

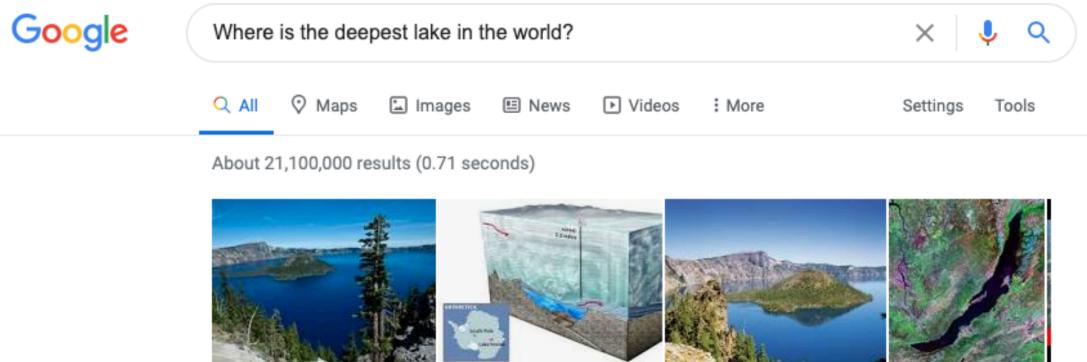
# 0220 Lecture 11&12

Lecture 11. Question Answering  
Lecture 12.

## Question Answering (QA)



- 인간이 언어로 하는 질문에 자동적으로 대답하는 시스템을 구축하는 것
- 2가지 유형
  1. 정답을 포함하고 있는 문서 찾기
    - : 기존의 정보 검색, 웹 검색으로 처리 가능



### Siberia

Lake **Baikal**, in Siberia, holds the distinction of being both the deepest lake in the world and the largest freshwater lake, holding more than 20% of the unfrozen fresh water on the surface of Earth.

세상에서 제일 깊은 호수는? → 객관적인 정답 존재

## 2. 문서에서 정답 찾기

### : MRC (Machine reading comprehension)-읽기 이해

The screenshot shows a Google search results page. The search query is "How can I protect myself from COVID-19?". The top result is a snippet from CDC.gov. It starts with a general statement about prevention and then lists ten specific actions to help prevent the spread of COVID-19. A "Learn more on cdc.gov" button is present, along with a note about informational purposes.

The best way to prevent illness is to avoid being exposed to this virus. Learn how COVID-19 spreads and practice these actions to help prevent the spread of this illness.

To help prevent the spread of COVID-19:

- Cover your mouth and nose with a mask when around people who don't live with you. Masks work best when everyone wears one.
- Stay at least 6 feet (about 2 arm lengths) from others.
- Avoid crowds. The more people you are in contact with, the more likely you are to be exposed to COVID-19.
- Avoid unventilated indoor spaces. If indoors, bring in fresh air by opening windows and doors.
- Clean your hands often, either with soap and water for 20 seconds or a hand sanitizer that contains at least 60% alcohol.
- Get vaccinated against COVID-19 when it's your turn.
- Avoid close contact with people who are sick.
- Cover your cough or sneeze with a tissue, then throw the tissue in the trash.
- Clean and disinfect frequently touched objects and surfaces daily.

[Learn more on cdc.gov](#)

For informational purposes only. Consult your local medical authority for advice.

코로나 19를 스스로 예방하기 위해 어떻게 해야하는가? → 문서를 읽어 스스로 답을 찾아야 하는 경우

텍스트 페이지를 읽어 정답을 도출해내는 것

QA 분야는 text뿐만 아니라 이미지 영역으로 확장되고 있음

## Visual QA



What color are her eyes?  
What is the mustache made of?



How many slices of pizza are there?  
Is this a vegetarian pizza?

## Stanford Question Answering Dataset(SQuAD)

### 1. 질문의 유형

#### 1) Factoid type questions (누가 언제 어디서 무엇을 어떻게)

EX. 중앙대학교는 어디에 있는가? 흑석

#### 2) List type questions

EX. 응용통계학과에 있는 수업은? [기초통계학, 수리통계학 ...]

#### 3) confirmation questions (yes or no)

EX. 오늘은 일요일인가? yes

#### 4) casual questions (why, how)

EX. 오늘 왜 늦었나? 늦잠 자서

#### 5) Hypothetical questions (특정한 정답 정해지지 X)

EX. 러시아랑 중국이랑 전쟁 나면 어떻게 될까?

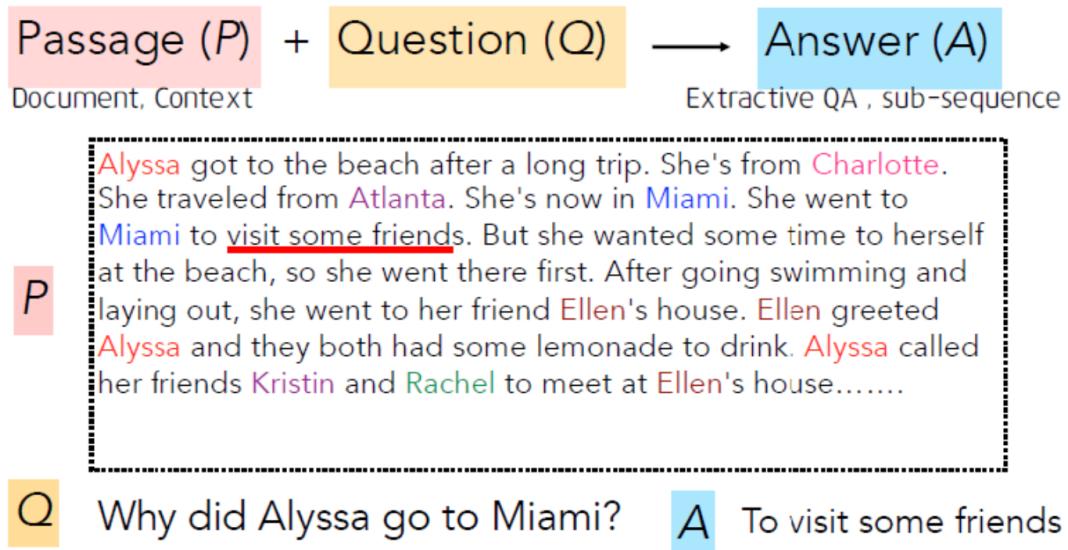
#### 6) complex questions

EX. 우주 너머에는 뭐가 있을까?

- SQuAD는 Factoid type questions에 해당
- English Wikipedia로 부터 수집

- MRC에서 가장 널리 쓰이는 데이터셋으로 알려짐

## 2. SQuAD



$(P, Q) \rightarrow A$

P: Passage document, text

Q: question

A: answer

- Answer는 Passage 속 하위 sequence로 구성 (text안에 정답이 있다!)

---

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?

**gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

**graupel**

Where do water droplets collide with ice crystals to form precipitation?

**within a cloud**

- answer: 위키피디아 문서를 바탕으로 크라우드소싱 인력들이 해당하는 질문과 답변 생성

### 3. Evaluation

- 3가지 답변 sampling
- exact match (0 or 1) : 예측 답과 실제 답이 일치하면 1 아니면 0
- F1 (partial credit): 예측된 답과 실제 답의 중첩 토큰 계산 (겹치는 토큰이 얼마나 있는지), 좀 더 안정적인 평가 방법

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:  $\max\{0.67, 0.67, 0.61\} = 0.67$

## 4. 한계

- passages 내에서만 정답을 찾도록 하는 질문 구성
  - : 여러 문서들을 비교하여 진짜 정답을 찾는 것이 아닌 특정 문서 내에서 정답 도출  
→ 잘못 기입된 문서일 경우 잘못된 답 도출
- 하지만 여전히 QA 문제에 있어 가장 많이 사용되는 데이터셋
- TriviaQA, Natural Questions, HotpotQA 등 QA 문제를 위한 여러 데이터셋 존재

## QA model

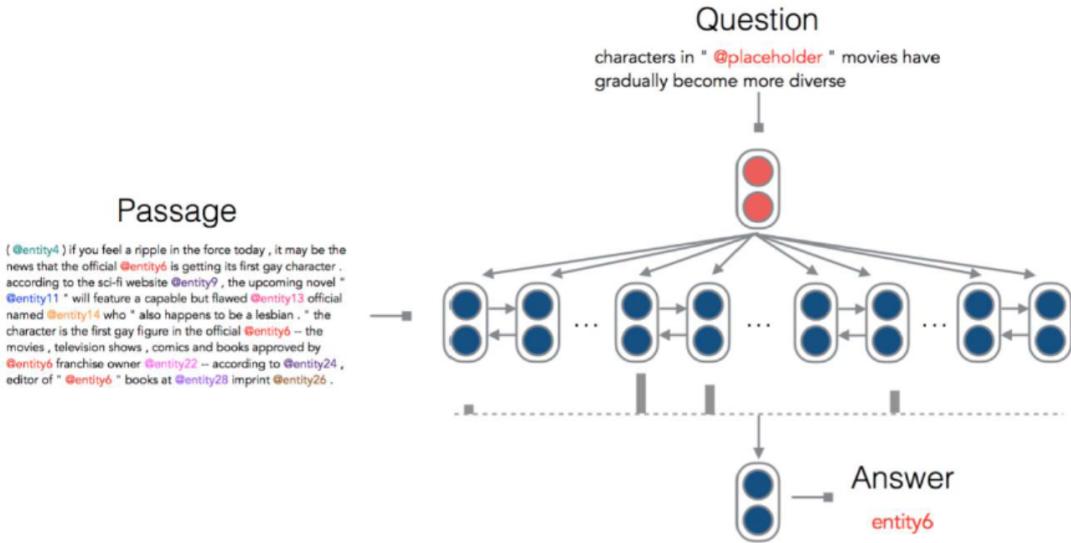
- Input:  $C = (c_1, c_2, \dots, c_N), Q = (q_1, q_2, \dots, q_M), c_i, q_i \in V$
- Output:  $1 \leq \text{start} \leq \text{end} \leq N$

C: context (paragraph)

Q: questions

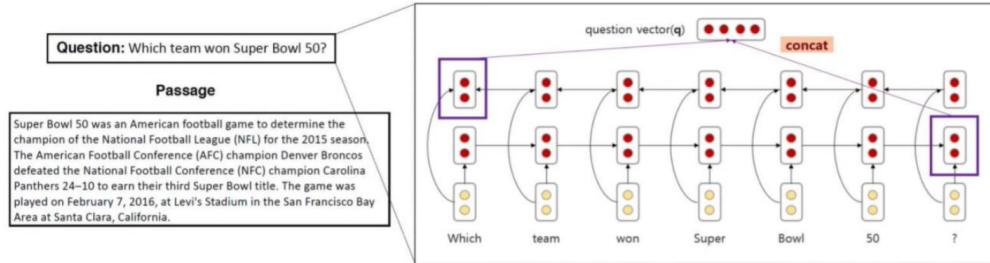
### 1. Standford Attentive Reader

- 질문에 대한 응답을 찾는 모델 구축 위해 Bi LSTM with attention 적용



- 과정

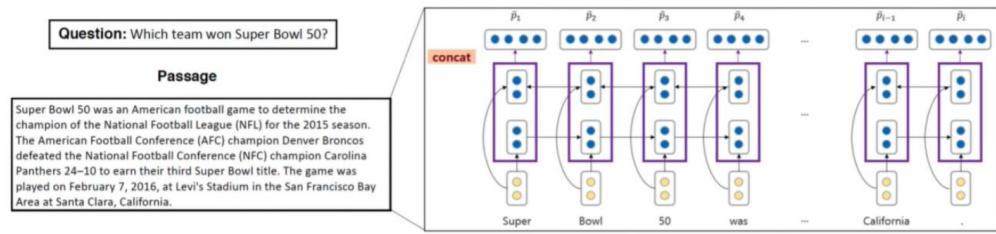
### 1. Question vector 생성



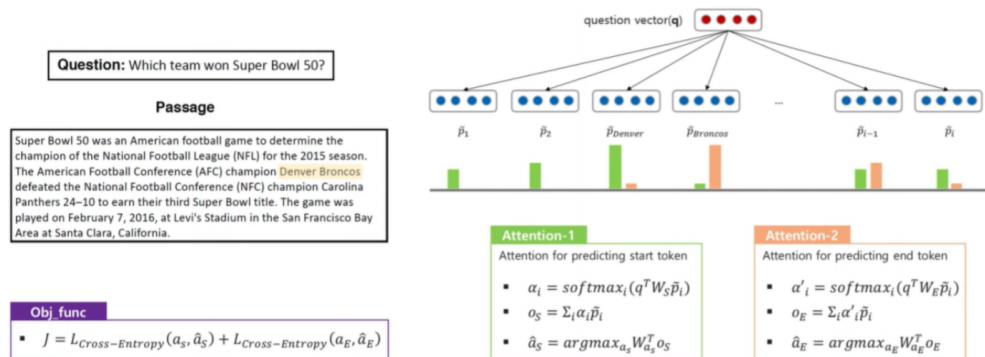
- 주어진 question 문장에 있는 단어들의 embedding을 사전에 학습된 glove에서 가져와 one-layer BiLSTM에 넣음
- BiLSTM의 각 방향 마지막 hidden state를 concat하여 **question vector**를 얻음

### 2. Passage vector 생성

- 단어들의 embedding을 사전에 학습된 glove에서 가져와 one-layer BiLSTM에 넣음
- 각각의 포지션에서 BiLSTM의 hidden state를 concat하여 문장 내 단어 개수 만큼의 **passage vector**를 생성



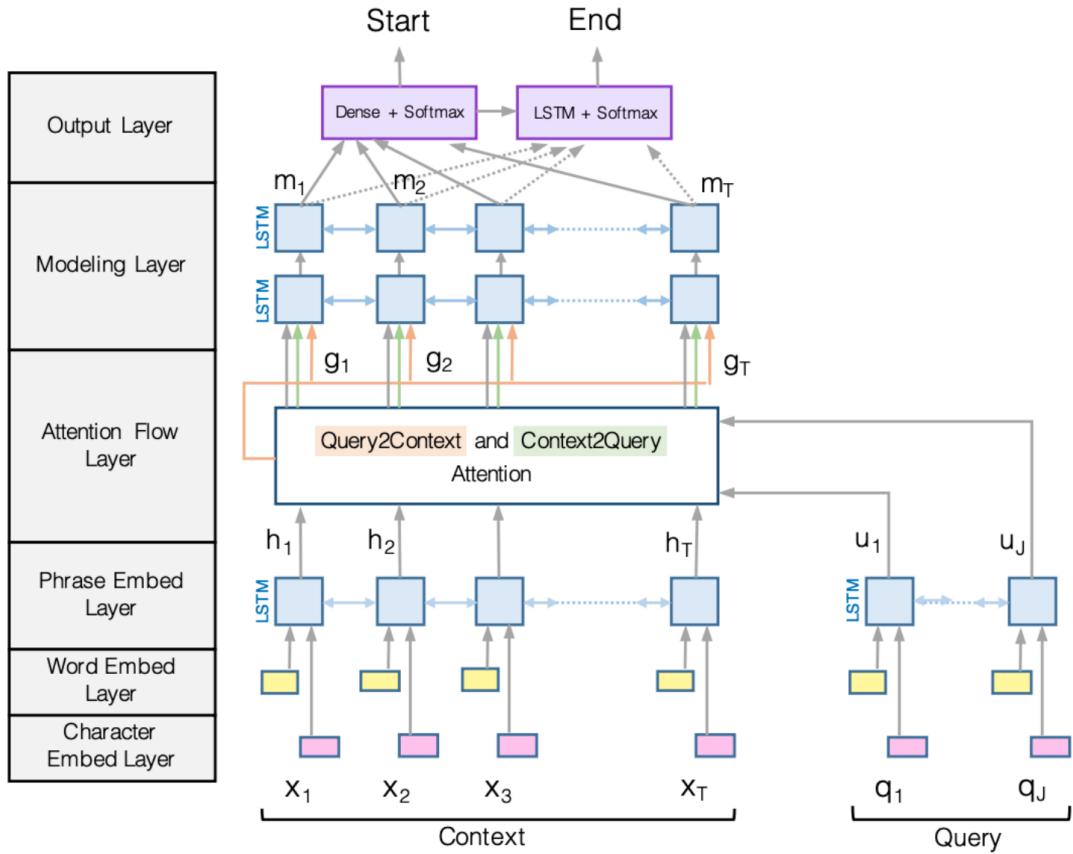
### 3. Attention



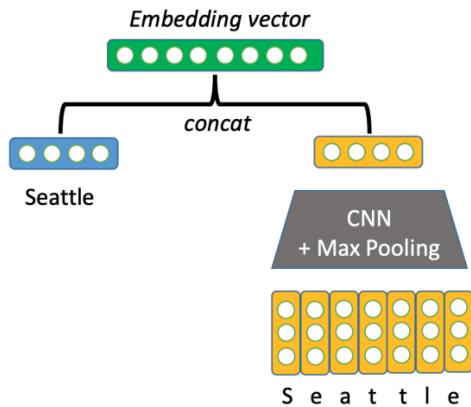
- $\alpha_i$ : i 개의  $p$  벡터와 한 개의  $q$  벡터를 이용하여 어텐션을 적용한 후 소프트 맥스를 취함.
- $o_S$ :  $\alpha_i$ 와  $p_i$  벡터를 곱하여 모두 더함
- $a_S$ :  $o_S$ 에 linear transform 을 취함
- 이렇게 start token 과 end token 구하기

## BiDAF : Bi-Directional Attention Flow for Machine Comprehension

- Question에서 Passage로 한 방향으로만 진행되는 Standford Attentive reader 와 달리 **attention이 양방향**으로 적용된 모델



[Embedding]



### 1. Character Embedding Layer

charCNN 을 사용하여 각 단어를 vector space 에 mapping

### 2. Word Embedding Layer

pre-trained word embedding model을 사용하여 각 단어를 vector space 에 mapping

### 3. Phrase Embed layer

BiLSTM 에 넣어 2d-dimension vector

#### [Attention Layer]

- Query to Context : Context의 어떤 정보가 Query 와 관련이 있는지
- Context to Query : Query 의 어떤 정보가 Context 와 관련이 있는지
- query 와 context(=passage) 를 single feature vector로 요약하지 않고, **query 와 context 를 연결시킴**
- **shared matrix 활용**