?
- Focus Detection
  - Find the question words that are replaced by the answer
- Relation Extraction
  - Find relations between entities in the question

Question Processing

What are the two states you could be reentering if you're crossing Florida's northern border?

- **Question Type: What (Factoid)**
- **Answer Type: Location-State**
- **Query: two states, border, Florida, north**
- **Focus: the two states**
- **Relations: borders(Florida, ?x, north)**

Answer Type Taxonomy

- 6 coarse classes
  - ABBEVIATION, ENTITY, DESCRIPTION, HUMAN LOCATION, NUMERIC
- 50 finer classes
  - LOCATION: city, country, mountain…
  - HUMAN: group, individual, title, description
  - ENTITY: animal, body, color, currency…

- **NER Performance?**
- **Can all types be predefined?**

Ambiguity

What do bats eat? – *food*, *plant*, or *animal*

Xin Li, Dan Roth. 2002. Learning Question Classifiers. COLING'02

Passage Retrieval

- Step 1: IR engine retrieves documents using query terms
- Step 2: Segment the documents into shorter units
  - Something like paragraphs
- Step 3: Passage ranking
  - Use answer type to help re-rank passages

Features for Passage Ranking

- Number of named entities of the right type in passage
- Number of query words in passage
- Number of question *n*-grams also in passage
- Proximity of query keywords to each other in passage
- Longest sequence of question words
- Rank of the document containing passage

Answer Processing

- Run an answer-type named-entity tagger on the passages

Q: Who was Queen Victoria's second son?

→ Answer Type: Person

Passage:

The Marie biscuit is named after Marie Alexandrovna, the da ughter of Czar Alexander II of Russia and wife of Alfred, the second son of Queen Victoria and Prince Albert

Features for Ranking Candidate Answers

- **Answer type match**: Candidate contains a phrase with the correct answer type
- **Pattern match**: Regular expression pattern matches the candidate.
- **Question keywords**: # of question keywords in the candidate
- **Keyword distance**: Distance in words between the candidate and query keywords
- **Novelty factor**: A word in the candidate is not in the query
- **Apposition features**: The candidate is an appositive to question terms
- **Punctuation location**: The candidate is immediately followed by a comma, period, quotation marks, semicolon, or exclamation mark
- **Sequences of question terms**: The length of the longest sequence of question terms that occurs in the candidate answer

Learning Surface Text Patterns

DISCOVERER

1.0 when <ANSWER> discovered <NAME>
1.0 <ANSWER> ' s discovery of <NAME>
1.0 <ANSWER> , the discoverer of <NAME>
1.0 <ANSWER> discovers <NAME> .
1.0 <ANSWER> discover <NAME>
1.0 <ANSWER> discovered <NAME> , the
1.0 discovery of <NAME> by <ANSWER>.
0.95 <NAME> was discovered by <ANSWER>
0.91 of <ANSWER> ' s <NAME>
0.9 <NAME> was discovered by <ANSWER> in

BIRTHYEAR

1.0 <NAME> (<ANSWER> -)
0.85 <NAME> was born on <ANSWER> ,
0.6 <NAME> was born in <ANSWER>
0.59 <NAME> was born <ANSWER>
0.53 <ANSWER> <NAME> was born
0.5 - <NAME> (<ANSWER>
0.36 <NAME> (<ANSWER> -
0.32 <NAME> (<ANSWER>) ,
0.28 born in <ANSWER> , <NAME>
0.2 of <NAME> (<ANSWER>

Question type	Number of questions	MRR on TREC docs
BIRTHYEAR	8	0.48
INVENTOR	6	0.17
DISCOVERER	4	0.13
DEFINITION	102	0.34
WHY-FAMOUS	3	0.33
LOCATION	16	0.75

State of the Art in TREC-9

- 50-byte answers
  - MRR=0.58
  - No correct answer was found for 34% of questions

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N}$$

- 250-byte answers
  - MRR=0.76
  - No correct answer was found for 14% of questions

Knowledge-based QA

- Build a semantic representation of the query
  - Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data or resources
  - Geospatial databases
  - Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
  - Restaurant review sources and reservation services
  - Scientific databases

NL to SPARQL

- Answers: Databases of Relations
  - born-in("Emma Goldman", "June 27 1869")
  - author-of("Cao Xue Qin", "Dream of the Red Chamber")
  - Draw from Wikipedia infoboxes, DBpedia, FreeBase, etc.

- Questions: Extracting Relations in Questions

  Whose granddaughter starred in E.T.?

  → (acted-in ?x "E.T.") (granddaughter-of ?x ?y)

Machine Reading Comprehension

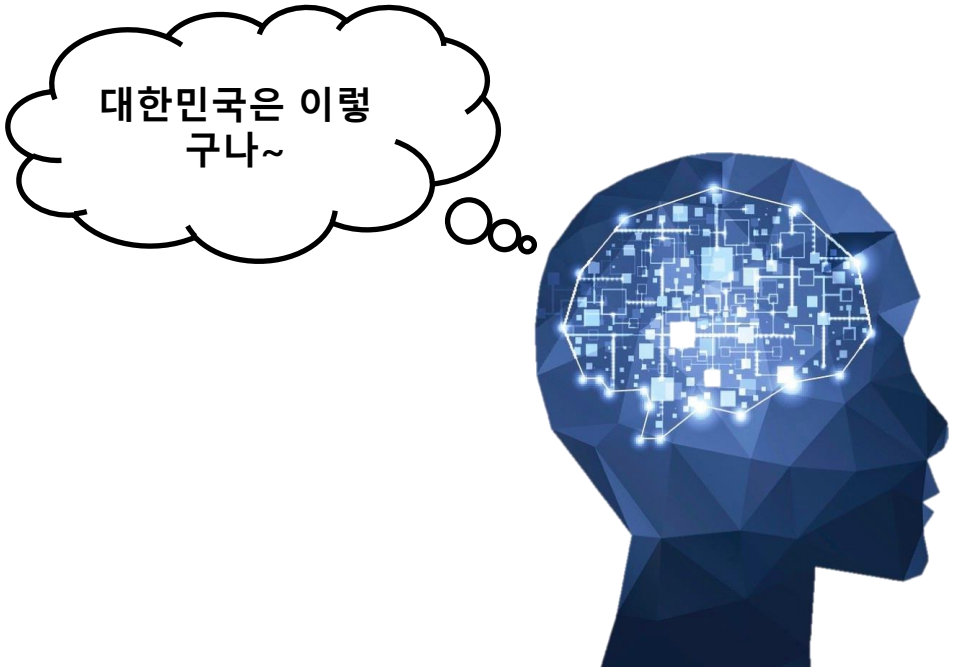
건국대학교 컴퓨터공학부 /
KAIST 전산학부 (겸직)

김학수

자연어처리연구실

Machine Reading Comprehension

- 주어진 문서를 기계가 이해하고 관련 질문을 답변해주는 시스템



대한민국

위키백과, 우리 모두의 백과사전.

다른 뜻에 대해서는 대한민국 (동음이의) 문서를 참조하십시오.

대한민국(듣기 (도움말·정보)), **大韓民國**, 영어: Republic of Korea; **ROK**, 문화어: 남조선; **南朝鮮**), 약칭으로 **한국**(**韓國**), **남한**(**南韓**)은 동아시아의 한반도 남부에 있는 공화국이다. 서쪽으로는 서해를 사이에 두고 중화인민공화국이, 동쪽으로는 동해를 사이에 두고 일본이 있으며 북쪽으로는 조선민주주의인민공화국과 맞닿아 있다. 수도는 서울특별시이며, 국기는 태극기, 국가는 애국가, 공용어는 한국어이다.

대한민국은 한반도의 북위 38도선 이남 지역 거주자들의 자유로운 선거(5.10 총선거)를 통하여 1948년 8월 15일에 공식적인 민주주의 국가로 출범하였다. 대한민국헌법전문에 따르면 대한국민은 3.1운동으로 건립된 대한민국 임시 정부의 법통을 계승한다. 대한민국은 1948년 12월 유엔 총회 결의 제195호를 통해 유엔으로부터 한반도 주민 대다수의 자유로운 의사에 따라 탄생한 합법 정부이자 한반도에서 유일한 정부로 승인을 받았다. 1991년 대한민국과 조선민주주의인민공화국은 동시에 UN에 가입하였다. 한편 국제법상의 통설과 대한민국의 헌법재판소 판결에 따르면, 조선민주주의인민공화국이 UN에 가입하였다 하여 다른 가맹국들에게도 그 국가성을 승인해야 할 의무가 있는 것은 아니며, 대한민국은 조선민주주의인민공화국의 국가성을 원칙적으로 부정하고 있다.

대한민국은 한국 전쟁 이래 일명 '한강의 기적'이라고 불리는 높은 경제 발전을 이룩하며, 1990년대에 이르러 세계적인 경제 강국으로 발전하였다. 2015년 구매력 기준 1인당 국민 총소득(GDP)은 36,601달러로[4] 세계은행에서 고소득 국가로 분류되었고, 2016년 유엔의 인간 개발 지수(HDI) 조사에서 세계 18위로 '매우 높음'으로 분류되었다.[3] 또한, 국제 통화 기금(IMF)에서는 대한민국을 선진 경제국으로

대한민국
大韓民國
국기 / 국장
국가 애국가

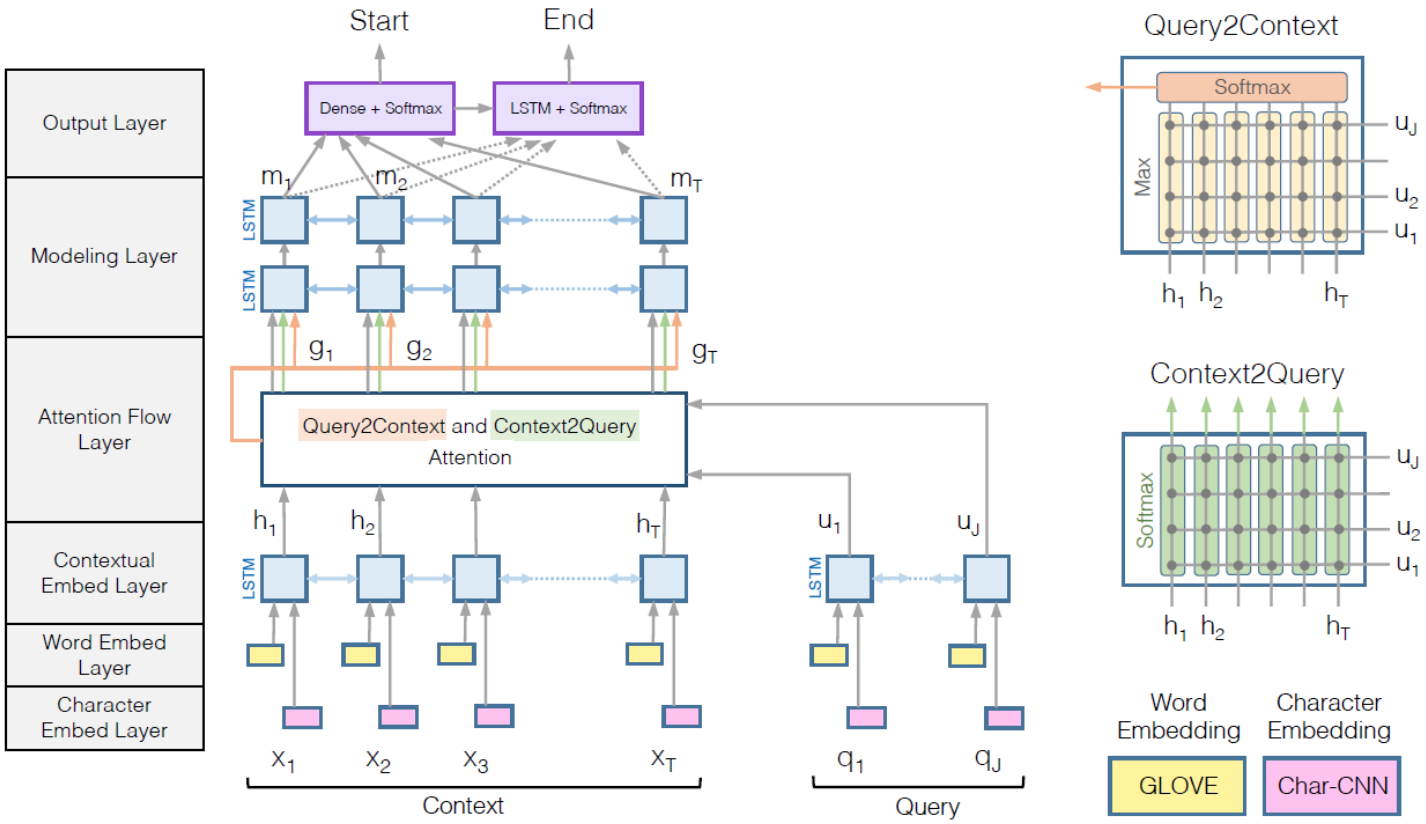
MRC 예제

Q: 대한민국 서쪽에는 어느 나라가 있나?

대한민국(大韓民國, 영어: Republic of Korea; ROK, 문화어: 남조선; 南朝鮮), 약칭으로 한국(韓國), 남한(南韓)은 동아시아의 한반도 남부에 있는 공화국이다. 서쪽으로는 서해를 사이에 두고 중화인민공화국이, 동쪽으로는 동해를 사이에 두고 일본이 있으며 북쪽으로는 조선민주주의인민공화국과 맞닿아 있다. 수도는 서울특별시이며, 국기는 태극기, 국가는 애국가, 공용어는 한국어이다.

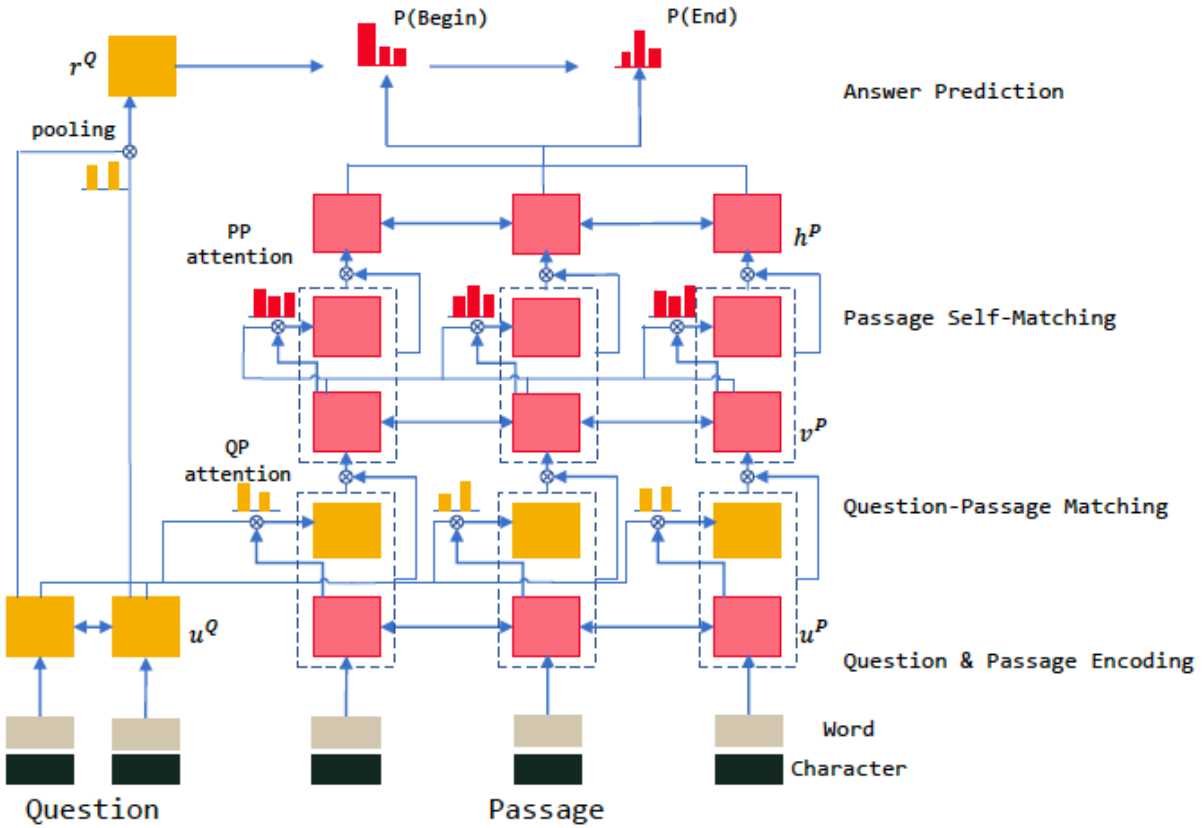
Stage 1: Initial Models

- Bi-Directional Attention Flow (Seo et al., 2017)



Stage 1: Initial Models

- R-Net (Wang et al., 2017)



모델들의 공통적인 구조

- **Encoding layer**
  - 질의와 문맥을 벡터로 표현
- **Co-attention layer**
  - 상호 Attention을 통해 문맥과 질의 간의 관계 파악
- **Output layer**
  - 질문에 해당하는 정답 단어의 시작과 끝 위치 출력

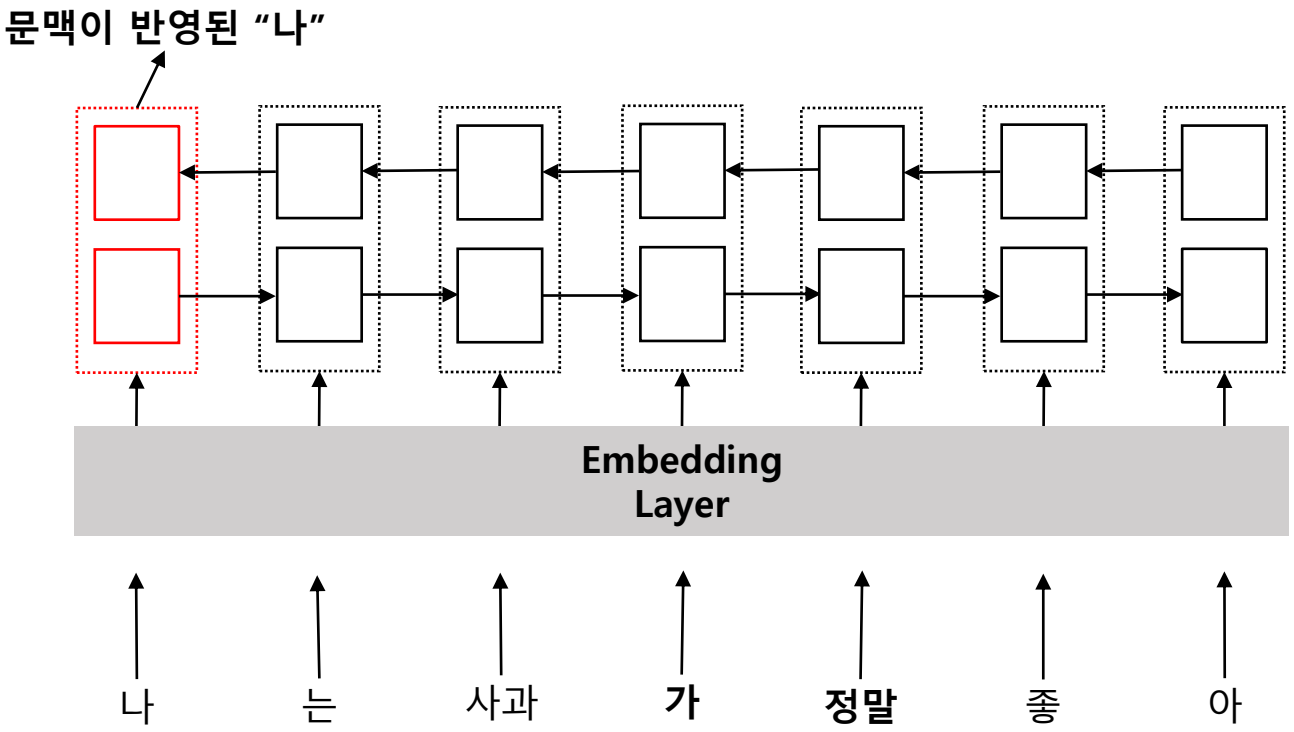
Encoding Layer

- Word Embedding
  - 단어를 기계가 알아들을 수 있는 벡터로 표현

- Recurrent Neural Network
  - 단어 벡터들을 연결하여 문장 벡터를 생성

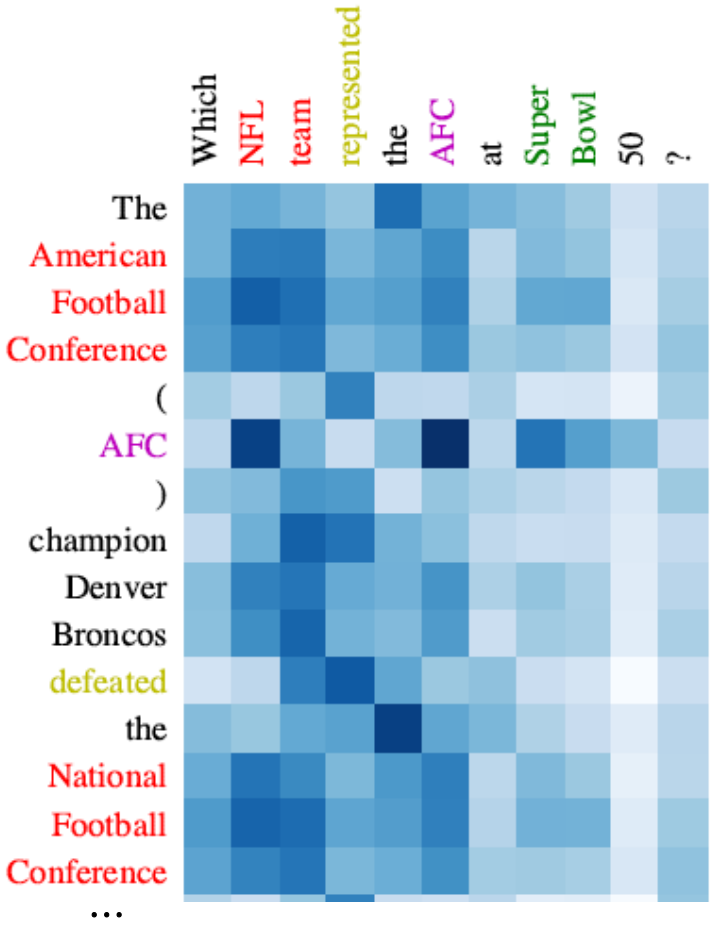
Sentence Embedding based on LSTM

- 입력된 임베딩 값을 통해 문맥 정보를 반영



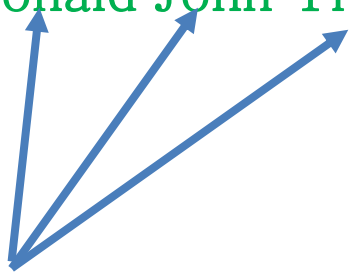
Co-Attention Layer

- Attention mechanism



Attention Mechanism

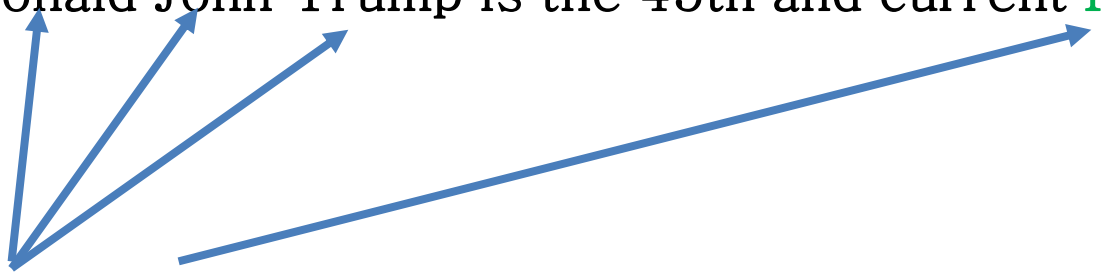
C Donald John Trump is the 45th and current President of the USA.



Q Who leads the United States?

Attention Mechanism

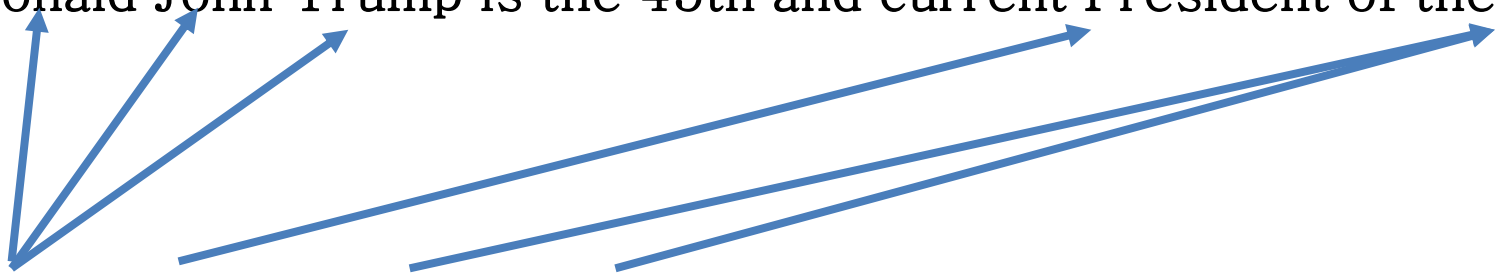
C Donald John Trump is the 45th and current President of the USA.



Q Who leads the United States?

Attention Mechanism

C Donald John Trump is the 45th and current President of the USA.



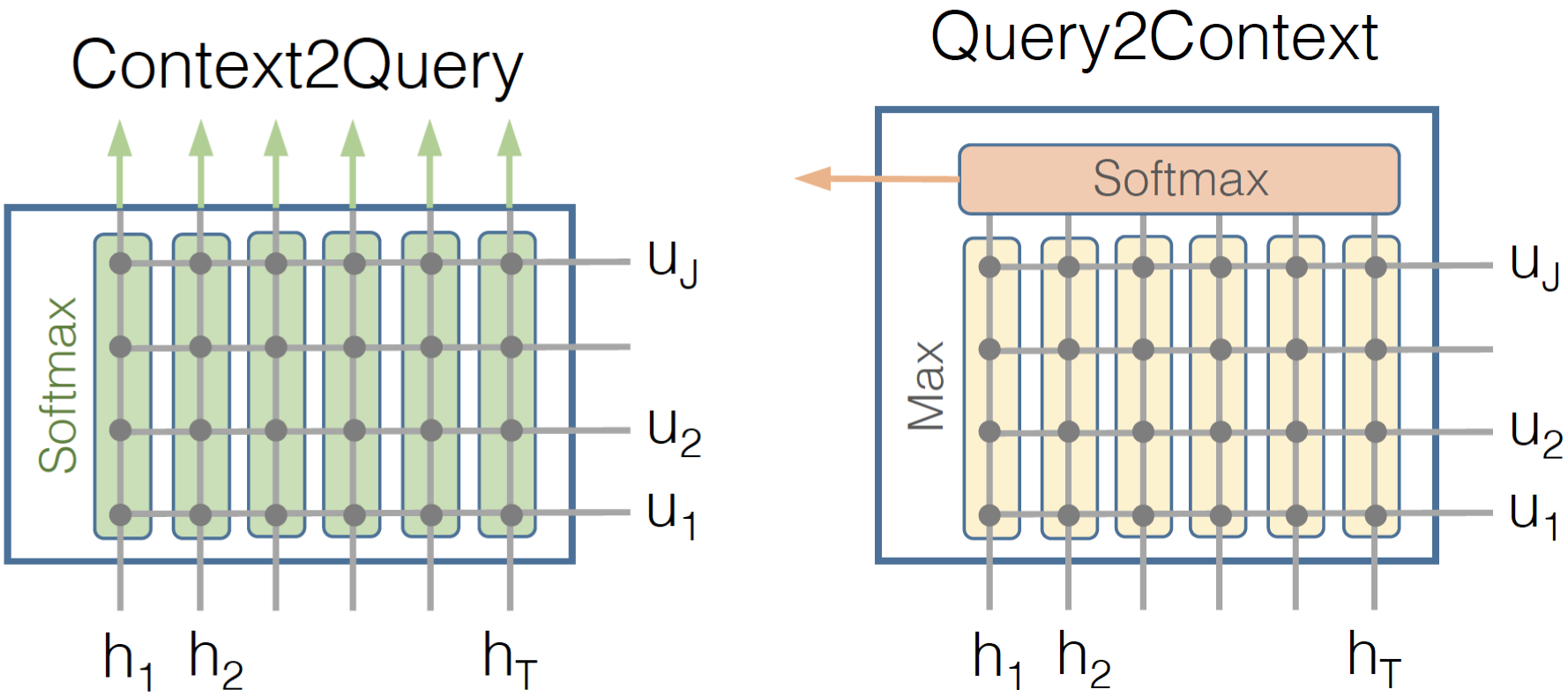
Q Who leads the United States?

Representative Co-Attention Methods

- Bi-directional Attention
- Fully-aware Attention
- Self Attention

Bi-directional Attention

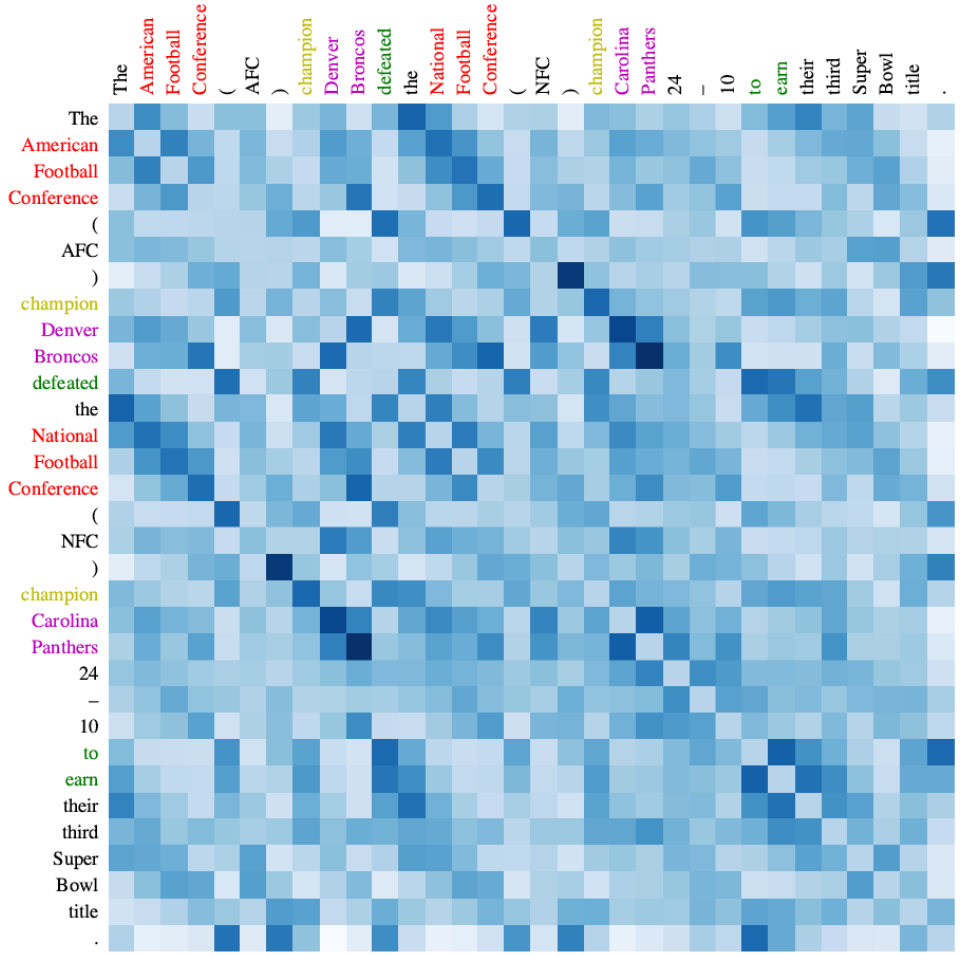
- Bi-Directional Attention Flow (Seo et al., 2017)



Self Attention

- R-Net (Wang et al., 2017)
  - 자기 **자신과 Attention**
  - 기존 Attention은 **서로 다른 문장** 간의 관계
  - Self Attention은 **같은 문장 내에 단어들** 간의 관계

Self Attention



Output Layer

- 정답이 나타나는 곳의 위치를 찍어주기

C Donald John Trump is the 45th and current President of the USA.

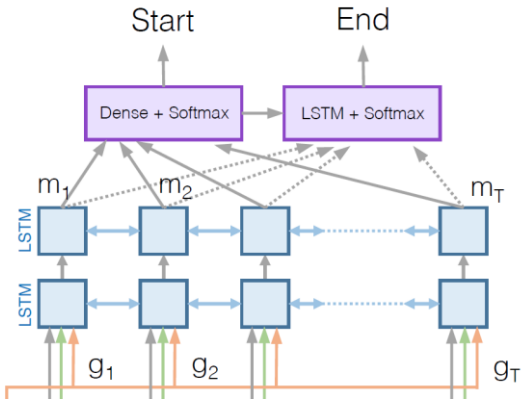
Q Who leads the United States?

How to Point Answers

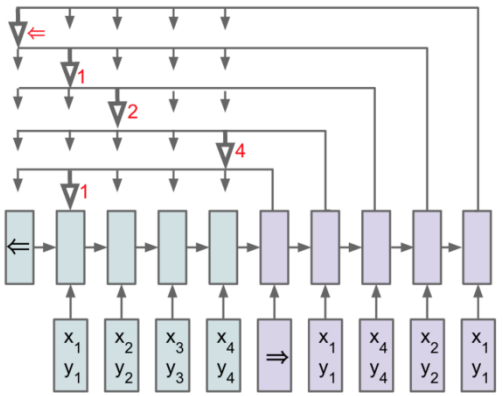
- Start, End의 확률 분포를 통해 계산

$$\mathbf{p}^1 = \mathrm{softmax}(\mathbf{w}_{(\mathbf{p}^1)}^\top [\mathbf{G}; \mathbf{M}])$$

$$\mathbf{p}^2 = \mathrm{softmax}(\mathbf{w}_{(\mathbf{p}^2)}^\top [\mathbf{G}; \mathbf{M}^2])$$

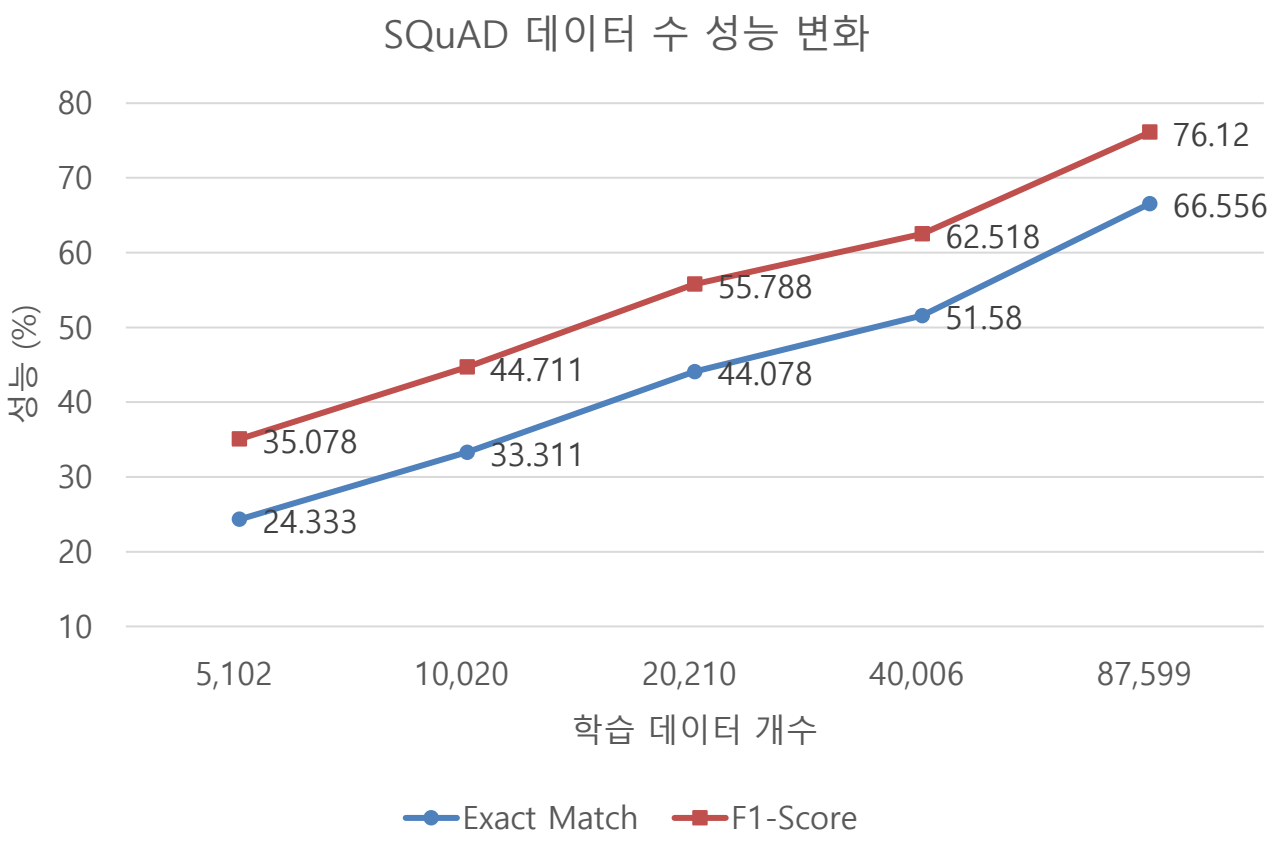


- Pointer Networks 통해 계산

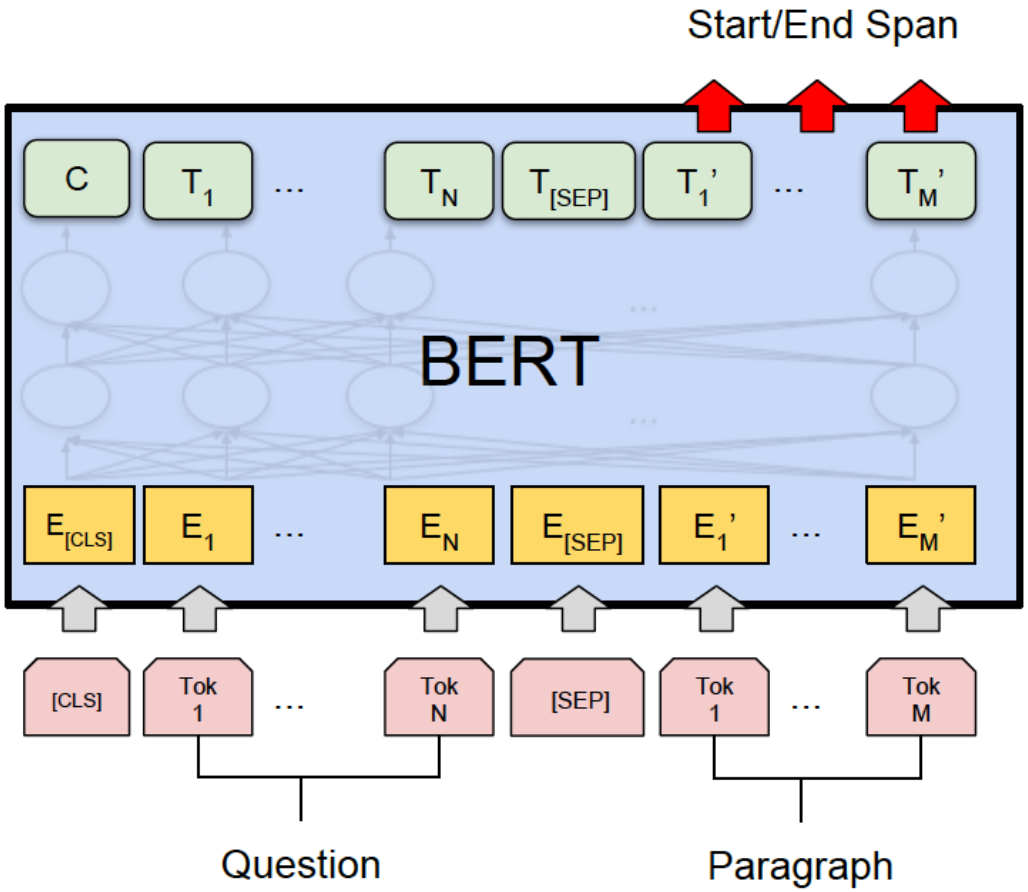


Size of Training Data for MRC

- BiDAF(single model)



Stage 2: Span Prediction Using PTM



SQuAD v1.1 Performance

System	Dev		Test	
	EM	F1	EM	F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

SQuAD 2.0

New **SQuAD2.0** combines the 100,000 questions in SQuAD1.1 with over 50,000 new, unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering. SQuAD2.0 is a challenging natural language understanding task for existing models, and we release SQuAD2.0 to the community as the successor to SQuAD1.1. We are optimistic that this new dataset will encourage the development of reading comprehension systems that know what they don't know.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452

SQuAD 2.0 Models

- No Answer를 답해줄 수 있는 옵션 추가
  - No-answer probability를 통한 무응답
  - Answer Verifier
  - No-answer span 추가

No-answer probability

- Loss function (Clark et al., 2018)

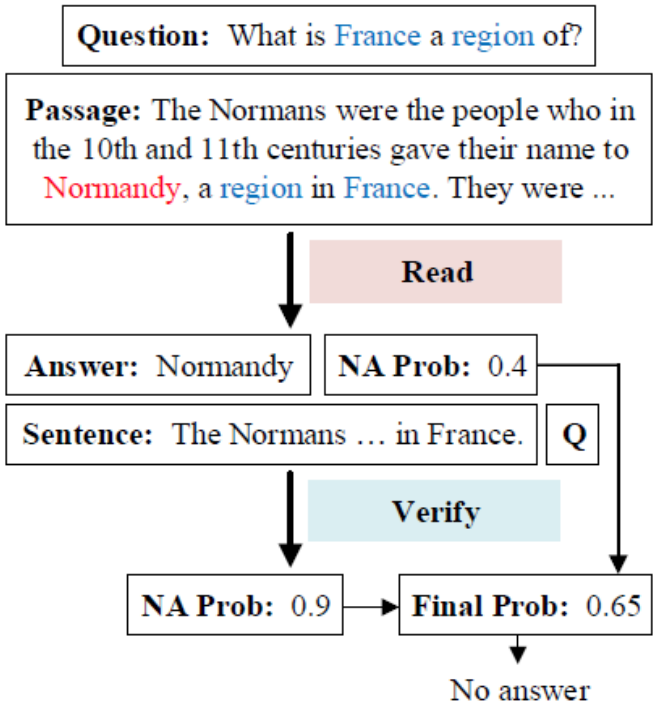
$$\mathcal{L}_{joint} = -\log\left(\frac{(1-\delta)e^{z} + \delta e^{s_a g_b}}{e^{z} + \sum_{i=1}^{n}\sum_{j=1}^{n} e^{s_i g_j}}\right)$$

No-answer Prob.

Answer Prob. (Start, End)

Answer Verifier

- Read + Verify: Machine Reading Comprehension with Unanswerable Questions (Hu et.al 2018)



MRC 성능을 향상 시키는 방법

- 모델 개선을 통한 성능 향상
  - 모델 구조 개선 및 추가 자질 사용
- 학습 데이터 수를 증가시켜 성능 향상
  - 기계독해 학습 데이터를 증가시켜 향상
- 학습 방법 개선을 통한 성능 향상
  - 기계독해와 질의 생성을 통한 상호 피드백 학습

새로운 도메인 적응

- 데이터의 도메인이 바뀔 시 성능 감소가 발생함

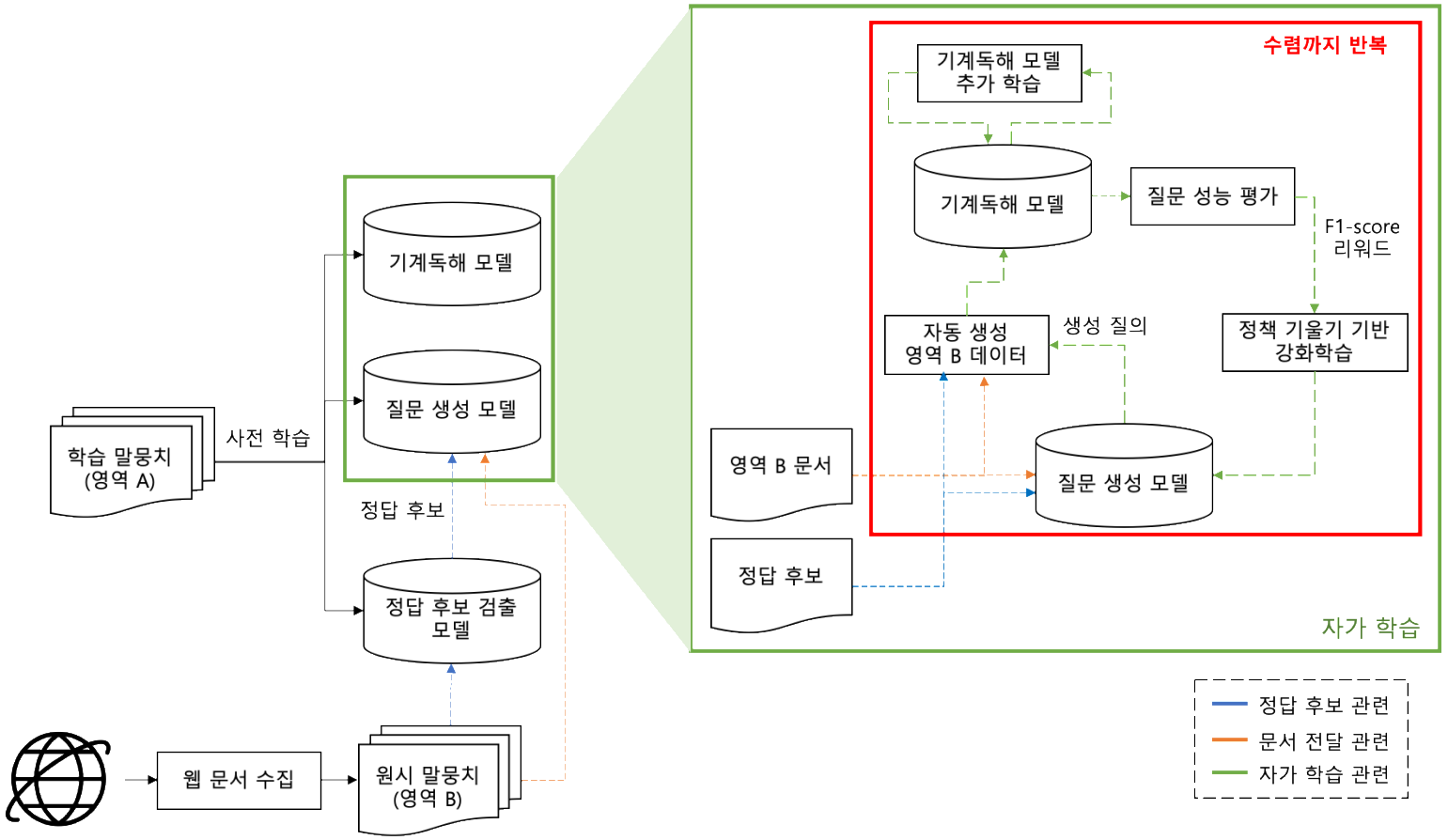
학습 데이터	평가 데이터	Exact Match(%)	F1-score(%)
일반 상식	일반 상식 평가 데이터	76.86	91.09
	뉴스, 민원 평가 데이터	45.66	68.70
뉴스, 민원	뉴스, 민원 평가 데이터	75.65	90.23

일반 상식 데이터 : KorQuAD v1.0
뉴스, 민원 데이터 : 자체 구축한 데이터

데이터 이슈

- 기계독해에서 충분한 성능을 달성하려면 60,000여개의 데이터가 필요
- 데이터를 구축하려면 시간 및 비용이 많이 발생
- 데이터를 자동으로 생성하고 검증하여 성능을 향상시킬 필요가 있음

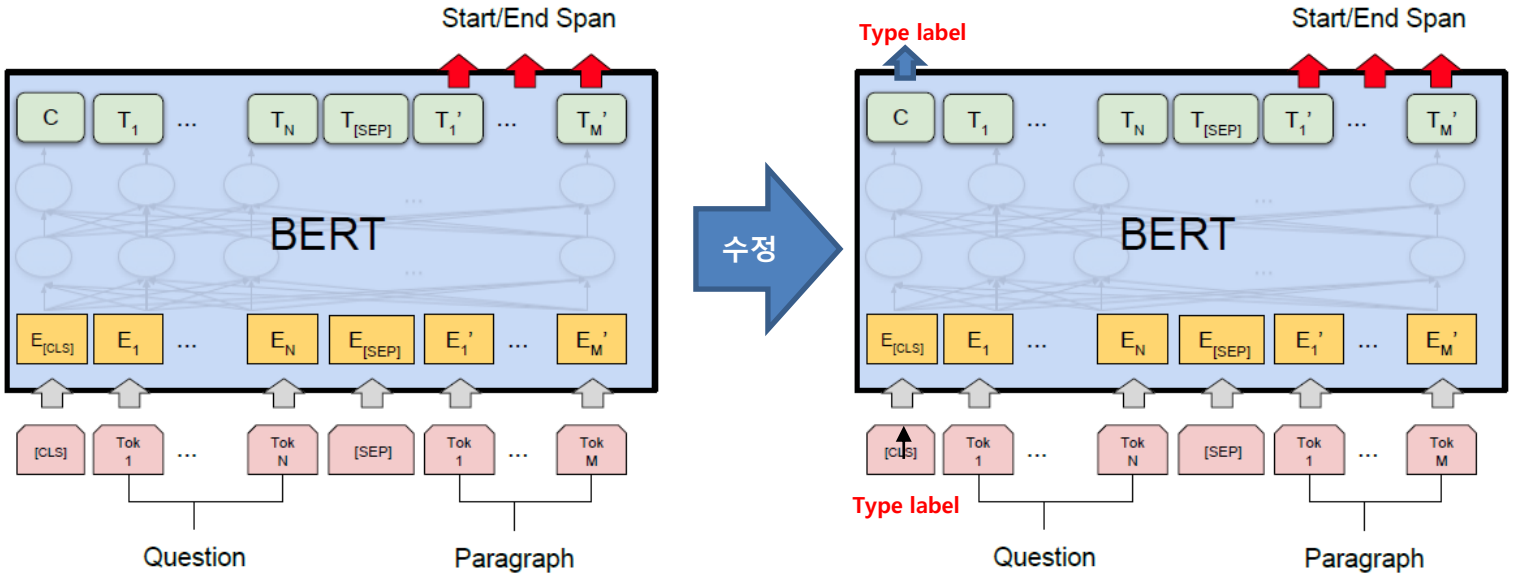
자가학습형 기계독해 구조도



강화학습을 통한 자가학습 순서

1) 기계독해 모델(1) 초기학습
2) 질의 생성 모델 초기학습
3) 비정형 말뭉치 수집
4) 정답 추출 모델 및 질의 생성 모델을 통해 기계학습 데이터 자동 생성
5) 생성된 질의와 수집 문서를 기계독해 모델(1)을 통해 검증 -> F1-score 반환
6) F1-score를 통해 질의 생성 모델 강화학습
7) 기존 학습 데이터를 통해 기계독해 모델(2) 1step 학습
8) 생성된 질의와 수집 문서를 통해 기계독해 모델(2) 1step 학습
9) 기계독해 모델(2) 평가
10) 기계독해 모델(2)의 성능이 모델(1)보다 높을 경우 모델 (2)를 모델 (1)에 복사
11) 수렴까지 4~10 반복

오류 누적을 방지하기 위한 기계독해



실험 데이터

- 영역 1: KorQuAD v1.0
  - 학습 데이터 : 62,411개
  - 평가 데이터 : 5,774개

- 영역 2:. 뉴스, 민원 말뭉치
  - 평가 데이터 : 5,322개

실험 데이터의 차이

- KorQuAD v1.0

Title : 차범근
Context : 대한민국의 축구 선수로 독일 분데스리가에서 활약하며 아시아인으로서는 역대 최다득점이 되는 리그 통산 98골을 기록하였다....
Question : 차범근이 분데스리가에서 기록한 통산 골의 갯수는?
Answer : 98골

- 뉴스, 민원 말뭉치

Title : 보건복지부 민원
Context : 안녕하세요? 보건복지 정책에 관심을 갖고 우리부 홈페이지를 방문해 주셔서 감사합니다. 문의하신 사항에 대하여 답변 드립니다. 초/중등교육법 또는 고등교육법에 따른 각종 학교인 경우에 학교생활 추가급여가 인정됨 ...
Question : 대안학교 등 각종학교에 재학중인 경우도 추가급여 대상인가요?
Answer : 학교생활 추가급여가 인정

실험 결과

- Baseline
  - KorQuAD v1.0 학습 데이터
- QG-Training
  - KorQuAD + 뉴스, 민원 문맥을 통한 생성 질의
- Self-Training (제안모델)
  - 자가학습 적용

Model	KorQuAD v1.0		뉴스, 민원 말뭉치	
	Exact Match (%)	F1-Score (%)	Exact Match (%)	F1-Score (%)
Baseline	76.86	91.09	45.66	68.70
QG-Training	76.95(+0.09)	91.15(+0.06)	58.19(+12.54)	79.96(+11.26)
Self-Training	77.21(+0.35)	91.23(+0.14)	59.17(+13.51)	80.57(+11.87)

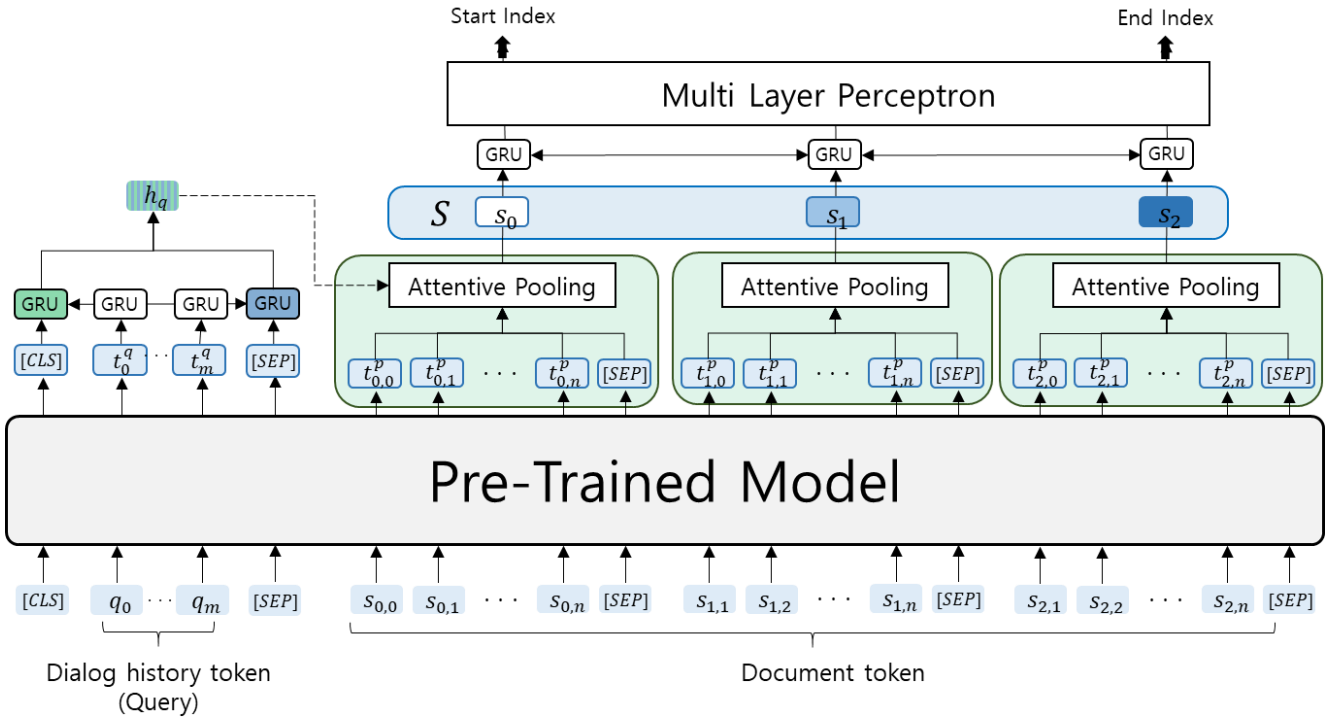
기존 MRC의 문제점

- 주로 간결한 답변을 추출
  - 실제 애플리케이션에서는 문장 형태의 답변이 필요한 경우도 존재
  - 긴 길이의 구간(Long Span)을 효과적으로 추출할 수 있어야 함

- 입력 시퀀스에서 하나의 답변을 추출하는 구조적인 제한이 존재
  - ➔ 문서에서 연속적이지 않은 구간이 필요한 경우도 존재
  - ➔ 다수의 구간(Multi Span)을 추출할 수 있어야 함

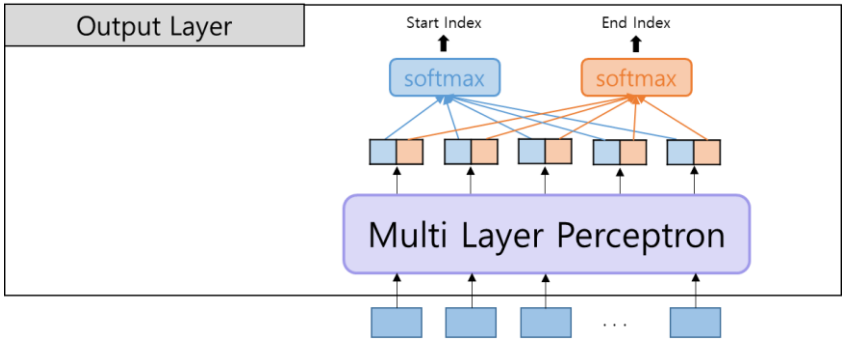
Long Span Prediction

- Long Span Prediction을 위한 모델 구조



Multi-span Prediction

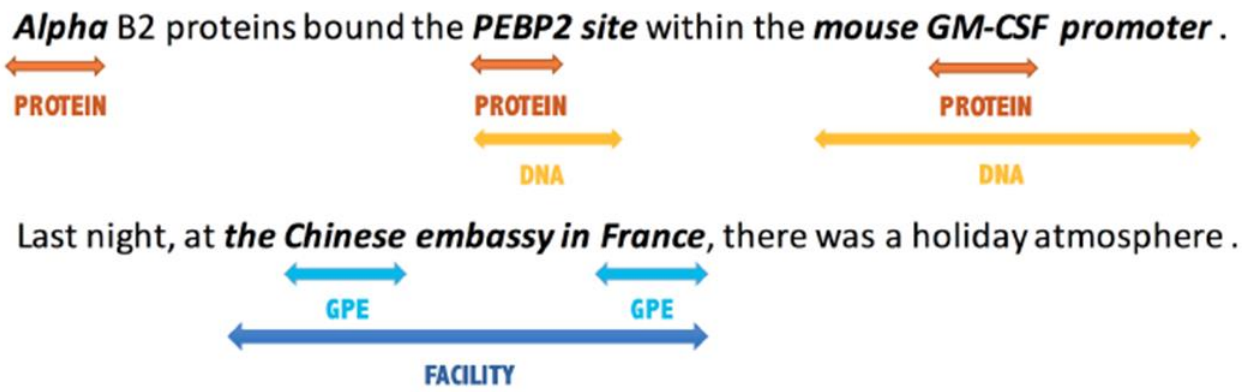
- 연구 방향
  - MRC framework는 대부분 Single Span Prediction에 최적화
  - 답변의 시작 및 끝 위치는 각각 독립적으로 학습됨



  - 기존 Multi Span Prediction 작업에서는 주로 BIO tagging
    - ➔ 하지만 이 방법은 연속적인 구간 추출에 성능 하락을 야기시킴
    - ➔ Span Matrix를 활용하여 Single Span에서 성능이 하락하는 문제를 완화하면서 Multi Span 추출 하고자 함

Span Matrix?

- 중첩 개체명 추출 작업에서 제안된 방법



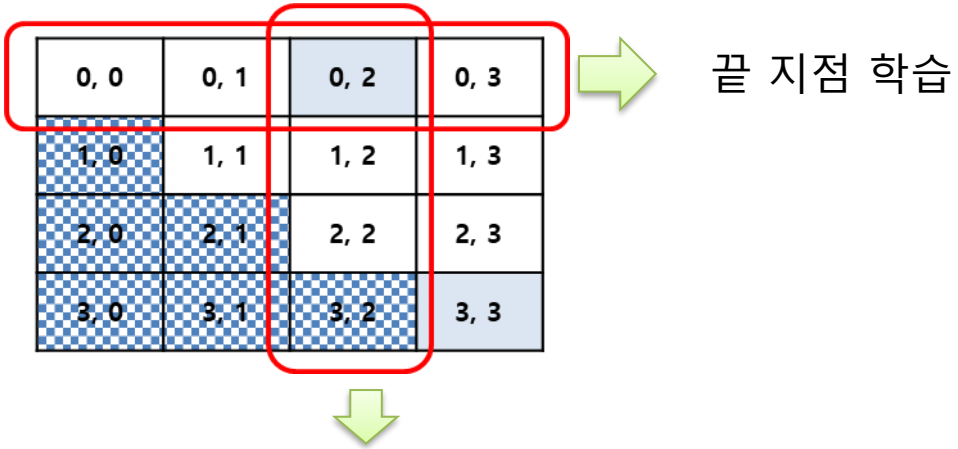
  - 모든 토큰의 조합을 통해 입력 Sequence에서 가능한 Span 생성
  - 생성한 Span 풀에서 유효한 개체명을 추출

➔ 이에 착안하여 Multi Span Prediction에 적용

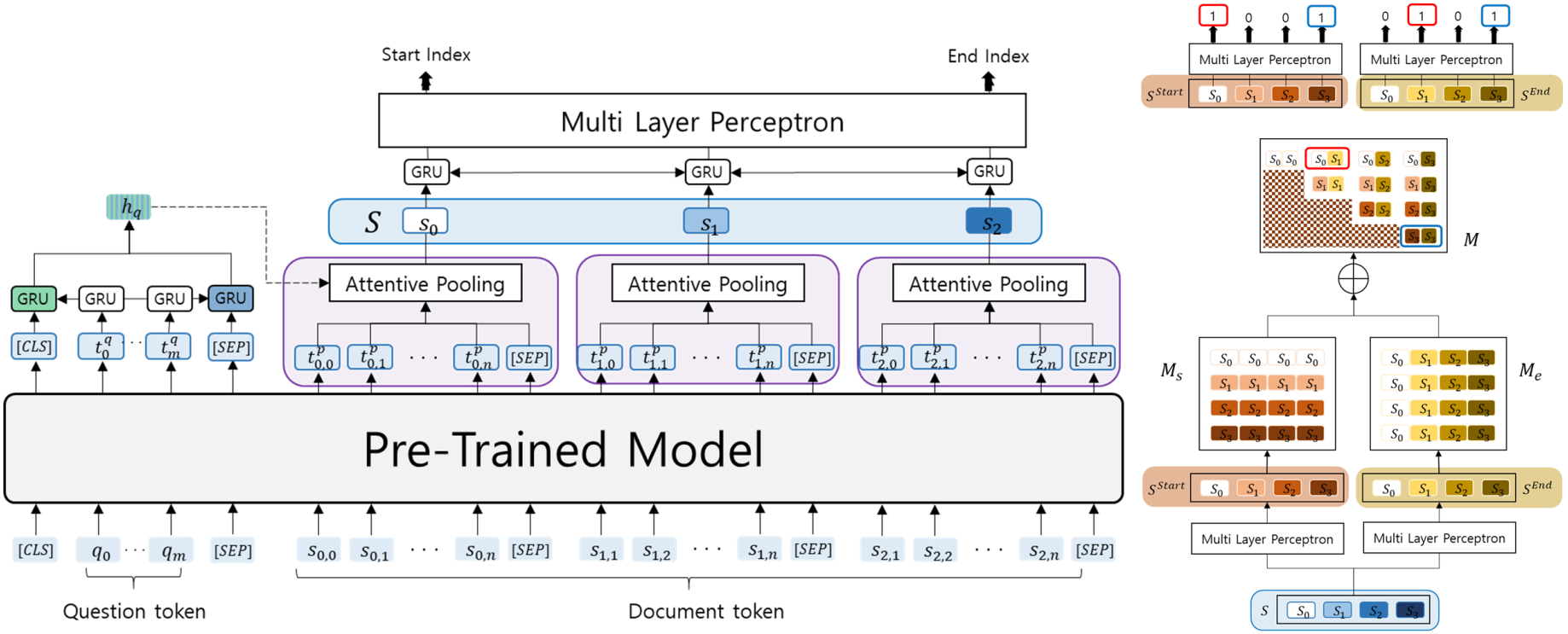
Li et al, "A Unified MRC Framework for Named Entity Recognition", ACL 2020

Span Matrix 적용 방법

- 중첩 개체명에서는 Span Matrix의 각 요소 별로 Binary Classification
  - 하나의 개체 안에 또 다른 개체가 포함되어있는 경우를 추출하기 위해 요소별 이진 분류로 학습
  - 다수의 구간으로 이루어진 서로 다른 지식은 중첩되는 범위가 존재하지 않음 → 유효한 Row/ Column에 대해서만 학습



Multi-span Prediction Model



실험 데이터

- MASH-QA
  - 의료 도메인 QA 데이터
  - 입력 질의에 대해 문장 단위의 답변을 추출 (Long Span)
  - 답변은 하나 이상의 범위로 구성되어 있음 (Single/Multi Span)

- QUOREF
  - 일반 도메인 QA 데이터
  - 입력 질의에 대해 토큰 단위의 답변을 추출 (Short Span)
  - 답변은 하나 이상의 범위로 구성되어 있음 (Single/Multi Span)

실험 결과

- MASH-QA

Model (MASH-QA)	Precision	Recall	F1 score
XLNet	56.05	19.73	29.19
RoBERTa	57.70	19.06	28.65
MultiCo (XLNet)	58.16	55.90	57.00
S-SMASE w/o Span Matrix (RoBERTa)	**65.31**	59.46	62.25
S-SMASE (RoBERTa)	61.37	**64.59**	**62.94**

- QUOREF

Model (QUOREF)	All		Single		Multi	
	EM	F1	EM	F1	EM	F1
RoBERTa	73.3	81.3	81.0	85.9	0.0	37.6
TASEIO+SSE (RoBERTa)	79.0	84.2	80.9	84.6	59.7	80.0
TASEIO (RoBERTa)	79.4	84.9	**81.4**	85.4	59.3	80.0
RoBERTa + Span Matix	**79.5**	**86.8**	**81.4**	**86.9**	**61.9**	**84.2**

Open-Domain Question Answering

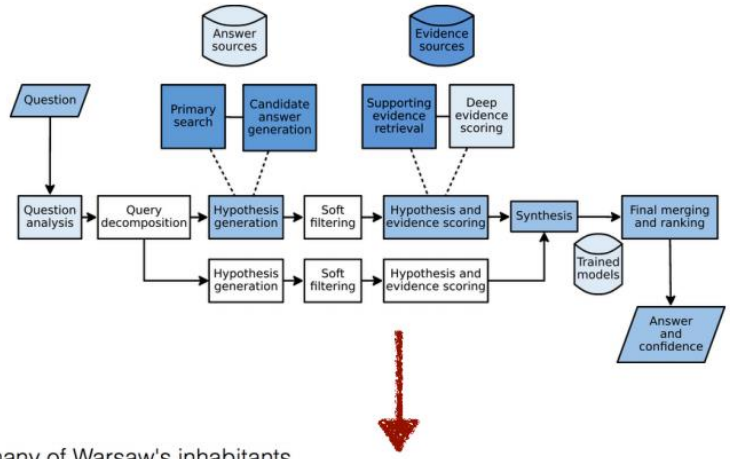
건국대학교 컴퓨터공학부 /
KAIST 전산학부 (겸직)

김학수

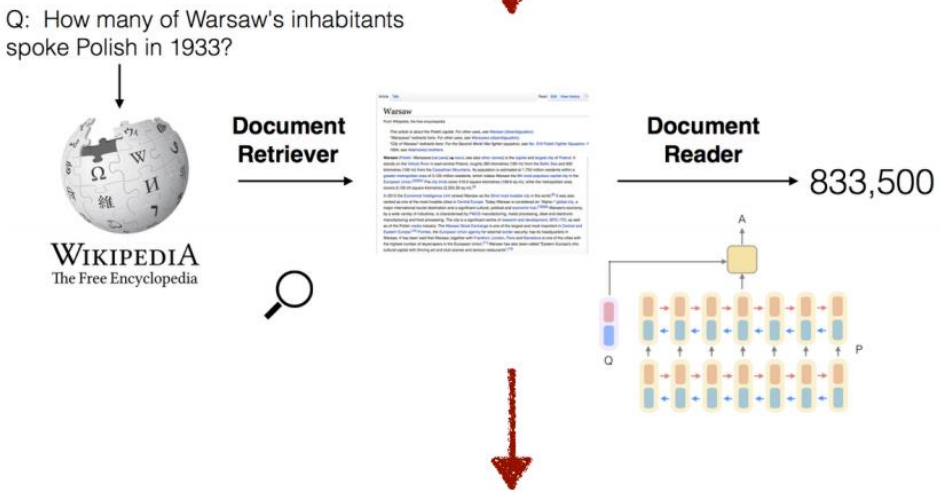
자연어처리연구실

Architectures of Open-Domain QA

Classical QA pipeline

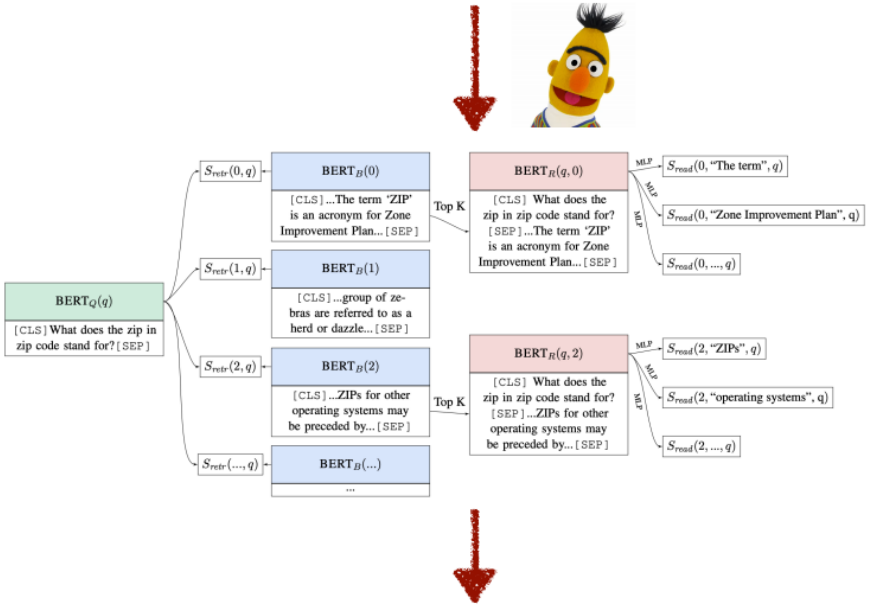


Two-stage Retriever-reader approaches

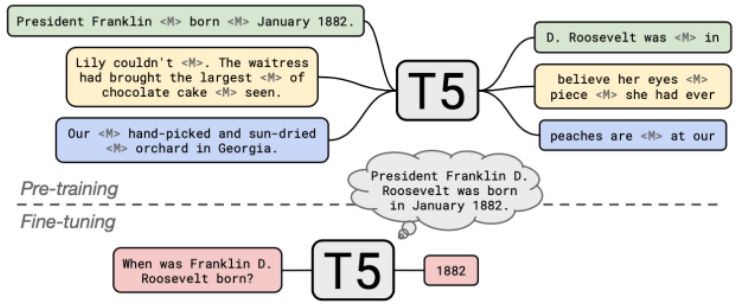


Architectures of Open-Domain QA

End-to-end learning

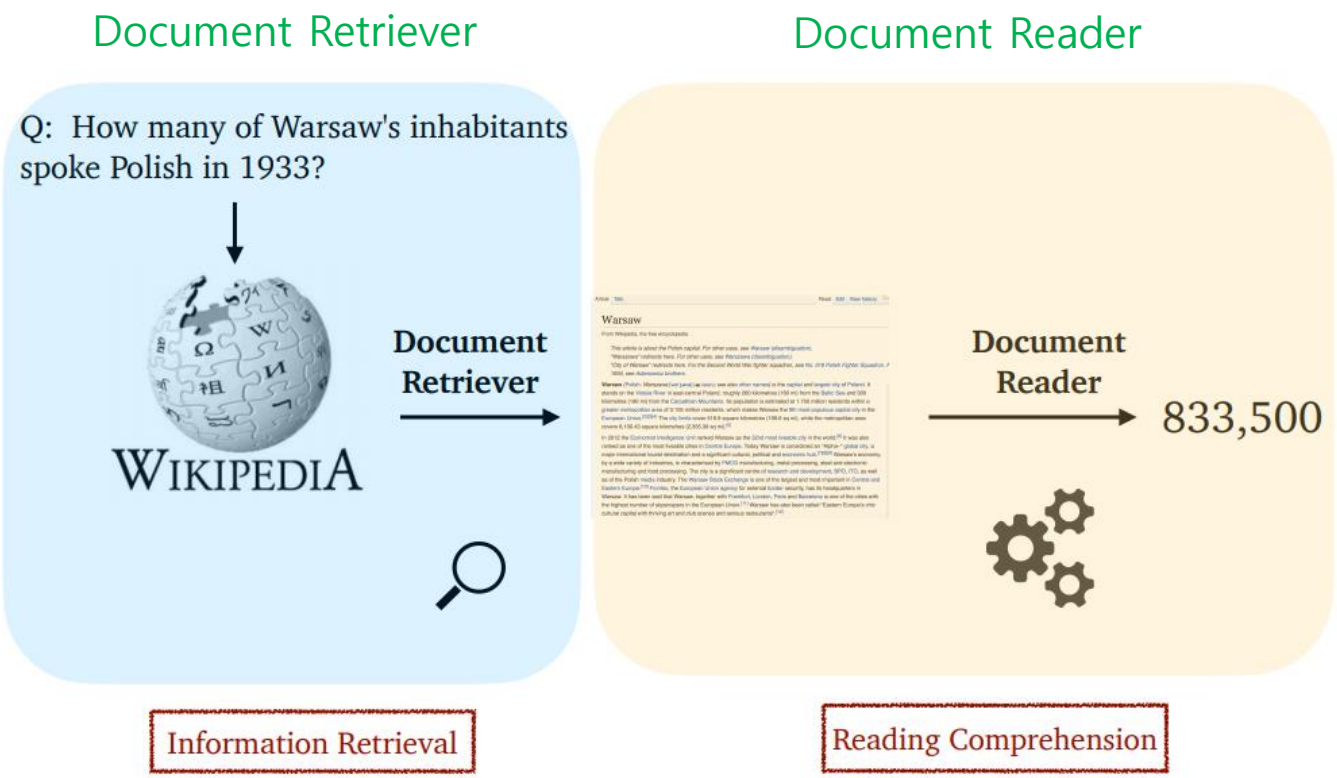


Retrieval-free models



Two-Stage Retriever-Reader

- DrQA: The first neural open-domain QA system



출처: Chen et al., Reading Wikipedia to Answer Open-domain Questions, 2017

DrQA

TF*IDF 랭킹 모델
문서 단위 검색

Training time:

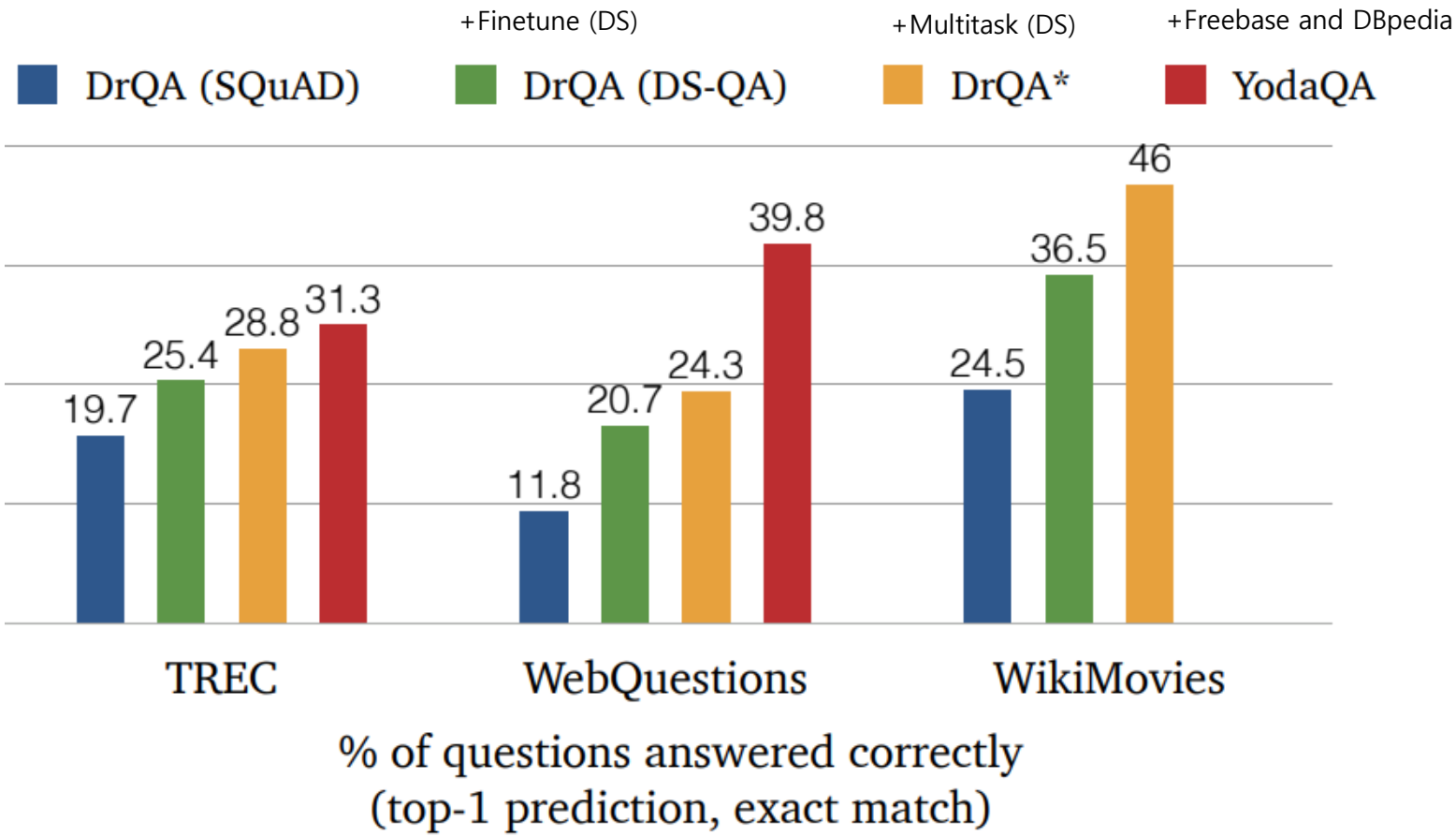
- Document retriever: not trained
- Document reader: a LSTM-based neural reading comprehension model trained on SQuAD + distantly-supervised data generated from QA datasets

Inference time:

- The document retriever returns *top 5 documents*
- The reader reads every (natural) paragraph in these 5 documents and predicts an answer and its span score.
- The system finally returns the answer with the highest (unnormalized) span score.

출처: Chen et al., Reading Wikipedia to Answer Open-domain Questions, 2017

DrQA



출처: Chen et al., Reading Wikipedia to Answer Open-domain Questions, 2017

Problems of DrQA

- DrQA considers retrieval at document level.

  Does paragraph-level retriever work better?

- Answers in the retrieved passages might not be directly comparable at inference time.

  Does multi-passage training help?

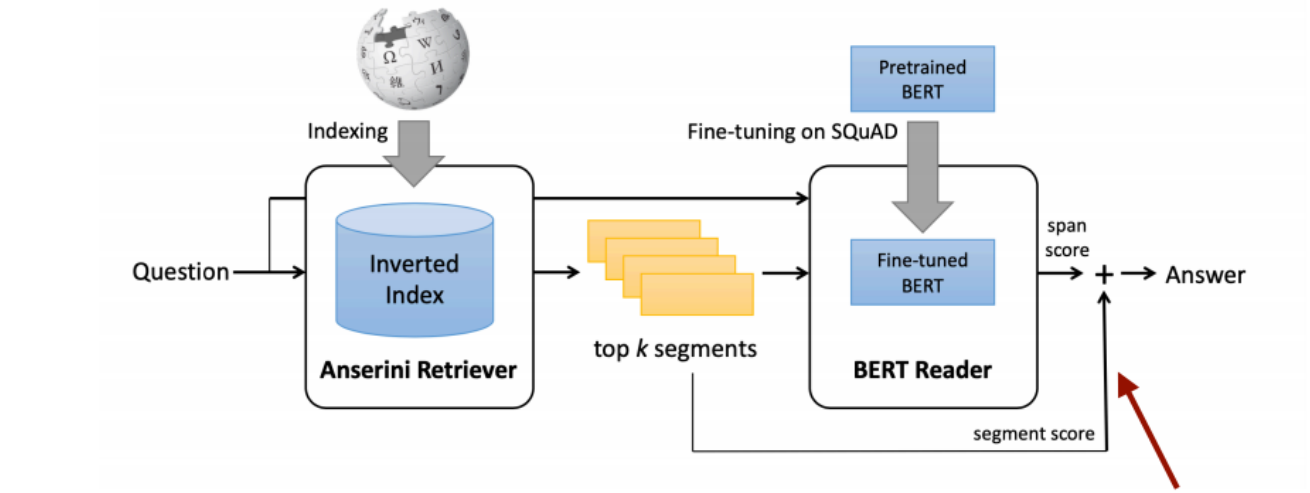
- The importance of each passage has been omitted.

  Can we use passage retriever scores or even train a better ranker?

- The retriever is not trained!

출처: Open-Domain QA Tutorial in ACL 2020

Paragraph-level Retrieving



Anserini Retriever
[Yang et al. 2017]:
Lucene with BM25, operated on 29.5M paragraphs

BERT Reader:
Trained on SQuAD

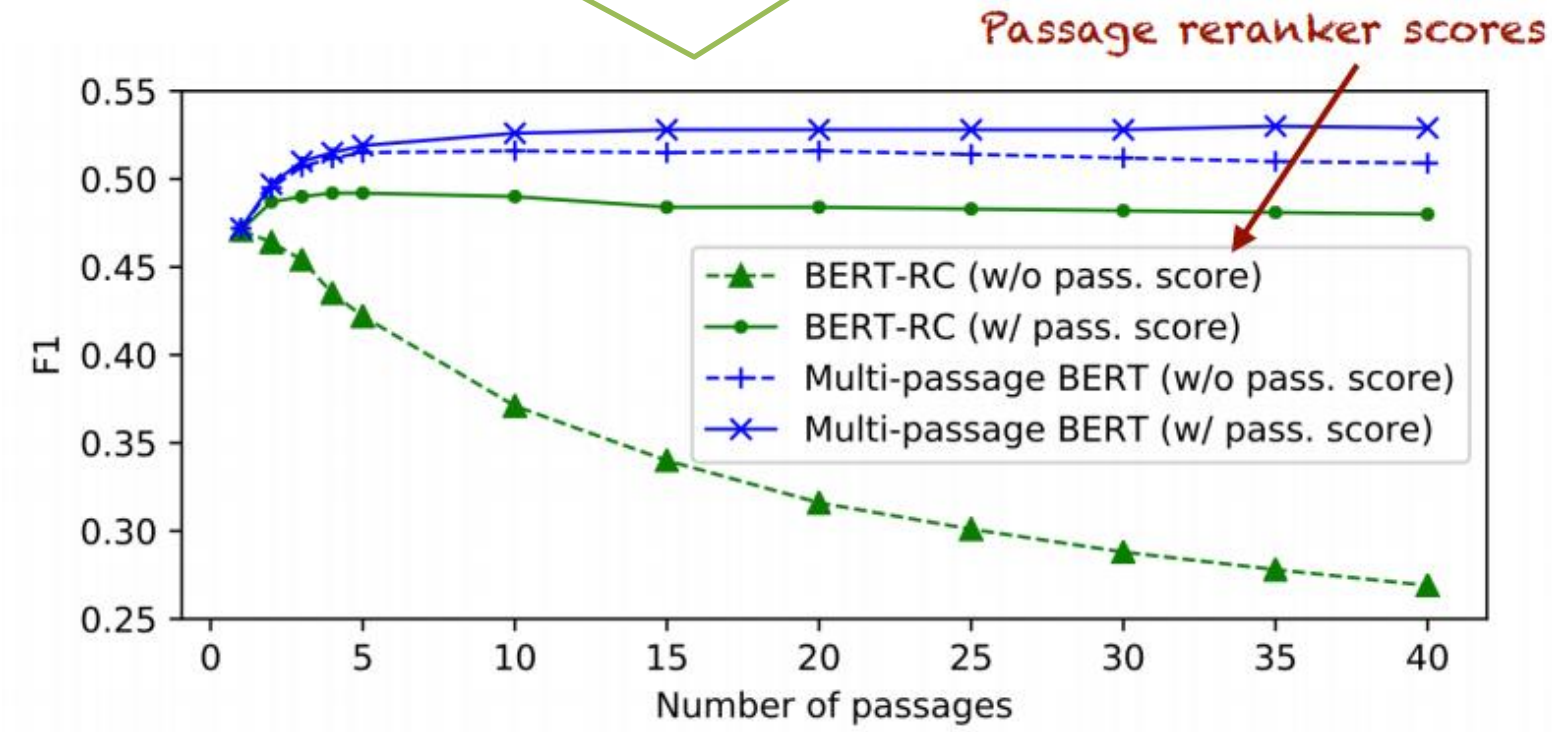
Scoring from both retriever and reader:
$S = (1 - \mu) \cdot S_{\text{Anserini}} + \mu \cdot S_{\text{BERT}}$

This system is only evaluated on SQuAD:
27.1 (DrQA, SQuAD only) $\longrightarrow$ **38.6**

출처: Yang et al., End-to-End Open-Domain Question Answering with BERTserini, 2019

Multi-passage Training

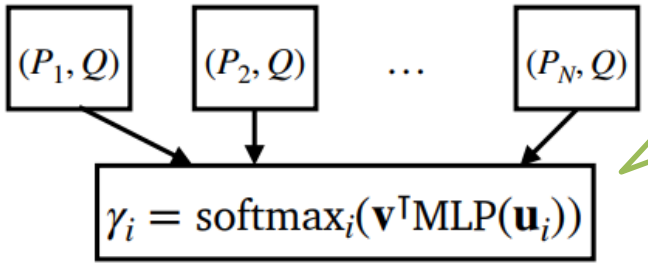
Shared normalization: process passages independently, but compute the span probability across spans in all passages in every mini-batch



출처: Wang et al., Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering, 2019

Training a Passage Re-Ranker

- Training a "deep" re-ranker model on retrieved passages can help further identify the relevance of the passages.



BERT의 [CLS] 토큰의 점수들을 global normalize 후 softmax!

$\mathbf{u}_i$ is a representation of (P_i, Q)

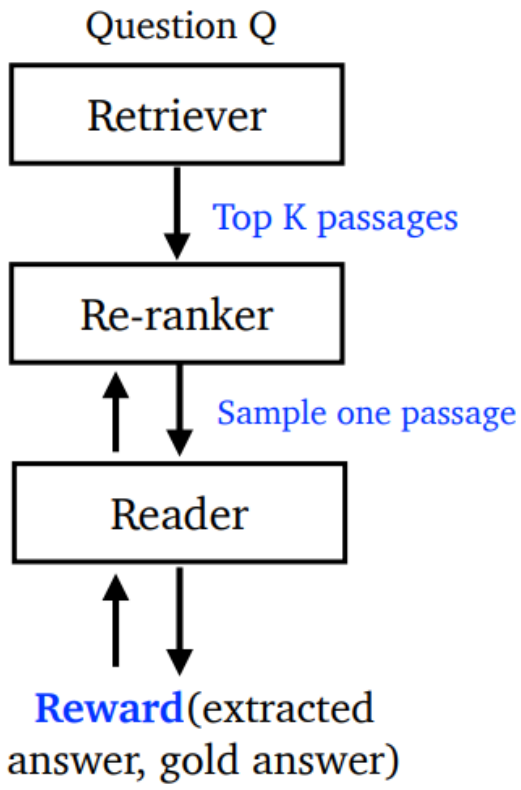
- This reranker can be easily trained using **distant supervision**: whether the passage contains the answer or not.

Reinforced Ranker-Reader

Algorithm 1 Reinforced Ranker-Reader (R^3)

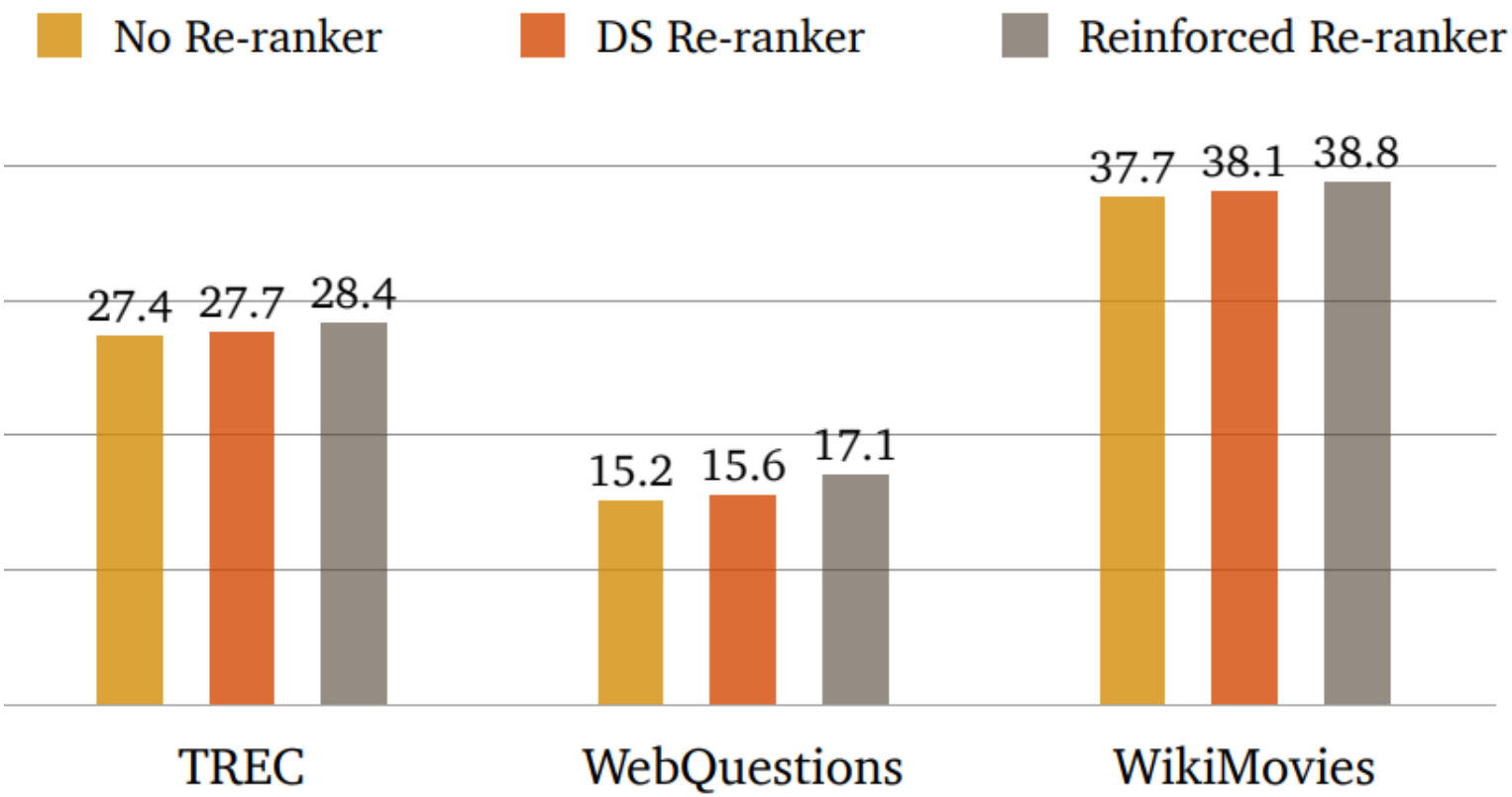
Input: $\mathbf{a}^g$, $\mathbf{q}$, passages from IR
Output: Θ
Initialize: $\Theta \leftarrow$ pre-trained Θ with a baseline method[6]
for each $\mathbf{q}$ in dataset **do**
  For question $\mathbf{q}$, sample K passages from the top N passages retrieved by IR model for training. [7]
  Randomly sample a positive passage $\tau \sim \pi(\tau|\mathbf{q})$
  Extract the answer $\mathbf{a}^{rc}$ through RC model
  Get reward r according to $R(\mathbf{a}^g, \mathbf{a}^{rc}|\tau)$.
  Updating Ranker (ranking model) through policy gradient $r\frac{\partial}{\partial\Theta}\log(\pi(\tau|\mathbf{q}))$
  Updating Reader (RC model) through supervised gradient $\frac{\partial}{\partial\Theta}L(\mathbf{a}^g|\tau,\mathbf{q})$
end for

$$\begin{cases} 2, & \text{if } \mathbf{a}^g == \mathbf{a}^{rc} \\ f1(\mathbf{a}^g, \mathbf{a}^{rc}), & \text{else if } \mathbf{a}^g \cap \mathbf{a}^{rc} != \emptyset \\ -1, & \text{else} \end{cases}$$

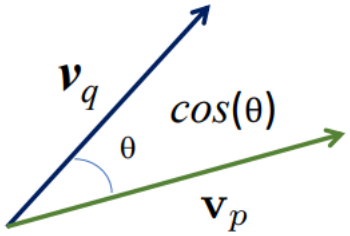


출처: Wang et al., R³: Reinforced Ranker-Reader for Open-Domain Question Answering, 2018

Experiments of R^3



Sparse Retrieval to Dense Retrieval



$d_1 >> d_2$

sparse repr: [0… 1 …. 1 …0..1] $\in \mathbb{R}^{d_1}$

dense repr: [1.03, -5.72, 6.42, .., 9.91] $\in \mathbb{R}^{d_2}$

sparse

"How many provinces did the Ottoman empire contain in the 17th century?"

"What part of the atom did Chadwick discover?"

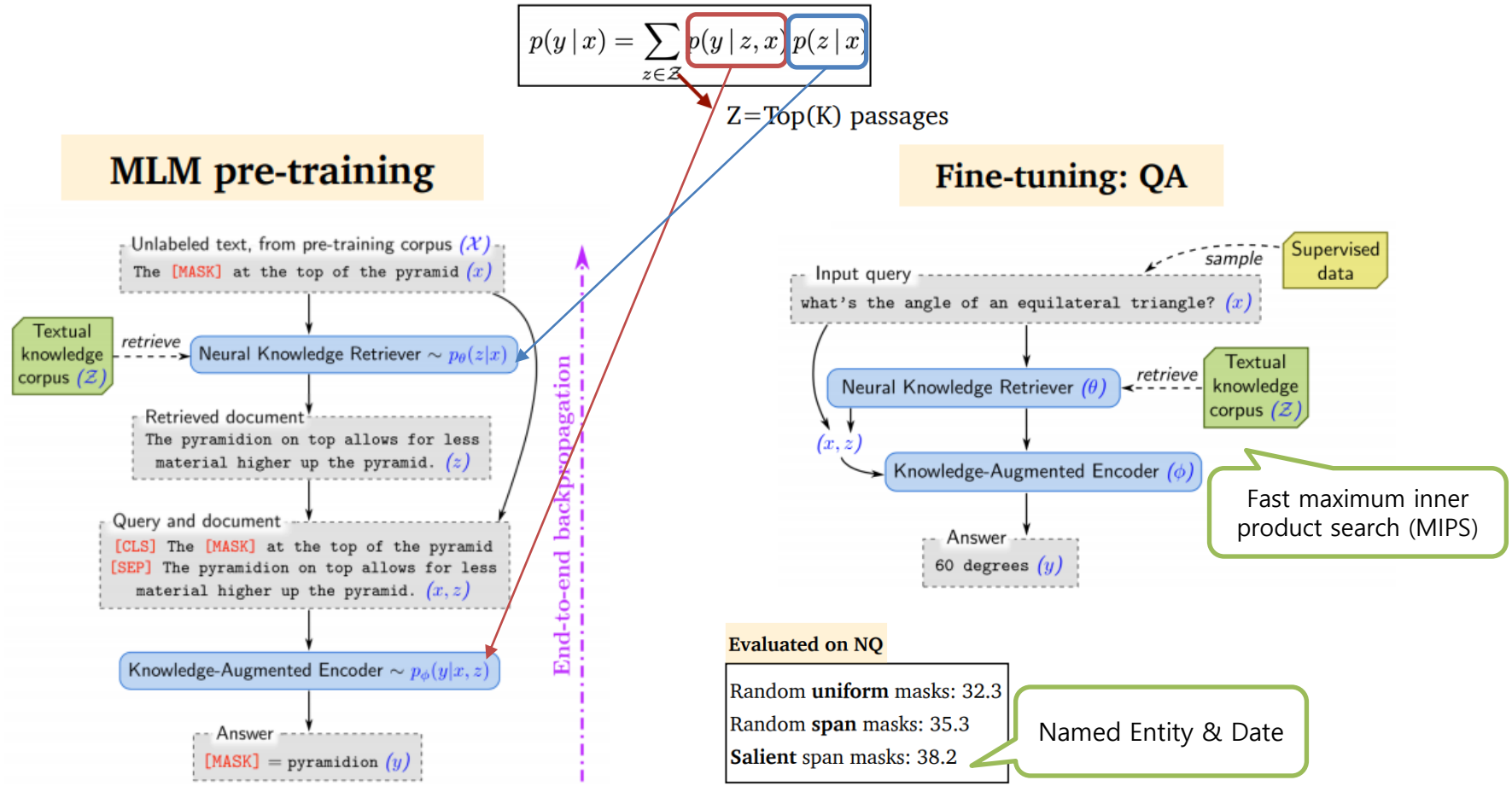
dense

"Who is the **bad guy** in lord of the rings?"

*Sala Baker is an actor and stuntman from New Zealand. He is best known for portraying the **villain** Sauron in the Lord of the Rings trilogy by Peter Jackson.*

출처: Open-Domain QA Tutorial in ACL 2020

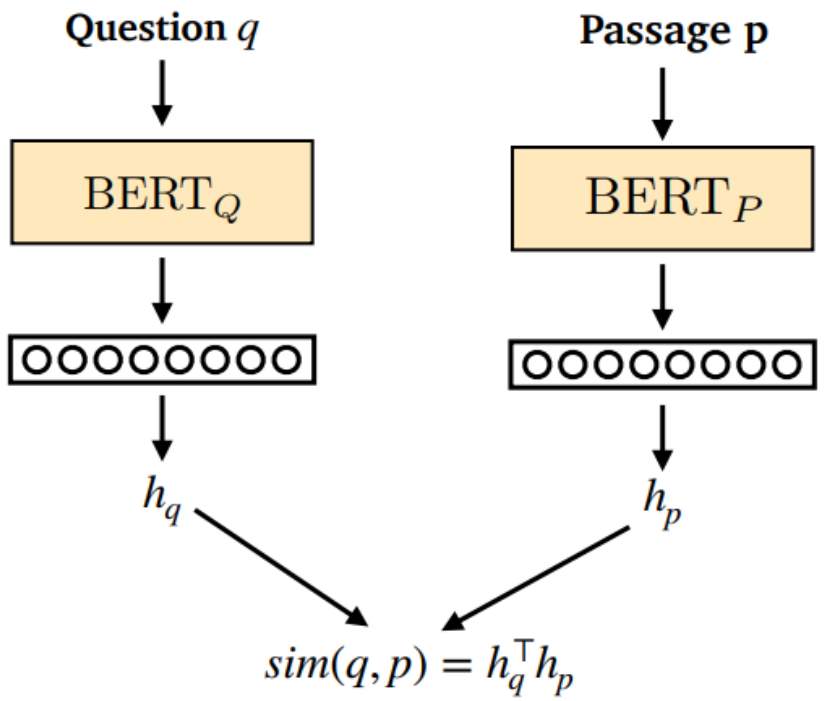
REALM: Retrieval-Augmented Language Model



출처: Guu et al., REALM: Retrieval-Augmented Language Model Pre-Training, 2020

DPR: Dense Passage Retrieval

Key contribution: you can train a dense retrieval only from a small number of Q/A pairs, without any pre-training!



핵심 아이디어!
Contrastive Learning

How to get positives and negatives?

$$\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^- \rangle\}_{i=1}^m$$

$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^n e^{\text{sim}(q_i, p_{i,j}^-)}}$$

후보 벡터는 FAISS(facebook에서 만든 유사도 측정 알고리즘)를 이용하여 고속으로 검색

출처: Karpukhin et al., Dense Passage Retrieval for Open-Domain Question Answering, 2020
Open-Domain QA Tutorial in ACL 2020

How to Train DPR

Positives

(1) Provided in the reading comprehension datasets

(2) Passages of high BM25 scores that contain the answer string

Negatives

(1) Random passages from the corpus

(2) Passages of high BM25 scores that DO NOT contain the answer string

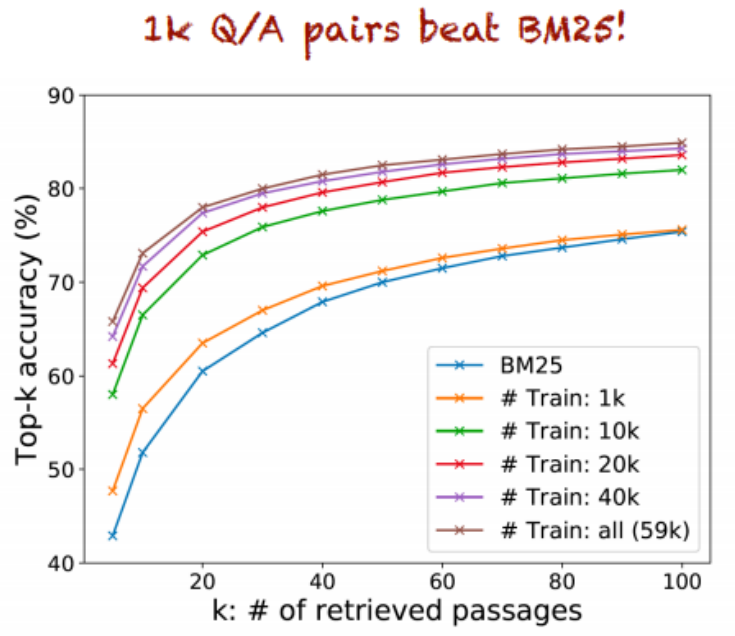
(3) Positive passages of OTHER questions

The best model uses (3) from the same mini-batch [in-batch negatives] and one passages from (2) [hard negatives].

출처: Karpukhin et al., Dense Passage Retrieval for Open-Domain Question Answering, 2020
Open-Domain QA Tutorial in ACL 2020

Experiments of DPR

Retriever performance on NQ



End-to-end QA performance

With a *multi-passage BERT reader* trained on the retrieved passages from the retriever:

- DPR is better than BM25 on NaturalQuestions, WebQuestions, TREC, TriviaQA but not SQuAD.
- DPR is better than REALM on bigger datasets (41.5 vs 39.2 on NQ). For smaller datasets (WebQ, TREC), it needs a mixed training with bigger datasets to outperform REALM.

Why? 인위적인 데이터 → 키워드 매칭 가능성 높음

출처: Karpukhin et al., Dense Passage Retrieval for Open-Domain Question Answering, 2020
Open-Domain QA Tutorial in ACL 2020

Retrieval-Free

- Key question: can we use **pre-trained language models** to act as "knowledge storage"?

Pertoni et al. Language Models as Knowlesge Bases?, 2019

- Instead of explicitly storing all the text and searching among their *dense* or *sparse* representations, can we query the LMs to obtain the answer directly?

- The LMs were pre-trained on Wikipedia (and other textual corpora) so they should be able to memorize a fair amount of information.

LMs as KBs?

Barack Obama was born in Honolulu.

출처: Open-Domain QA Tutorial in ACL 2020

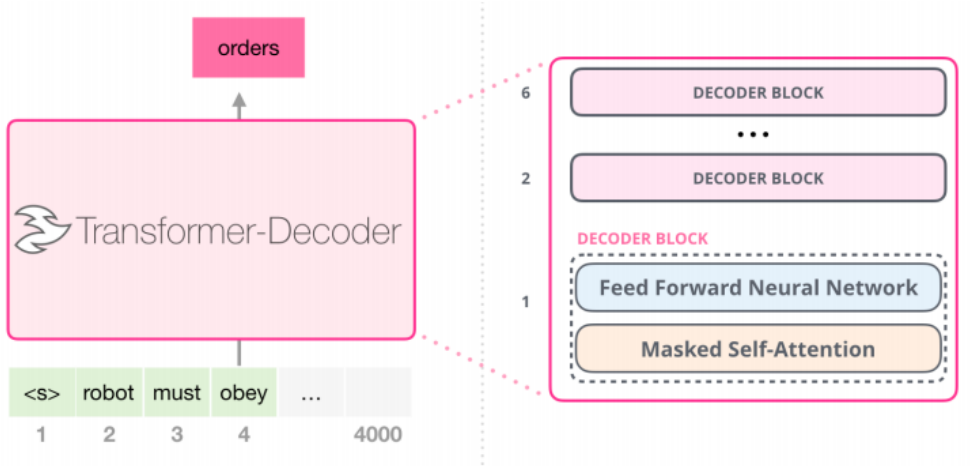
GPT-2

- GPT-2 is a *very large*, transformer-based language model trained on *a massive dataset*.

48 layers, hidden size 1600, 1.5B parameters

WebText: 8 million documents, excluding Wikipedia (!)

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, \ldots, s_{n-1})$$



출처: Rardford et al., Language Models are Unsupervised Multitask Learners, 2019
Open-Domain QA Tutorial in ACL 2020

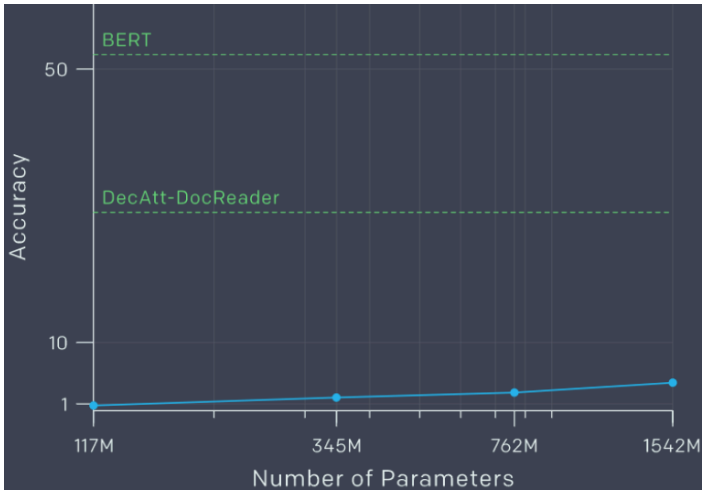
GPT-2: Zero-Shot QA

- Evaluated on Natural Questions and no training at all

63.1% on the 1% of questions

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

4% accuracy but no fine-tuning



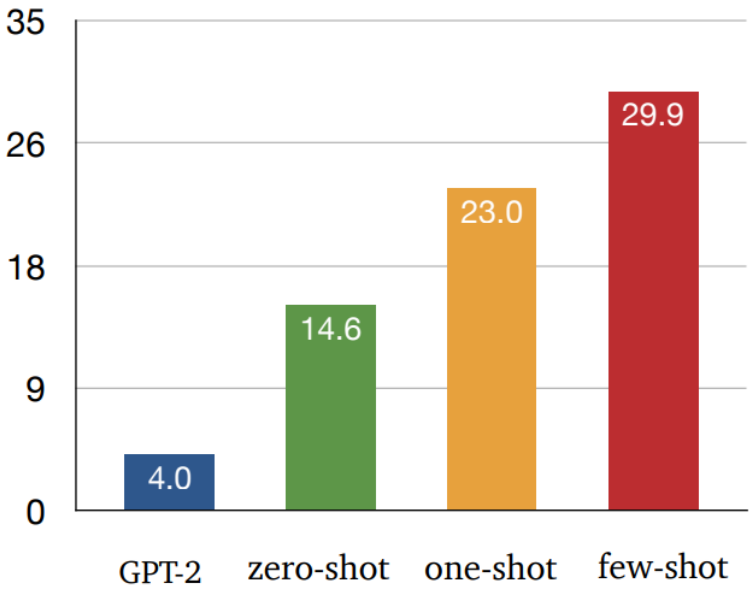
출처: Rardford et al., Language Models are Unsupervised Multitask Learners, 2019
Open-Domain QA Tutorial in ACL 2020

GPT-3: Few-Shot Learner

96 layers, hidden size 12288, **175B** parameters

Larger corpora: Common Crawl + WebText + Books + **English Wikipedia**

Evaluated on Natural Questions:



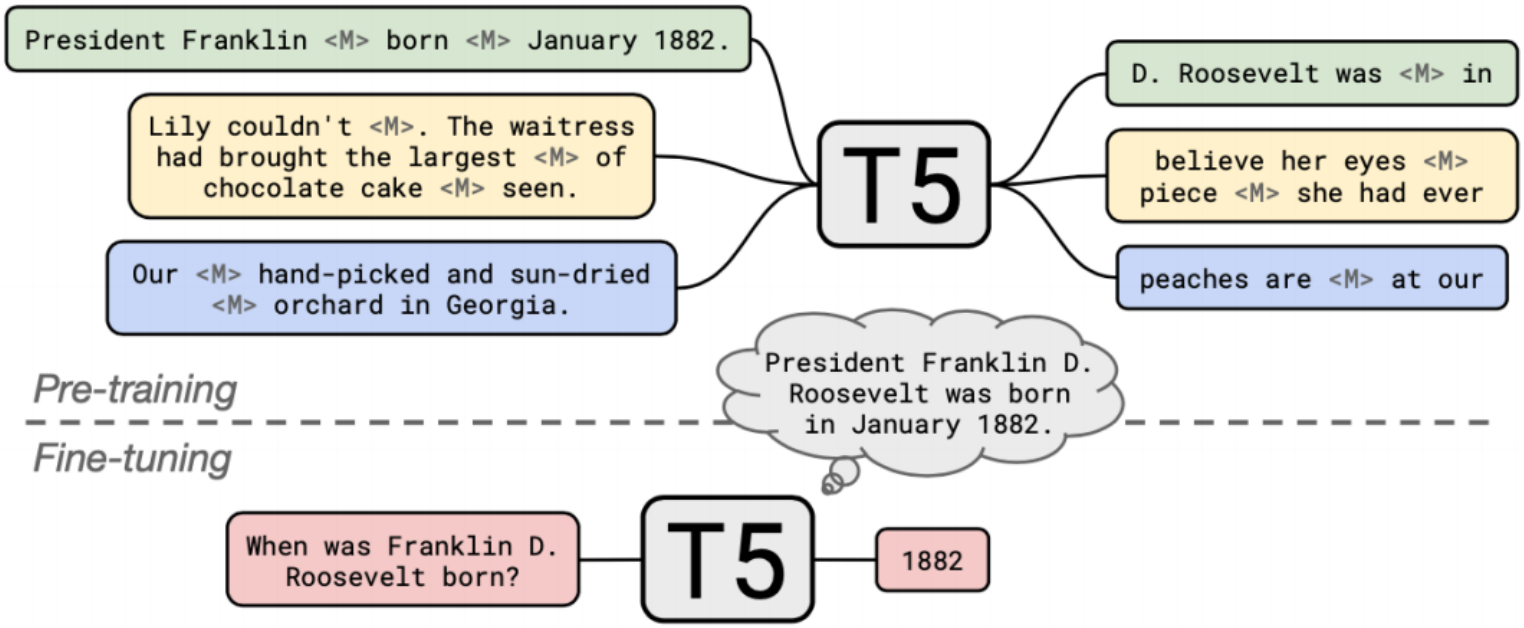
Few-shot learner

- No weight updates
- $Q_1, A_1, Q_2, A_2, .., Q_K, A_K, Q$?
- One-shot setting is a special case when only **one** example is given.

Argument (Prompt) Learning!

출처: Brown et al., Language Models are Few-Shot Learners, 2020
Open-Domain QA Tutorial in ACL 2020

T5: Fine-tuning leads to improved performances



*: Pre-trained on a multitask mixture including an **unsupervised "span corruption" task** on unlabeled text as well as supervised translation, summarization, classification, and reading comprehension tasks

출처: Roberts et al., How Much Knowledge Can You Pack into the Parameters of Language Models?, 2020
Open-Domain QA Tutorial in ACL 2020

Experiments of Fine-Tuned T5

		Natural Questions	WebQuestions	TriviaQA
		NQ	WQ	TQA
	Chen et al. (2017)	–	20.7	–
	Lee et al. (2019)	33.3	36.4	47.1
	Min et al. (2019a)	28.1	–	50.9
	Min et al. (2019b)	31.8	31.6	55.4
	Asai et al. (2019)	32.6	–	–
	Ling et al. (2020)	–	–	35.7
	Guu et al. (2020)	40.4	40.7	–
	Févry et al. (2020)	–	–	53.4
	Karpukhin et al. (2020)	**41.5**	42.4	57.9
220M	T5-Base	27.0	29.1	29.1
770M	T5-Large	29.8	32.2	35.9
3B	T5-3B	32.1	34.9	43.4
11B	T5-11B	34.5	37.4	50.1
	T5-11B + SSM	36.6	**44.7**	**60.5**

Named Entity & Date

BERT-base = 110M parameters

출처: Roberts et al., How Much Knowledge Can You Pack into the Parameters of Language Models?, 2020
Open-Domain QA Tutorial in ACL 2020

질의응답

Homepage: http://nlp.konkuk.ac.kr
E-mail: nlpdrkim@konkuk.ac.kr