KOREA UNIVERSITY

DSBA CS224n 2021 Study

[Lecture 11]
Question Answering

DSBA Data Science & Business Analytics

고려대학교 산업경영공학과

Data Science & Business Analytics Lab

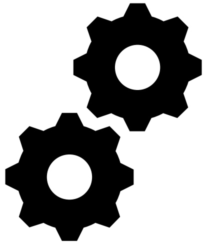
발표자 : 김선우

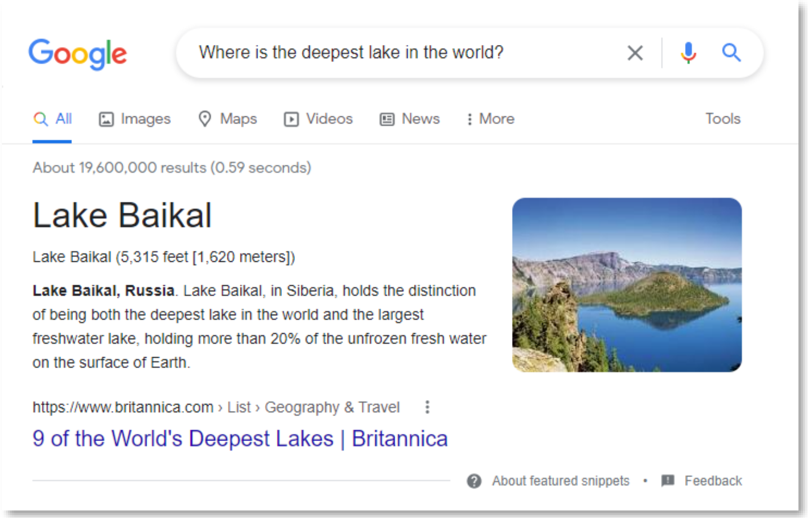
00 발표 목차

Contents

01 Question Answering

What is question answering?

Question(Q) ⟶  ⟶ Answer(A)

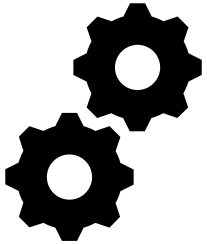


사람의 언어로 된 **질문**에 자동적으로 **답**할 수 있는 시스템을 만들자!

01 Question Answering

Question answering : a taxonomy

Question(Q) $\longrightarrow$



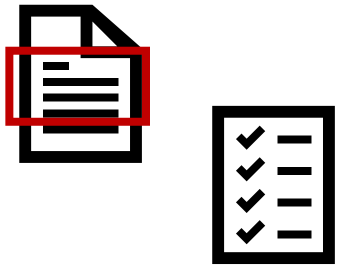
$\longrightarrow$ Answer(A)

정보 소스

질문 유형

- ✓ Factoid vs non-factoid
- ✓ Open-domain vs closed domain
- ✓ Compositional vs simple

답 유형



Y/N

01 Question Answering

Question answering in deep learning era

대부분의 state-of-the-art question answering 시스템들은 end-to-end train과 pre-train된 language model 위해 build

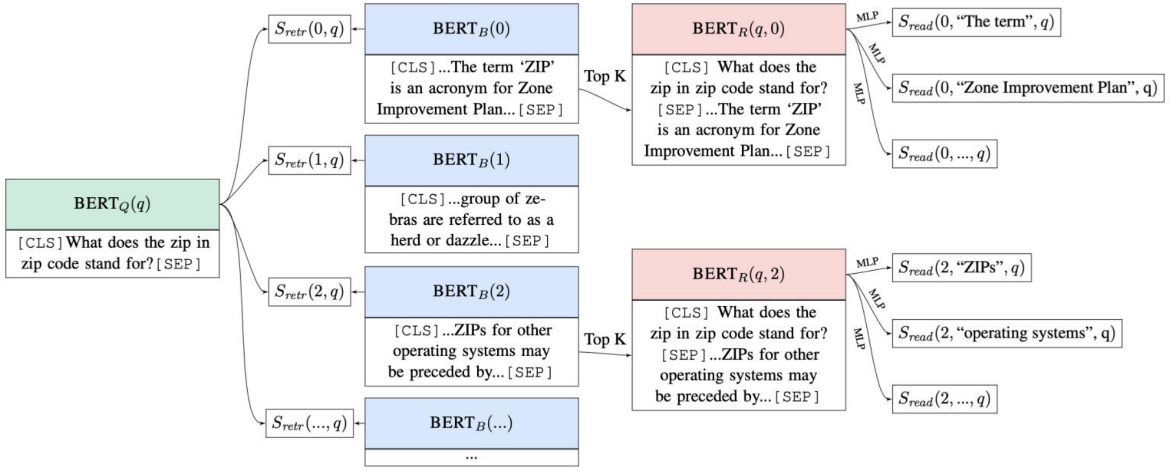


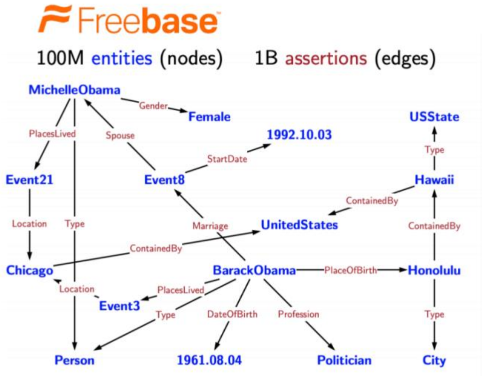
Image credit: (Lee et al., 2019)

01 Question Answering

Beyond textual QA problems

오늘날에는 unstructured text에 기반한 질문에 답하고자 한다

Knowledge based QA



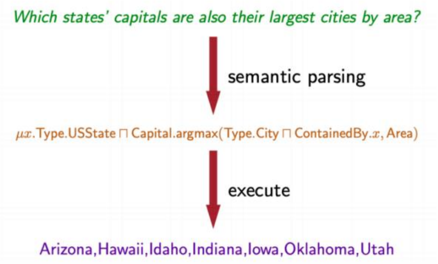
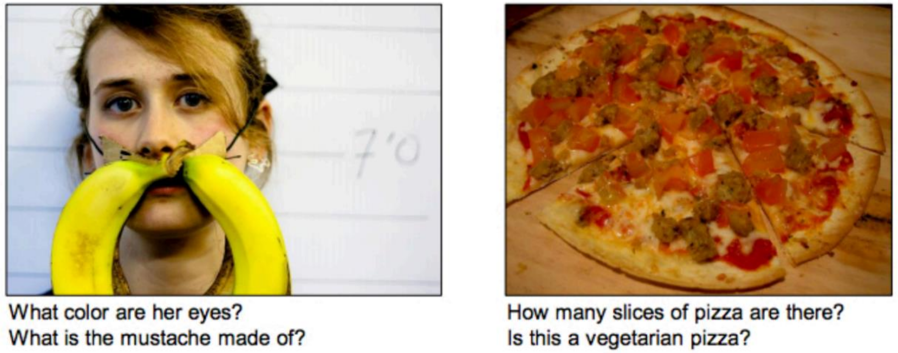


Image credit: Percy Liang

Visual QA



(Antol et al., 2015): Visual Question Answering

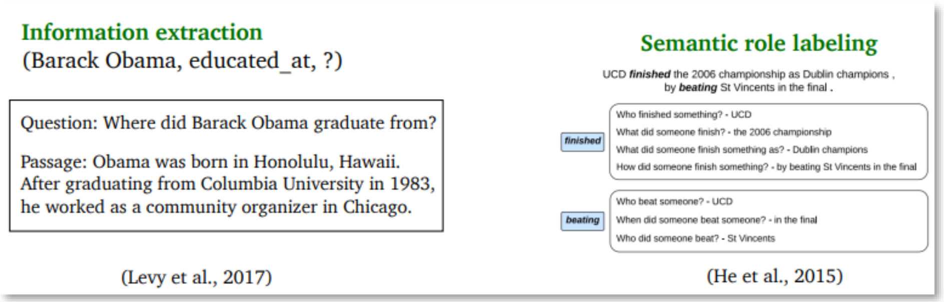
02 Reading Comprehension

What is reading comprehension?

Kannada language is the official language of Karnataka and spoken as a native language by about 66.54% of the people as of 2011. Other linguistic minorities in the state were Urdu (10.83%), Telugu language (5.84%), Tamil language (3.45%), Marathi language (3.38%), Hindi (3.3%), Tulu language (2.61%), Konkani language (1.29%), Malayalam (1.27%) and Kodava Takk (0.18%). In 2007 the state had a birth rate of 2.2%, a death rate of 0.7%, an infant mortality rate of 5.5% and a maternal mortality rate of 0.2%. The total fertility rate was 2.2.

Q: Which linguistic minority is larger, Hindi or Malayalam?

A: Hindi



$$(P, Q) \rightarrow A$$

Text로 이루어진 문단을 **이해**하고
해당 내용에 대한 질문에 **답**하자!

- ✓ 얼마나 컴퓨터가 사람의 언어를 잘 이해하는지 평가할 수 있는 testbed
- ✓ 많은 다른 NLP task도 reading comprehension 문제로 단순화 가능

Reading Comprehension

Stanford question answering dataset (SQuAD)

(passage, question, answer) triple 100k개

한계: <u>각 answer passage의 span</u>

Stanford Question Answering Dataset (SQuAD)

Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

Along with non-governmental and nonstate schools, what is another name for private schools?

Gold answers: ① independent ② independent schools ③ independent schools

Estimated human performance
EM = 82.3, F1 = 91.2

Evaluation

✓ Exact Match (EM)

3개의 answer에 대해 span이 존재하는지에 따라 0/1accuracy 값 획득

✓ F1

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$$

$$\mathrm{F1} = \frac{2 * precision * recall}{precision * recall}$$

	Actual	
Predicted	True positive(TP)	False Positive(FP)
	False negative(FN)	True Negative(TN)

02 Reading Comprehension

Stanford question answering dataset (SQuAD)

Q: What did Tesla do in December 1878?

A: {left Graz, left Graz, left Graz and severed all relations with his family}
Prediction: {left Graz and served}

Exact match: max{0, 0, 0} = 0
F1: max{0.67, 0.67, 0.61} = 0.67

1.예측된 answer을 각각의 gold answer와 비교
(a, an, the, .(구두점) 제거)
2. Max score 계산
3. 모든 example에 대해 EM과 F1 평균

02 Reading Comprehension

How can we build a model to solve SQuAD?

(passage, paragraph, context), (question query) 같은 의미로 사용

input

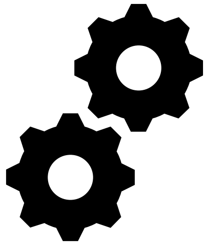
Context

Query

$$C = (c_1, c_2, \ldots, c_N)$$
$$Q = (q_1, q_2, \ldots, q_M)$$
$$c_i, q_i \in V$$

N : 100~200, M: 10~15

Problem formulation

→  →

output

Answer

Context의 일부분(span)

$1 \leq$ start $\leq$ end $\leq N$

A family of LSTM-based models with attention (2016-2018)

Fine-tuning BERT-like models for reading comprehension(2019+)

02 Reading Comprehension

LSTM-based vs BERT models

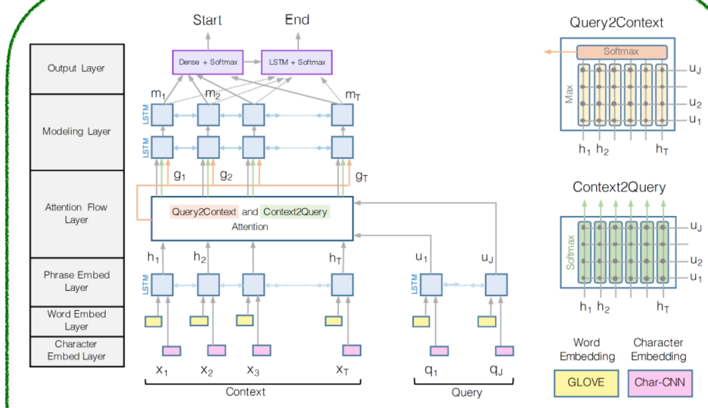


Image credit: (Seo et al, 2017)

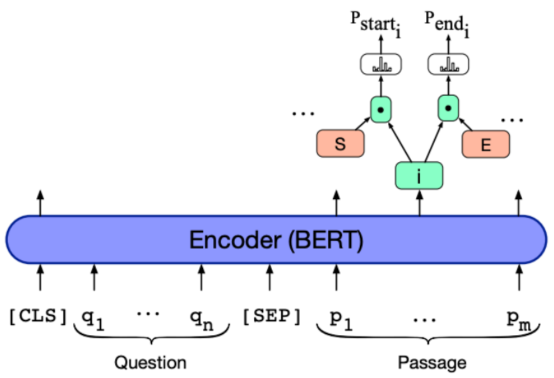


Image credit: J & M, edition 3

03 Reading Comprehension

Recap : Seq2seq model with attention

Machine Translation

- ✓ Source, target 문장
- ✓ Autoregressive decoder

(word-by-word target 문장 생성)

- ✓ Source 문장의 어떤 단어가 현재의 target 단어와 가장 관련 있을까?

Reading Comprehension

- ✓ Passage, question

(길이는 다를 수 있다)

- ✓ Two classifier

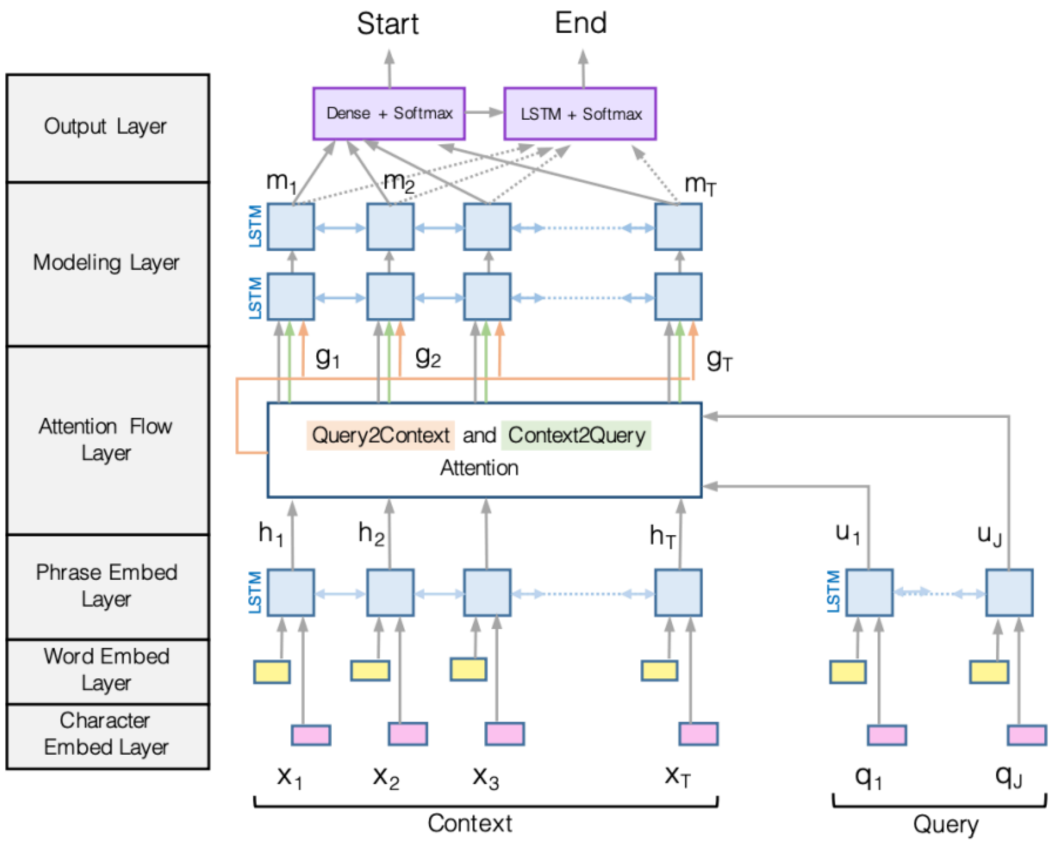
(정답의 start, end 위치만 예측)

- ✓ Passage의 어떤 단어들이 question(의 어떤 단어)와 가장 관련 있을까?

Attention is the key ingredient

03 Reading Comprehension

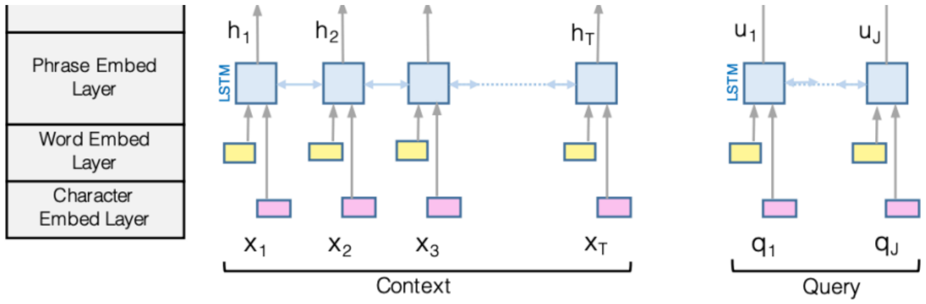
The Bidirectional Attention Flow model(BiDAF)

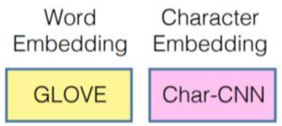


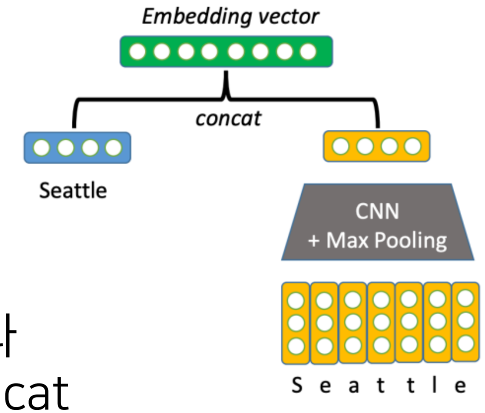
(Seo et al., 2017): Bidirectional Attention Flow for Machine Comprehension

02 Reading Comprehension

BiDAF : Encoding







✓ context와 query의 각 단어에 대해 word embedding (GloVe)과 character embedding(CNNs over character embeddings) concat

$e(c_i) = f([GloVe(c_i); charEmb(c_i)])$ $e(q_i) = f([GloVe(q_i); charEmb(q_i)])$

f: high-way networks omitted here

✓ contextual embedding 생성 위해 context와 query에 대해 각각 bidirectional LSTM 사용

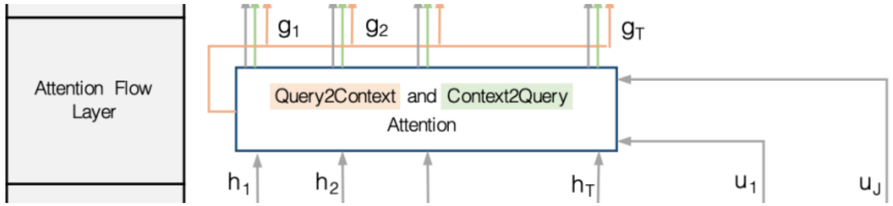
$$\vec{c}_i = LSTM(\vec{c}_{i-1}, e(c_i)) \in \mathbb{R}^H \qquad \vec{q}_i = LSTM(\vec{q}_{i-1}, e(q_i)) \in \mathbb{R}^H$$

$$\overleftarrow{c}_i = LSTM(\overleftarrow{c}_{i+1}, e(c_i)) \in \mathbb{R}^H \qquad \overleftarrow{q}_i = LSTM(\overleftarrow{q}_{i+1}, e(q_i)) \in \mathbb{R}^H$$

$$c_i = [\vec{c}_i, \overleftarrow{c}_i] \in \mathbb{R}^{2H} \qquad q_i = [\vec{q}_i, \overleftarrow{q}_i] \in \mathbb{R}^{2H}$$

02 Reading Comprehension

BiDAF:Attention



Query-to-context attention

query 단어 중 하나와 가장 관련 있는 context 단어들 선택하기

While Seattle's weather is very nice in summer, its weather is very rainy in winter, making it one of the most gloomy cities in the U.S. LA is …

Q: Which city is gloomy in winter?

Context-to-query attention

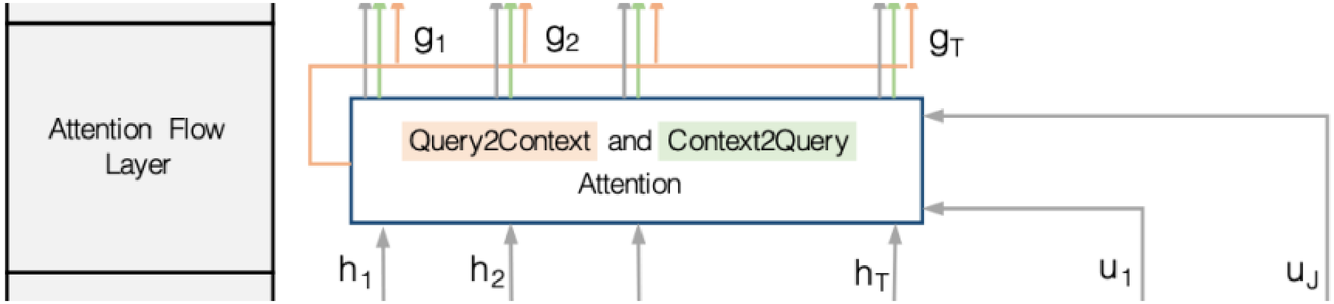
각각의 context 단어에 대해 가장 유사한 query 단어 찾기

Q: Who leads the United States?

C: Barack Obama is the president of USA.

02 Reading Comprehension

BiDAF:Attention



Input : query와 context의 contextual vector representation (c_i, q_j)
Output : context 단어들의 query-aware vector representation(g_i),
이전 layer의 contextual embedding

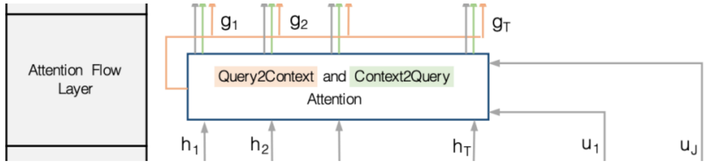
1. 모든 (c_i, q_j) 쌍에 대해 유사도 (similarity score)계산

$$S_{i,j} = w_{sim}^T [c_i; q_j; c_i \odot q_j] \in \mathbb{R} \quad w_{sim} \in \mathbb{R}^{6H}$$

2. Context-to-query attention (질문의 어떤 단어들이 c_i와 관련있는지), Query-to-context attention (문단에서 어떤 단어들이 질문 단어들과 관련있는지) 연산

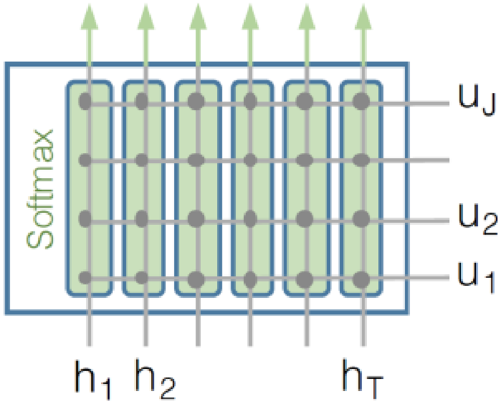
02 Reading Comprehension

BiDAF:Attention



Context-to-query attention

각 contex단어에 어떤 query 단어들이 더 관련있는지

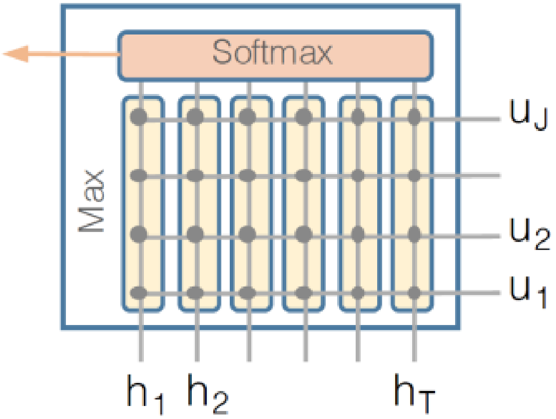


$$\alpha_{i,j} = softmax_j(S_{i,j}) \in \mathbb{R}$$

$$a_i = \sum_{j=1}^{M} \alpha_{i,j} q_j \in \mathbb{R}^{2H}$$

Query-to-context attention

하나의 query 단어에 대해 어떤 context 단어들이 가장 유사도 높은지



$$\beta_{i,j} = softmax_i(max_{j=1}^{M}(S_{i,j})) \in \mathbb{R}^N$$

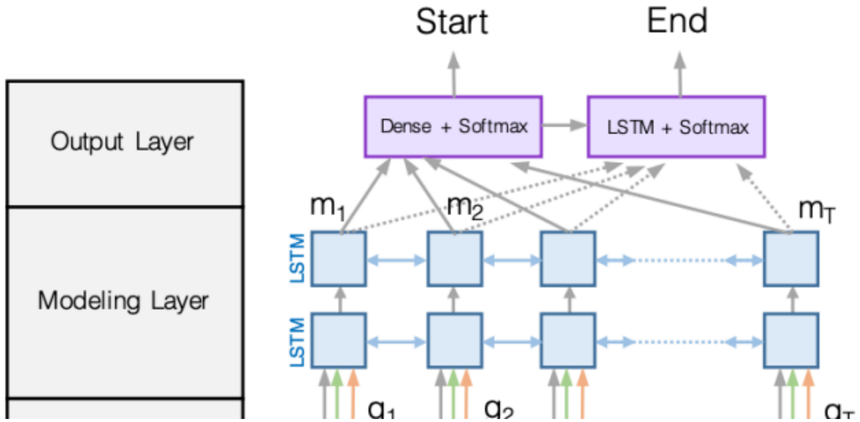
$$b_i = \sum_{i=1}^{M} \beta_i c_i \in \mathbb{R}^{2H}$$

Final output

$$g_i = [c_i; a_i; c_i \odot a_i; c_i \odot b_i] \in \mathbb{R}^{8H}$$

02 Reading Comprehension

BiDAF: Modeling and output layers



Final training loss

$L = -log\ p_{start}(s^*) - log\ p_{end}(s^*)$

✓ **Modeling layer** : g_i를 또다른 bidirectional LSTM의 2개의 layer로 전달

-Attention layer : query와 context 사이의 interaction 모델링

-Modeling layer : context 단어들 사이의 interaction 모델링

$$m_i = BiLSTM(m_i) \in \mathbb{R}^{2H}$$

✓ **Output layer** : start, end 위치를 예측하는 classifier

$p_{start} = softmax(w_{start}^T[g_i; m_i])$ $\quad p_{end} = softmax\left(w_{end}^T[g_i; m'_i]\right)$

$m'_i = BiLSTM(m_i) \in \mathbb{R}^{2H}$ $\quad w_{start}, w_{end} \in \mathbb{R}^{10H}$

02

Reading Comprehension

BiDAF:Performance on SQuAD

	EM	F1
No char embedding	65.0	75.4
No word embedding	55.5	66.8
No C2Q attention	57.2	67.7
No Q2C attention	63.6	73.7
Dynamic attention	63.5	73.6
BIDAF (single)	67.7	77.3
BIDAF (ensemble)	72.6	80.7

	Published[12]	LeaderBoard[13]
Single Model	EM / F1	EM / F1
LR Baseline (Rajpurkar et al., 2016)	40.4 / 51.0	40.4 / 51.0
Dynamic Chunk Reader (Yu et al., 2016)	62.5 / 71.0	62.5 / 71.0
Match-LSTM with Ans-Ptr (Wang & Jiang, 2016)	64.7 / 73.7	64.7 / 73.7
Multi-Perspective Matching (Wang et al., 2016)	65.5 / 75.1	70.4 / 78.8
Dynamic Coattention Networks (Xiong et al., 2016)	66.2 / 75.9	66.2 / 75.9
FastQA (Weissenborn et al., 2017)	68.4 / 77.1	68.4 / 77.1
BiDAF (Seo et al., 2016)	68.0 / 77.3	68.0 / 77.3
SEDT (Liu et al., 2017a)	68.1 / 77.5	68.5 / 78.0
RaSoR (Lee et al., 2016)	70.8 / 78.7	69.6 / 77.7
FastQAExt (Weissenborn et al., 2017)	70.8 / 78.9	70.8 / 78.9
ReasoNet (Shen et al., 2017b)	69.1 / 78.9	70.6 / 79.4
Document Reader (Chen et al., 2017)	70.0 / 79.0	70.7 / 79.4
Ruminating Reader (Gong & Bowman, 2017)	70.6 / 79.5	70.6 / 79.5
jNet (Zhang et al., 2017)	70.6 / 79.8	70.6 / 79.8
Conductor-net	N/A	72.6 / 81.4
Interactive AoA Reader (Cui et al., 2017)	N/A	73.6 / 81.9
Reg-RaSoR	N/A	75.8 / 83.3
DCN+	N/A	74.9 / 82.8
AIR-FusionNet	N/A	76.0 / 83.9
R-Net (Wang et al., 2017)	72.3 / 80.7	76.5 /84.3
BiDAF + Self Attention + ELMo	N/A	**77.9/ 85.3**
Reinforced Mnemonic Reader (Hu et al., 2017)	73.2 / 81.8	73.2 / 81.8

02 Reading Comprehension

Attention visualization

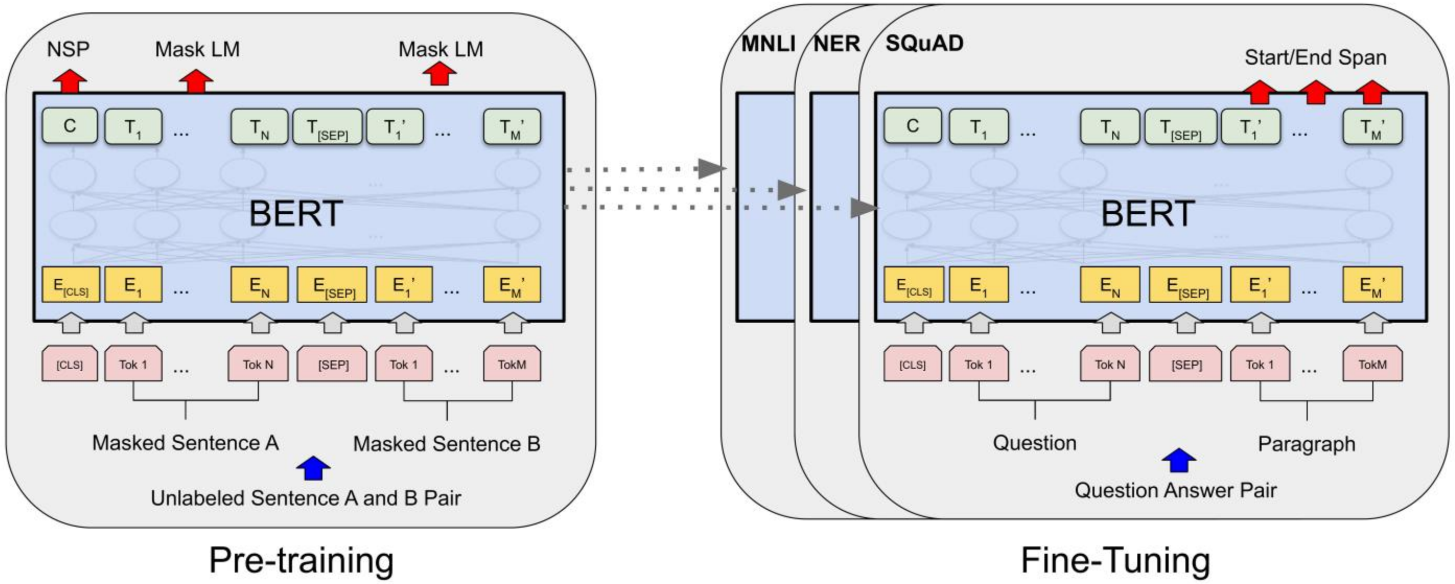
Attention score

02 Reading Comprehension

BERT for reading comprehension

BERT란? 대량의 text(Wikipedia + BooksCorpus) 에 pre-train된 deep bidirectional Transformer encoder

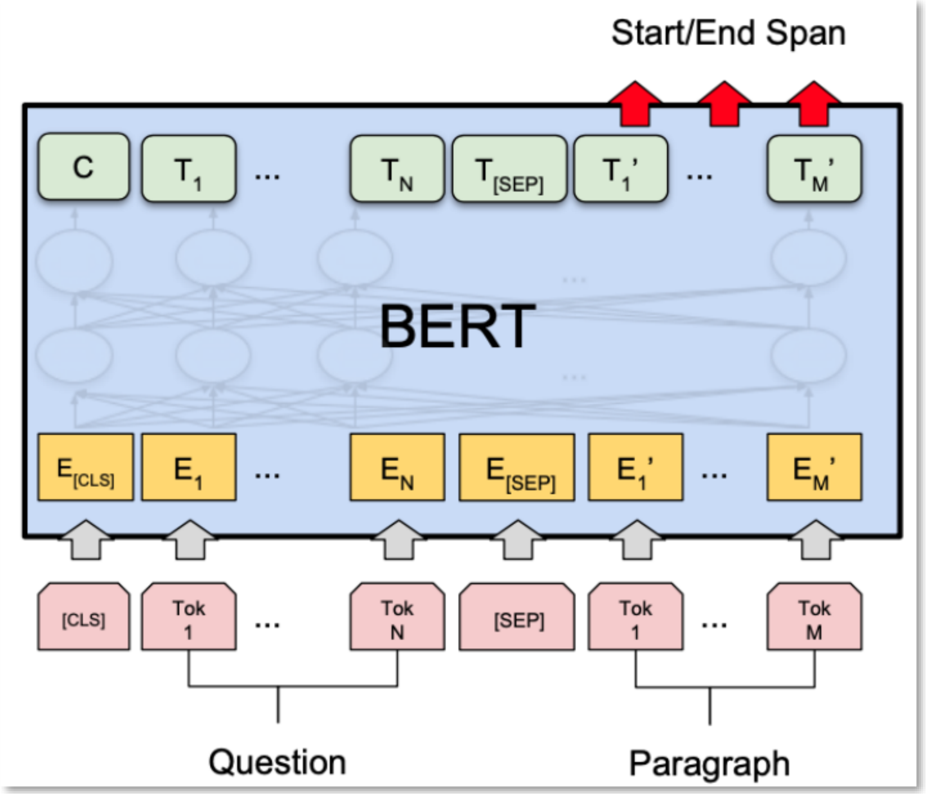
pre-train 1. Masked language model (MLM) 2. Next sentence prediction (NSP)



$BERT_{base}$: 12개의 layer, 110M parameters
$BERT_{large}$: 24개의 layer, 330M parameters

02 Reading Comprehension

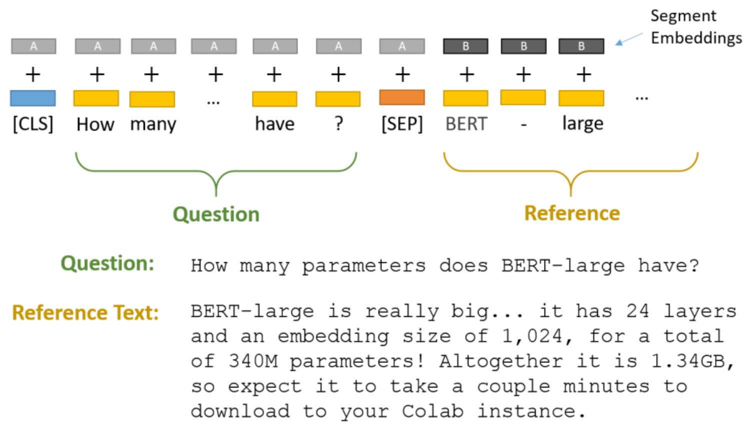
BERT for reading comprehension



Question = Segment A

Passage = Segment B

Answer = predicting two endpoints in segment B



Final training loss

$$L = -\log p_{start}(s^*) - \log p_{end}(e^*)$$

$$p_{start}(i) = softmax_i(w_{start}^T H)$$

$$p_{end}(i) = softmax_i(w_{end}^T H)$$

Where H = $[h_1, h_2, \cdots, h_N]$ 는 BERT에 의해 반환되는 paragraph의 hidden vector

02 Reading Comprehension

Comparisons between BiDAF and BERT models

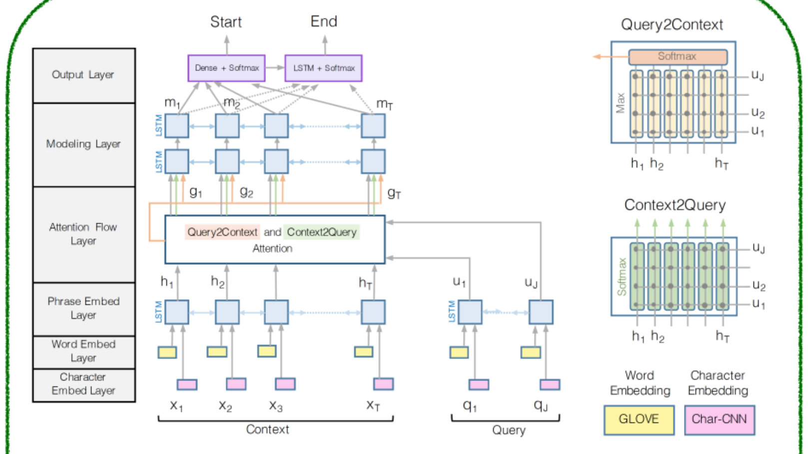


Image credit: (Seo et al, 2017)

BiDAF

- ✓ ~2.5M parameters
- ✓ 다수의 bidirectional LSTM 위에 build
- ✓ Only built on top of GloVe

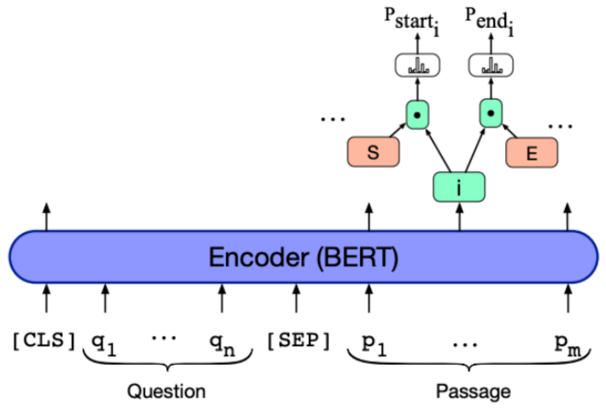


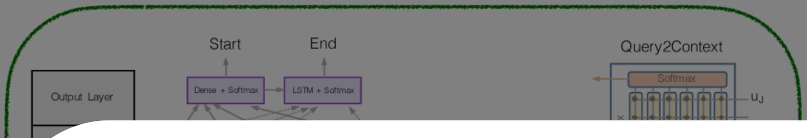
Image credit: J & M, edition 3

BERT models

- ✓ 110M or 330M parameters
- ✓ Transformer 위에 build
- ✓ Pre-trained

02 Reading Comprehension

Comparisons between BiDAF and BERT models



- ✓ Question과 passage 사이의 interaction model
- ✓ BERT는 question과 passage의 concatenation 사이에서 self-attention 사용 = attention(P, P) + attention(P, Q) + attention(Q, P) + attention(Q, Q)
- ✓ BiDAF에 passage에 대한 self-attention layer (attention(P,P)) 추가하면 성능이 좋아진다

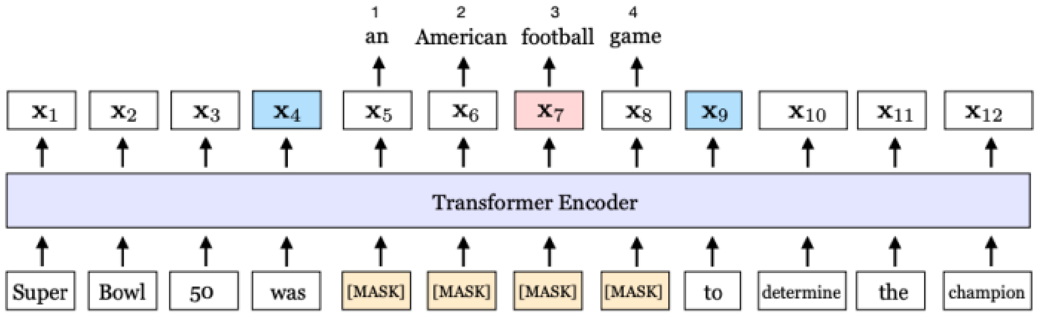
- ✓ ~2.5M parameters
- ✓ 다수의 bidirectional LSTM 위에 build
- ✓ Only built on top of GloVe

- ✓ 110M or 330M parameters
- ✓ Built on top of Transformers
- ✓ Pre-trained

02 Reading Comprehension

Can we design better pre-training objectives?

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$
$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$



- ✓ 15%의 랜덤 단어 대신 인접한 단어의 span을 마스킹하자
- ✓ endpoint 사이에 있는 마스크된 단어들을 예측하기 위해 span의 2개의 endpoint 사용 하자

= 2개의 endpoint에 span의 정보를 압축하여 넣자

(Joshi & Chen et al., 2020): SpanBERT: Improving Pre-training by Representing and Predicting Spans

02 Reading Comprehension

Is reading comprehension solved?

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Adversarial distracting sentence

	Match Single	Match Ens.	BiDAF Single	BiDAF Ens.
Original	71.4	75.4	75.5	80.0
AddSent	27.3	29.4	34.3	34.2
AddOneSent	39.0	41.8	45.7	46.9
AddAny	7.6	11.7	4.8	2.7
AddCommon	38.9	51.0	41.7	52.6

SQuAD에 대해서는 이미 인간보다 뛰어남! 그렇다면 reading comprehension은 다 해결된 문제라 볼 수 있을까? 아니다!

문제점

- ✓ Adversarial example에 대해 낮은 성능을 보임
- ✓ Out-of-domain distribution의 example에 대해 낮은 성능

(Jia and Liang, 2017): Adversarial Examples for Evaluating Reading Comprehension Systems

02 Reading Comprehension

Is reading comprehension solved?

한 데이터셋에 대해 train된 시스템들은 다른 데이터셋으로 generalize 못함

		Evaluated on				
		SQuAD	TriviaQA	NQ	QuAC	NewsQA
Fine-tuned on	SQuAD	**75.6**	46.7	48.7	20.2	41.1
	TriviaQA	49.8	**58.7**	42.1	20.4	10.5
	NQ	53.5	46.3	**73.5**	21.6	24.7
	QuAC	39.4	33.1	33.8	**33.3**	13.8
	NewsQA	52.1	38.4	41.7	20.4	**60.1**

(Sen and Saffari, 2020): What do Models Learn from Question Answering Datasets?

02 Reading Comprehension

Is reading comprehension solved?

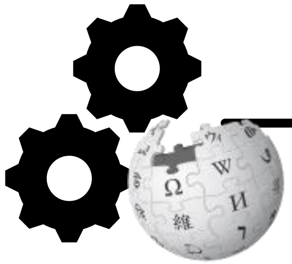
BERT-large model trained on SQuAD

Temporal	*MFT:* change in one person only	41.5	**C:** Both Luke and Abigail were writers, but there was a change in Abigail, who is now a model. **Q:** Who is a model? **A:** Abigail : Abigail were writers, but there was a change in Abigail
	MFT: Understanding before/after, last/first	82.9	**C:** Logan became a farmer before Danielle did. **Q:** Who became a farmer last? **A:** Danielle : Logan
Neg.	*MFT:* Context has negation	67.5	**C:** Aaron is not a writer. Rebecca is. **Q:** Who is a writer? **A:** Rebecca : Aaron
	MFT: **Q** has negation, **C** does not	100.0	**C:** Aaron is an editor. Mark is an actor. **Q:** Who is not an actor? **A:** Aaron : Mark
Coref.	*MFT:* Simple coreference, he/she.	100.0	**C:** Melissa and Antonio are friends. He is a journalist, and she is an adviser. **Q:** Who is a journalist? **A:** Antonio : Melissa
	MFT: Simple coreference, his/her.	100.0	**C:** Victoria and Alex are friends. Her mom is an agent **Q:** Whose mom is an agent? **A:** Victoria : Alex
	MFT: former/latter	100.0	**C:** Kimberly and Jennifer are friends. The former is a teacher **Q:** Who is a teacher? **A:** Kimberly : Jennifer
SRL	*MFT:* subject/object distinction	60.8	**C:** Richard bothers Elizabeth. **Q:** Who is bothered? **A:** Elizabeth : Richard
	MFT: subj/obj distinction with 3 agents	95.7	**C:** Jose hates Lisa. Kevin is hated by Lisa. **Q:** Who hates Kevin? **A:** Lisa : Jose

(Ribeiro et al., 2020): Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Open-domain question answering

Open-domain question answering

$Question(Q) \longrightarrow$  $\longrightarrow$ Answer(A)

Open- domain vs closed-domain?

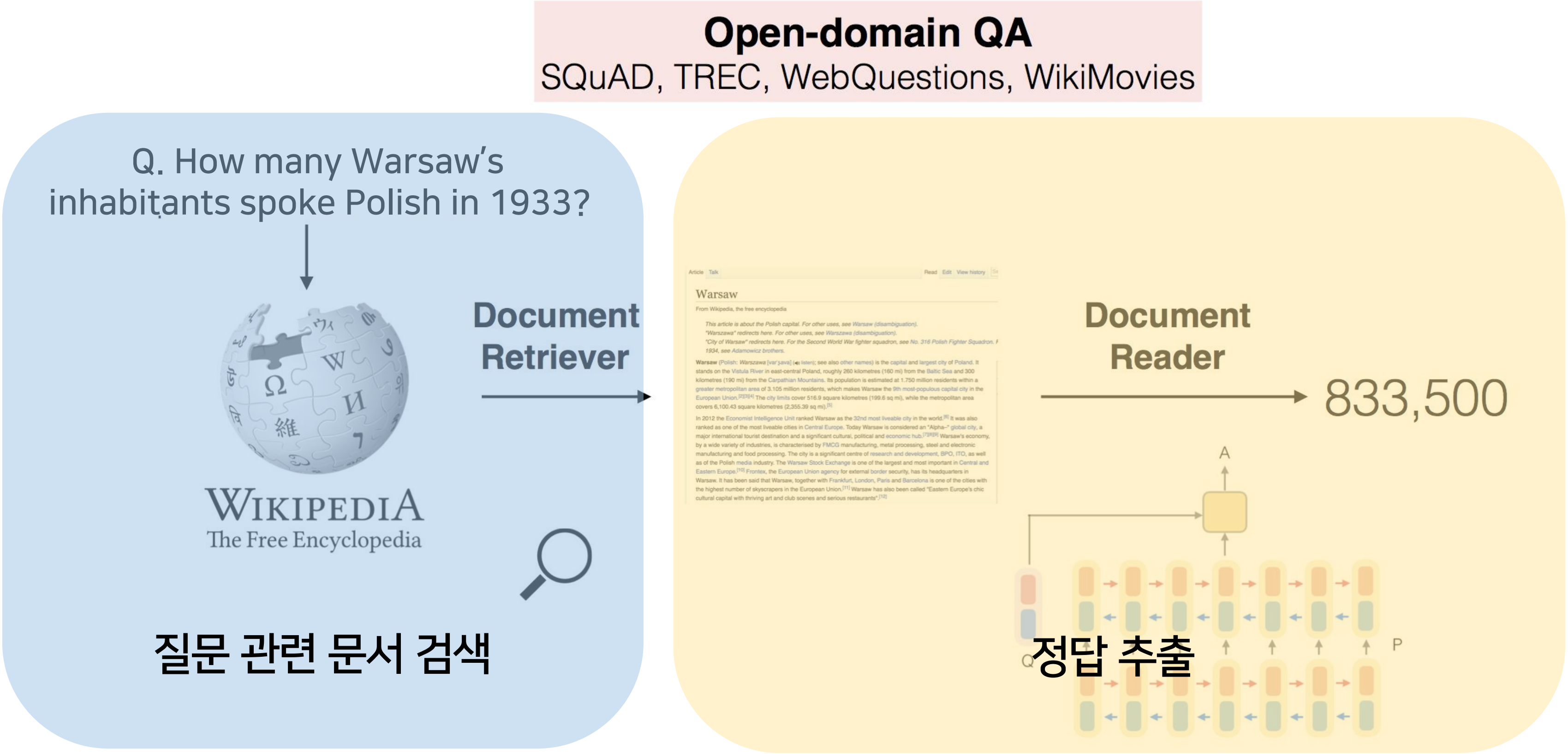
- ✓ passage가 주어져 있다고 가정하지 않음
- ✓ 대신 다량의 document에 접근 가능 (ex: Wikipedia)
- ✓ 정답이 어디에 위치해 있는지 모름

- ✓ Challenging but practical

어떠한 open-domain 질문에도 답하자!

03 Open-domain question answering

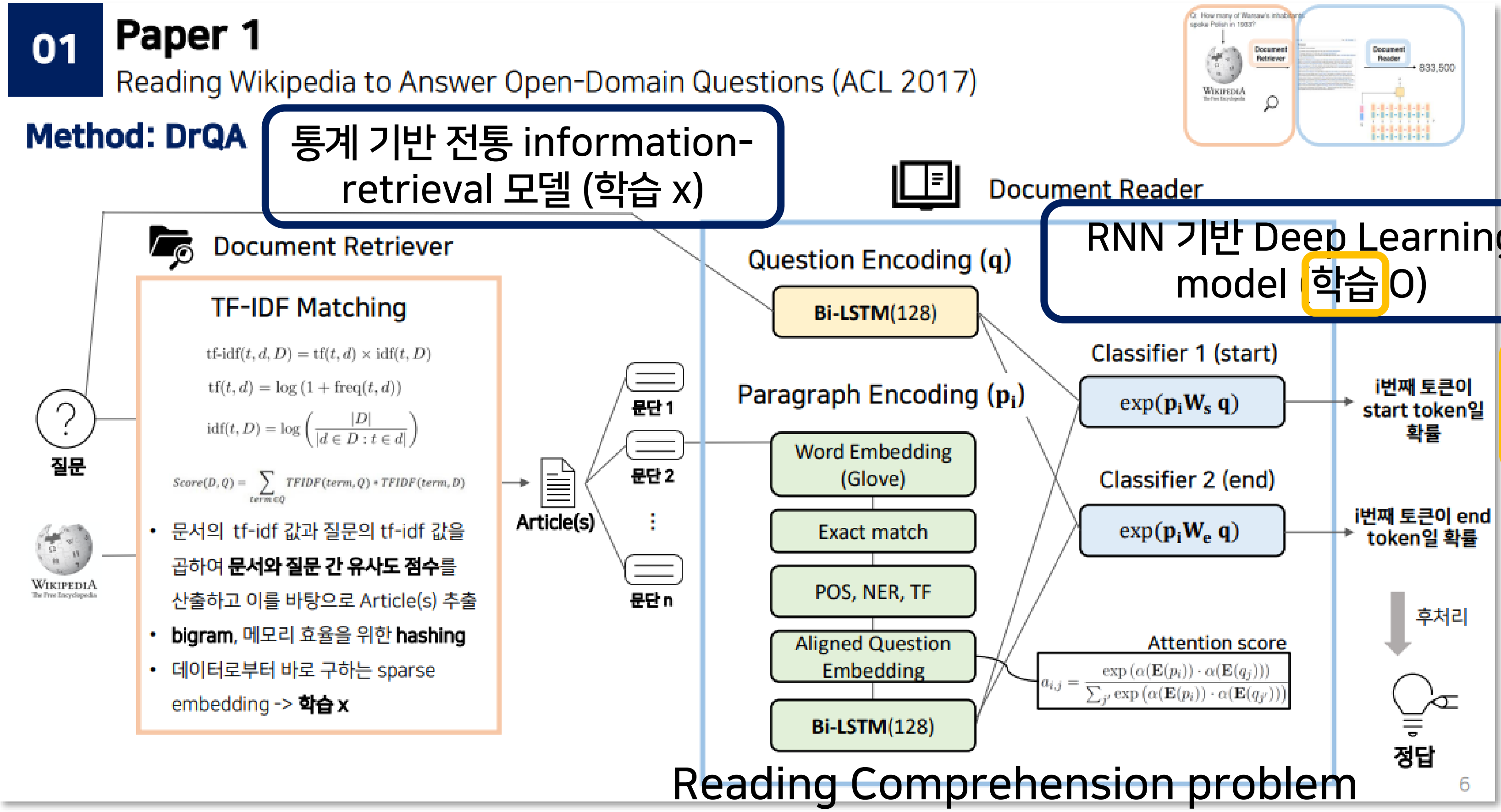
Retriever-reader framework



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

03 Open-domain question answering

Retriever-reader framework



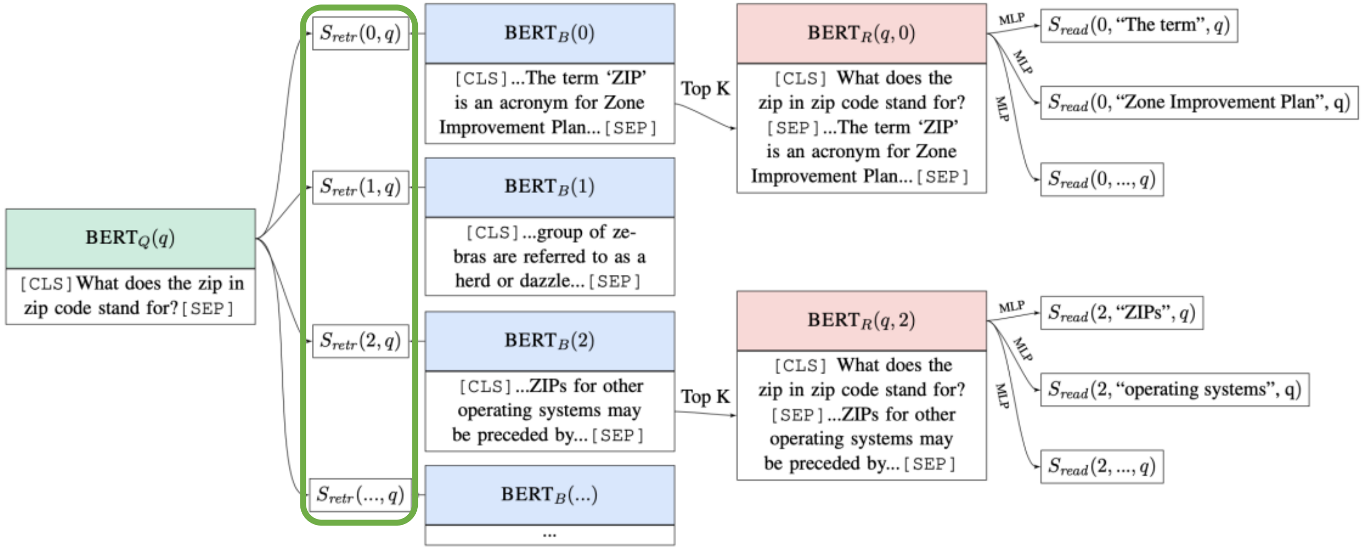
SQuAD,
Distantly Supervised Data (Q,A) →(P,Q,A)

고유경 석사과정 Open-Domain Question Answering Paper Review #1

03 Open-domain question answering

We can train the retriever too

Joint training of retriever and reader

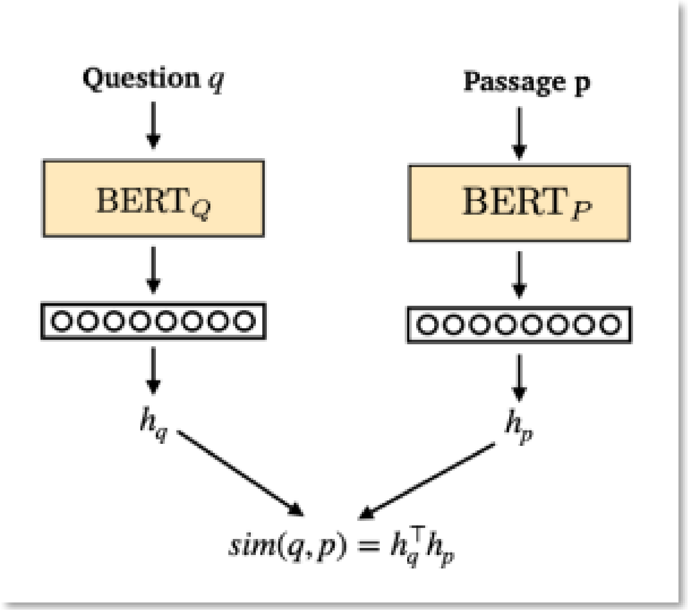


- ✓ question과 passage는 BERT를 이용해 encode될 수 있음
- ✓ Retrieval score = question representation과 passage representation 사이의 dot product
- ✓ BUT passage 수가 많을 때는 모델링이 쉽지 않음

Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering

03 Open-domain question answering

We can train the retriever too



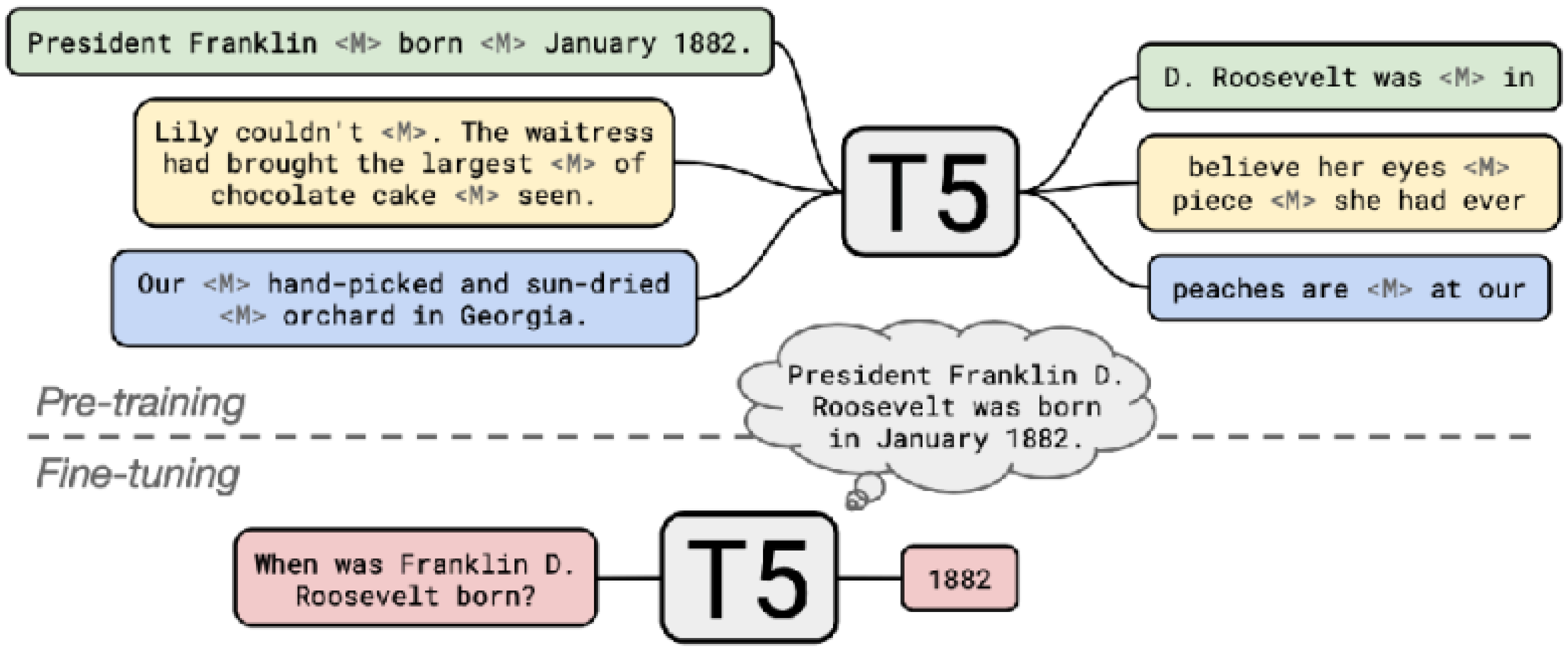
Dense passage retrieval (DPR)

(Q,A) 쌍 이용해 retriever 학습 가능

Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

03 Open-domain question answering

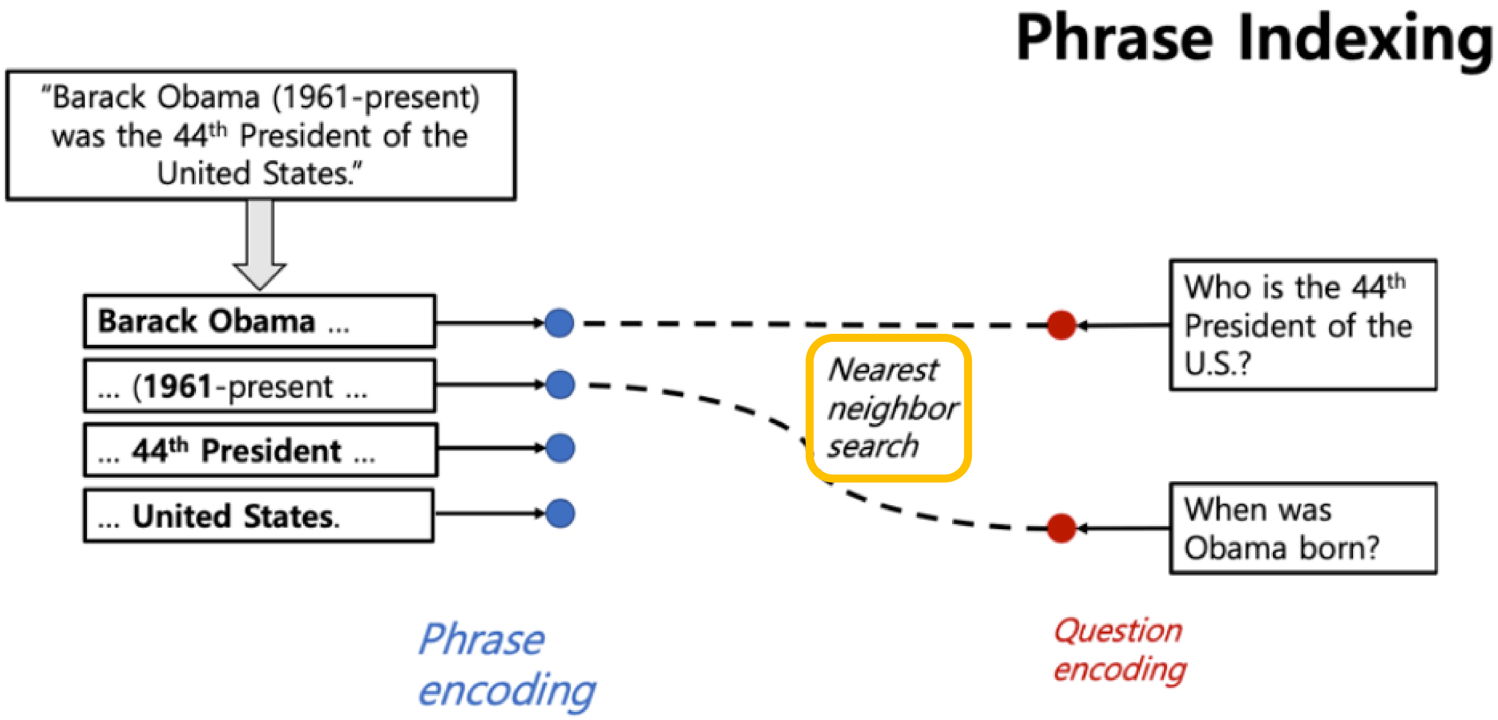
Without an explicit retrieval stage



Roberts et al., 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?

03 Open-domain (textual) question answering

Maybe the reader model is not necessary too!



Seo et al., 2019. Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index
Lee et al., 2020. Learning Dense Representations of Phrases at Scale 49 DensePhrases: Dem

감사합니다

00

Appendix

000