텍스트 세미나13&14

ToBig's 13기 강의자
김미성

Lecture 10. Question Answering

2020.11.25 (수)

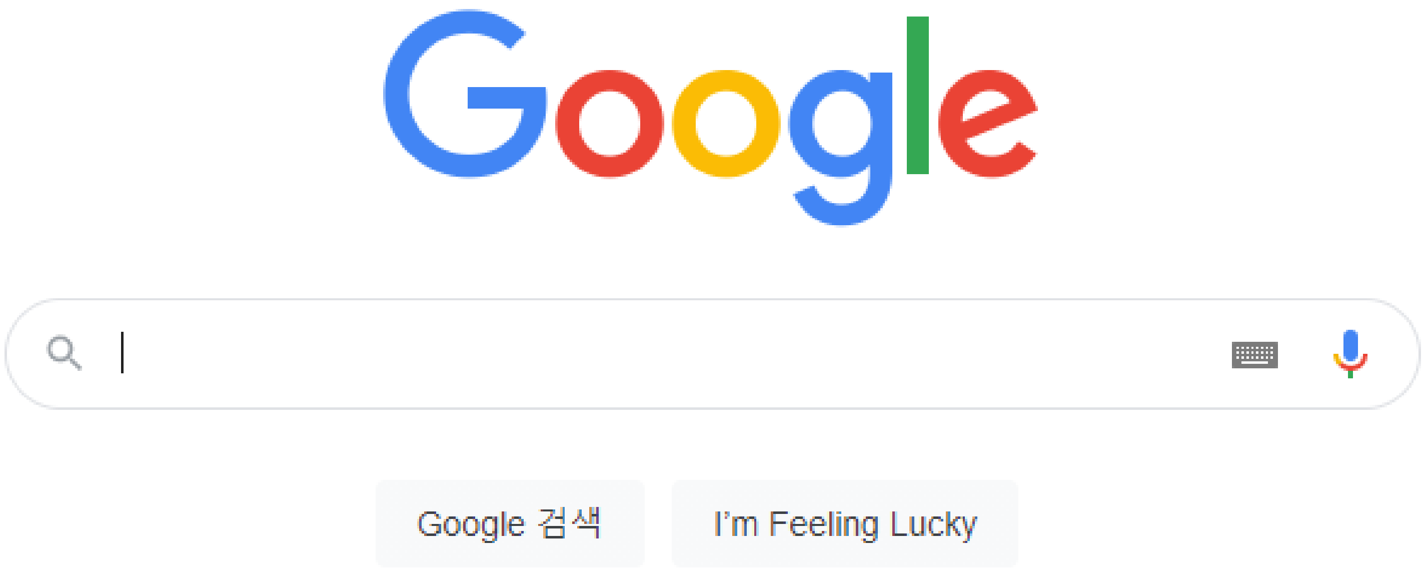
Contents

텍스트 세미나13&14

ToBig's 13기 강의자
김미성

Motivation & History

Unit 01 | Motivation & History



Unit 01 | Motivation & History

Unit 01 | Motivation & History

1. Motivation : Question answering

우리는 종종 questions 에 대한 answer 를 원한다.

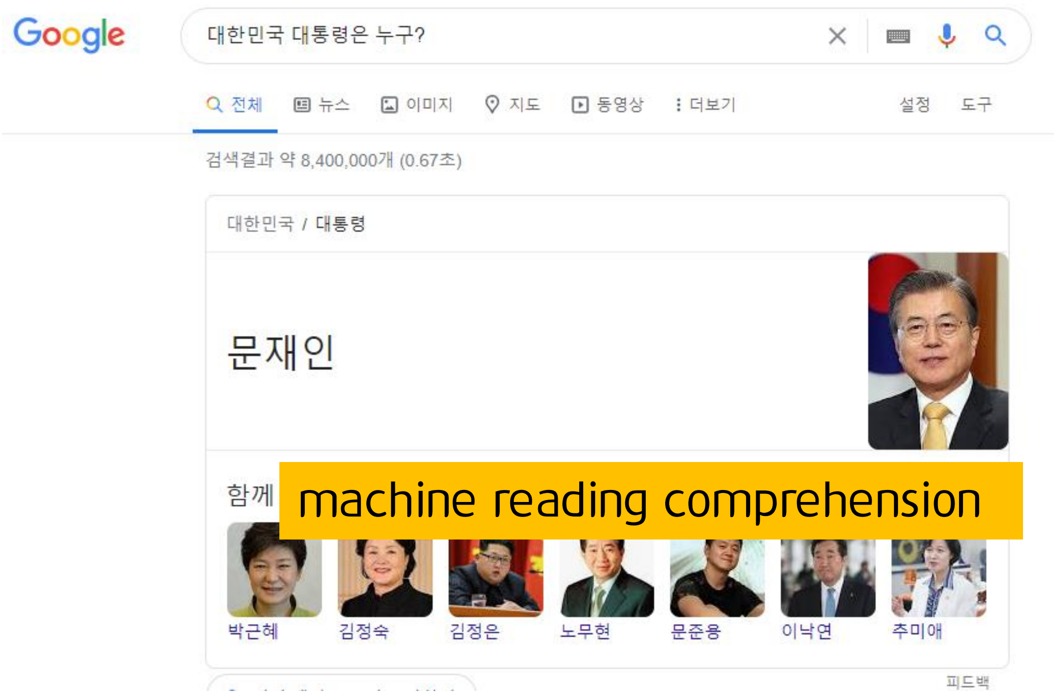
1. Finding documents that contain an answer (정답을 포함하고 있는 문서 찾기)

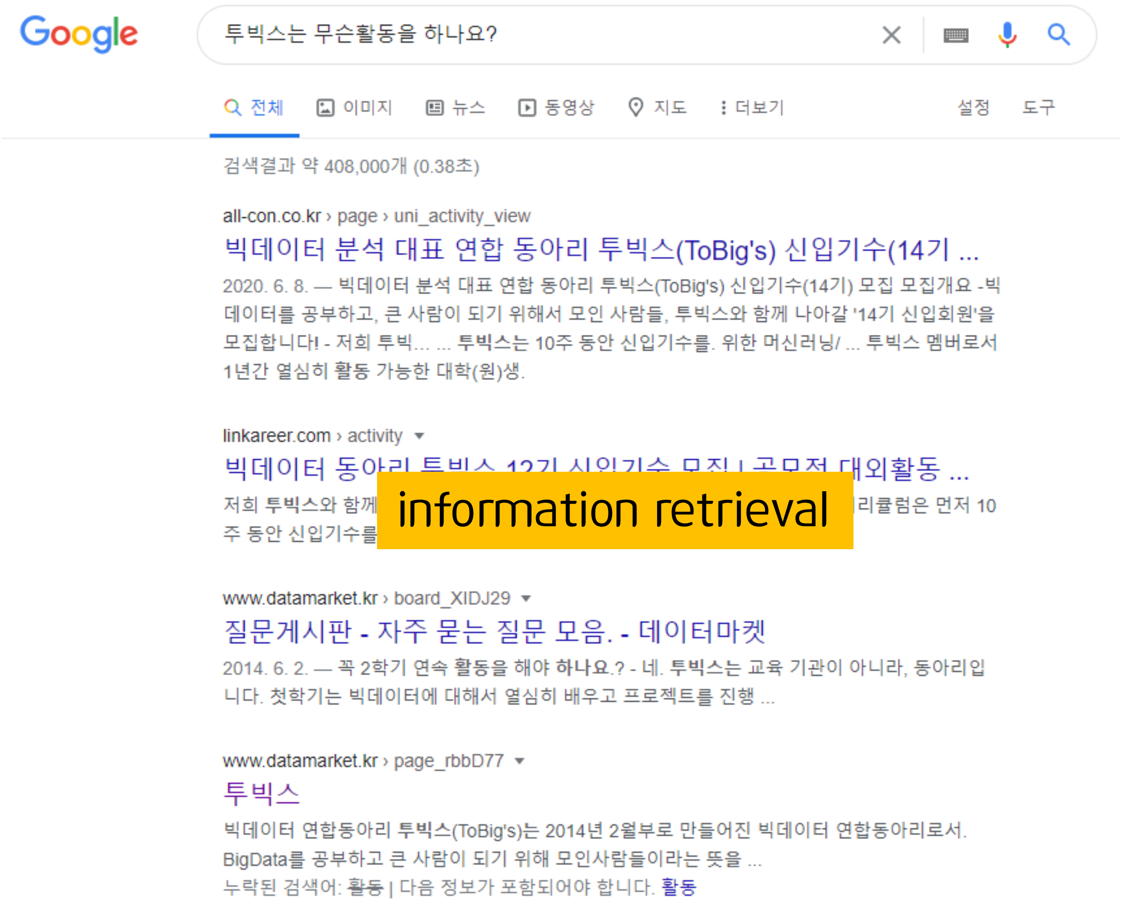
- 기존의 정보 검색(information retrieval = IR) / 웹 검색으로 처리할 수 있는 사항.

2. Finding an answer in a paragraph or a document (문서에서 정답 찾기)

- 이 문제를 흔히 읽기 이해(machine reading comprehension = MRC) 라고 한다.
- 오늘날 우리가 중심적으로 다룰 사항

Unit 01 | Motivation & History





2. MCTest Reading Comprehension

QA dataset은 이렇게 3개로 구성되어있다.

Passage (P) + Question (Q) ⟶ Answer (A)

Document, Context

Extractive QA , sub-sequence

P

Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house.......

Q Why did Alyssa go to Miami?

A To visit some friends

텍스트 세미나13&14

ToBig's 13기 강의자
김미성

Standford Question Answering Dataset (SQuAD)

Unit 02 | Stanford Question Answering Dataset (SQuAD)

1. Types of questions

1) **Factoid type questions** [what, which, when, who, how] → SQuAD
   Ex) 한국의 수도는? 서울
2) **List type questions** [list of facts or answers]
   Ex) 투빅스 멤버에는 누가 있는가? 미성, 예진, 재빈, 민준, 진석, 주원, 혜원, 혜빈, 민정, 예은, 경태, 혜린, 거홍 ….
3) **Confirmation questions** [yes or no]
   Ex) 오늘이 월요일 인가요? 아니
4) **Casual Questions** [why, how]
   Ex) 그는 오늘 왜 늦었어? 차가 막혀서
5) **Hypothetical Questions** [no specific answers]
   Ex) 남한과 북한이 전쟁이 나면 어떻게 될까? ???
6) **Complex Questions**
   Ex) 환경오염의 이유가 무엇일까? ???

Unit 02 | Stanford Question Answering Dataset (SQuAD)

2. Stanford Question Answering Dataset(SQuAD)

SQuAD(1.1) dataset은 이렇게 구성 되어있다.

Question: Which team won Super Bowl 50?

Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

100k examples

Answer must be a span in the passage

A.k.a. extractive question answering

1. 한 passage 당 5개의 질문 정도
2. 질문에 대한 답은 항상 구절에 나온 하위 sequence
3. 위키피디아 문서를 바탕으로 크라우드소싱 인력들이 해당하는 질문과 답변을 생성함.

➔ 이 dataset이 만들어지면서 급격히 발전!

2. Stanford Question Answering Dataset(SQuAD)

SQuAD(1.1) dataset의 평가 방법

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
Along with sport and art, what is a type of talent scholarship?
Gold answers: ① academic ② academic ③ academic
Rather than taxation, what are private schools largely funded by?
Gold answers: ① tuition ② charging their students tuition ③ tuition

1. 3가지의 답변을 sampling 한다.
2. Exact Match
- 예측된 답과 실제 답이 정확히 일치하면1, 그렇지 않으면 0
3. F1 score
- 더 안정적이고 좋은 평가 방법임
- 예측된 답과 실제답의 중첩토큰을 계산함.
4. Both metrics ignore punctuation and articles

Unit 02 | Stanford Question Answering Dataset (SQuAD)

2. Stanford Question Answering Dataset(SQuAD)

SQuAD(2.0) dataset

Question: Which team won Super Bowl 50?

Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

100k examples

Answer must be a span in the passage

A.k.a. extractive question answering

+ added unanswerable questions in SQuAD 2.0

2. Stanford Question Answering Dataset(SQuAD)

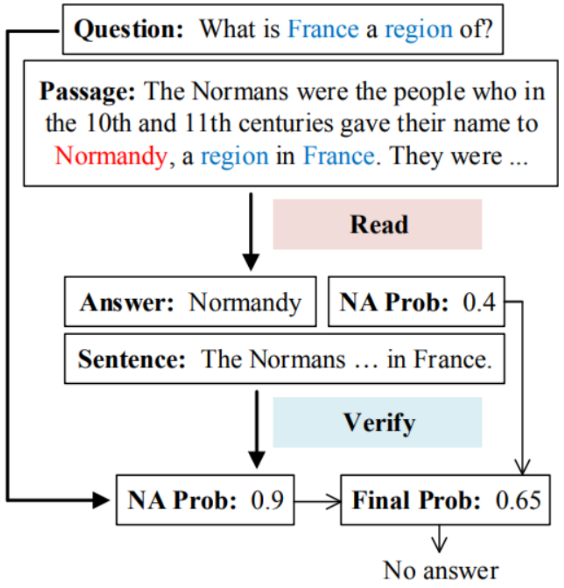
- No Answer를 어떻게 처리할 수 있을까?
  1. 임계값을 사용한다.

Unit 02 | Stanford Question Answering Dataset (SQuAD)

2. Stanford Question Answering Dataset(SQuAD)

- No Answer를 어떻게 처리할 수 있을까?

Machine Reading Comprehension with Unanswerable Questions (2019-AAAI)



1. Read-then-verify 시스템을 제안
2. Passage에 대한 question에 대한 정답과 no-answer 확률을 각각 도출함.
3. 이렇게 도출한 정답이 적절한지 확인을 한다.

- 질문: 프랑스는 어떤 지역입니까?
- 답변: 노르망디는 프랑스의 지역입니다.

https://ai-information.blogspot.com/2019/05/nl-037-read-verify-machine-reading.html

2. Stanford Question Answering Dataset(SQuAD)

- **SQuAD Limitations**
  - Span-based answers 만 존재 (no yes/no, counting, implicit why)
  - Passages 내에서만 정답을 찾도록 하는 질문 구성
    - ➢ 여러 문서들을 비교하여, 진짜 정답을 찾아낼 필요가 없음
    - ➢ 실제 마주하게 될 질문-답변 보다 잘 정립된 문법 구조
  - 동일 지시어(coreference) 문제를 제외하고는 Multi-fact 문제, 문장 추론 문제가 거의 없음
- Nevertheless, SQuAD is a well-targeted, well-structured, clean dataset
  - QA 문제를 푸는데 있어 가장 많이 사용되고, 경쟁하고 있는 데이터셋
  - 실제 시스템을 개발하기 위한 유용한 starting point

SQuAD

Home Explore 2.0 Explore 1.1

Stanford **Qu**estion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 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.

Explore SQuAD2.0 and model predictions

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

SQuAD 1.1, the previous version of the SQuAD dataset, contains 100,000+ question-answer pairs on 500+ articles.

Explore SQuAD1.1 and model predictions

SQuAD1.0 paper (Rajpurkar et al. '16)

Getting Started

We've built a few resources to help you get started with

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

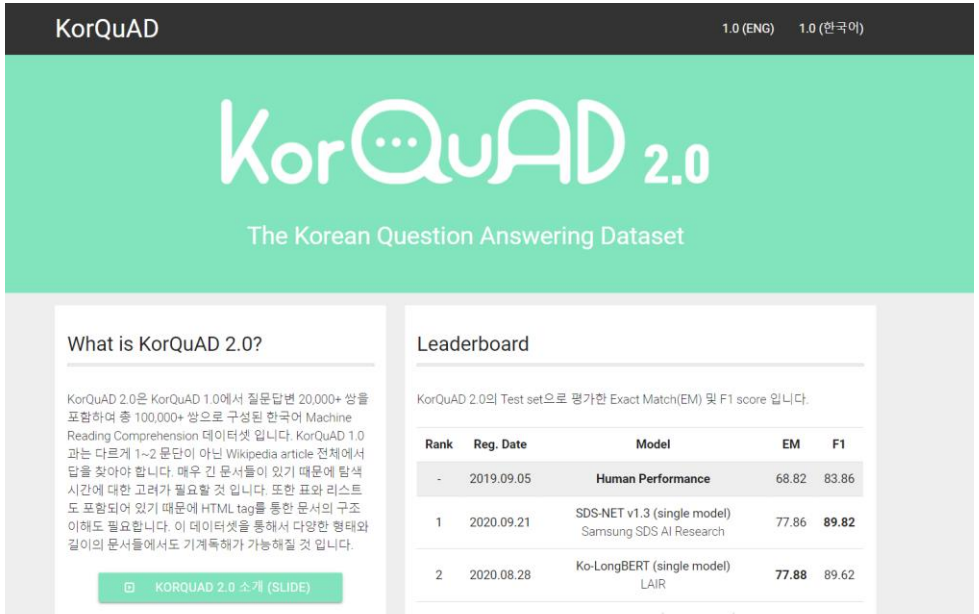
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
3 Jul 31, 2020	ATRLP+PV (ensemble) *Hithink RoyalFlush*	90.442	92.877
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839
4 Jun 21, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.420	92.799
4 Sep 11, 2020	EntitySpanFocus+AT (ensemble) *RICOH_SRCB_DML*	90.454	92.748
4 Mar 12, 2020	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.386	92.777

https://rajpurkar.github.io/SQuAD-explorer/

Unit 02 | Stanford Question Answering Dataset (SQuAD)

참고

- 한국어 위키백과를 대상으로 대규모 MRC 데이터 구축




```
{
  "creator": "KorQuAD",
  "data": [
    {
      "title": "존재와 무",
      "paragraphs": [
        {
          "context": "《존재와 무》 (l'êtreetlenéant:essaid'ontologiephénoménologique)는 프랑스의 철학자 샤르트르가 1943년 출판한 책이다. 사르트의 주된 목적은 개인의 존재에 앞서 개인의 실존을 주장하는 것이다. 이 책을 작성하면서 최우선적으로 염두한 것은 자유가 존재한다는 것이다. 1940년과 1941년에 전쟁 포로로 억류되어있던 시절 샤르트르는 마르틴 하이데거의 《존재와 시간》을 읽고 자기 자신만의 자유를 전개하였다. 하이데거의 영향을 받았음에도 샤르트르는 하이데거의 존재와의 가설적인 재조우와 비교하여 인간성(humanity)이 개인적인 성취의 상태를 달성할 수 있다는 방법에 회의적이었다.",
          "qas": [
            {
              "id": "9_f9_wiki_4511-1",
              "answers": [
                {
                  "answer_start": 69,
                  "text": "사르트르/nnp"
                }
              ],
              "question": "책 존재와 무의 저자가 누구야"
            },
            {
              "id": "9_f9_wiki_4512-1",
              "answers": [
                {
                  "answer_start": 60,
                  "text": "프랑스"
                }
              ],
              "question": "사르트르는 어느나라 철학자야"
            }
          ]
        }
      ]
    }
  ]
}
```


사진 출처 : http://blog.naver.com/PostView.nhn?blogId=jeanmy1102&logNo=221749234951

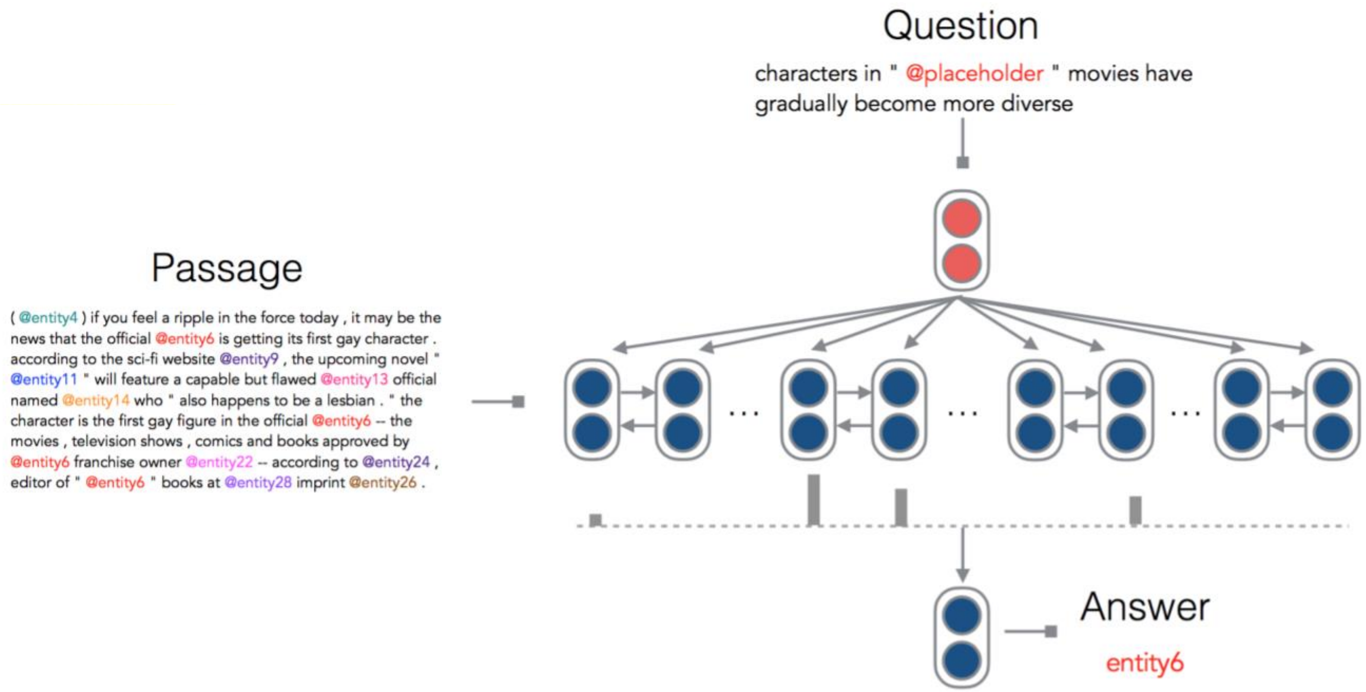
ToBig's 13기 강의자
김미성

QA model

Unit 03 | QA model

1. Stanford Attentive Reader

- 질의에 대한 응답을 찾는데 BiLSTM with attention을 적용

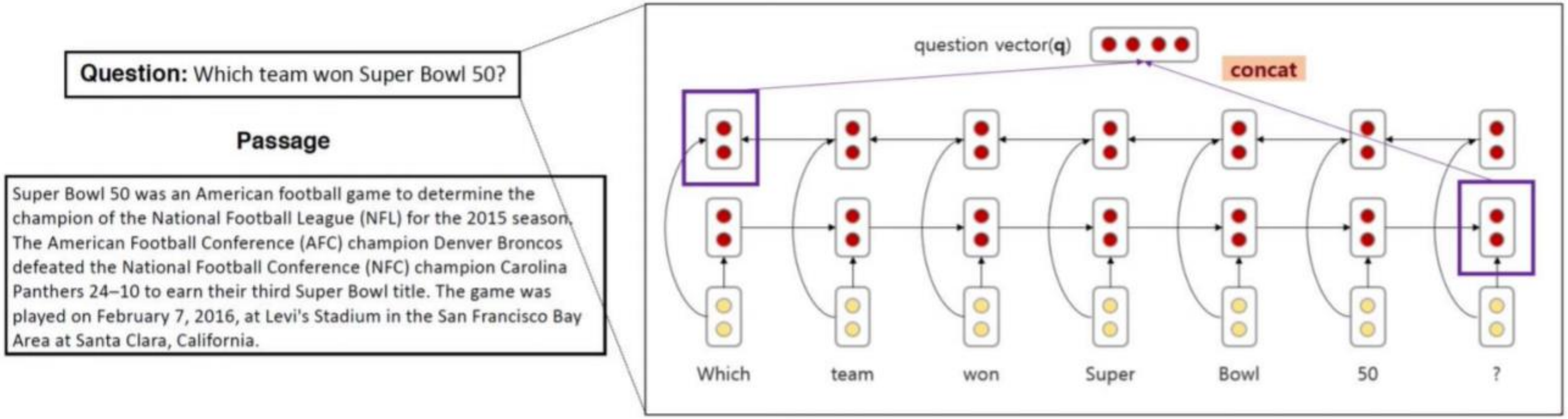


Unit 03 | QA model

1. Stanford Attentive Reader

1. Question 에 대한 vector 생성

- 주어진 문장에 있는 단어들을 사전에 학습된 Glove 에서 가져옴
- One layer BiLSTM 에 넣음.
- 각 방향의 마지막 hidden state를 concat 하여 고정된 벡터를 얻음.

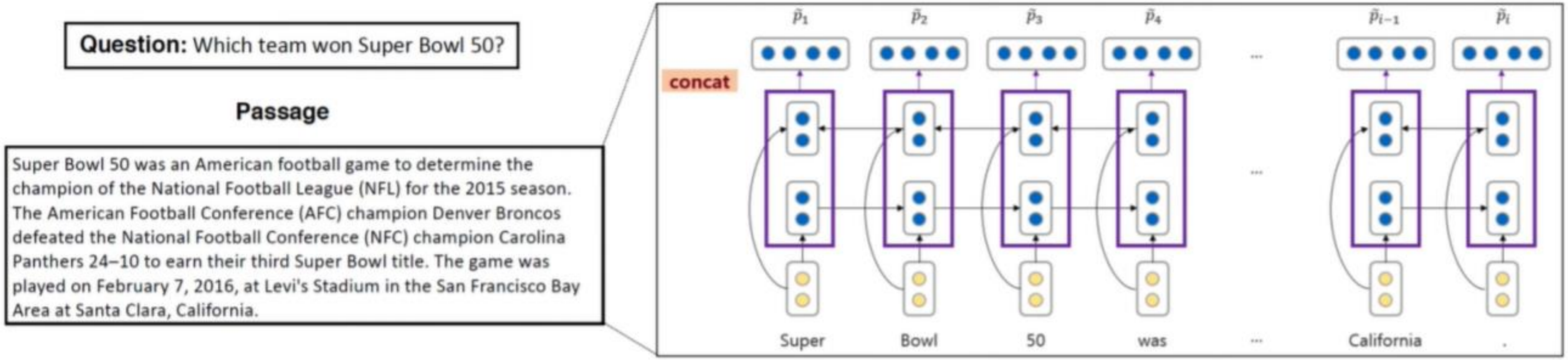


Unit 03 | QA model

1. Stanford Attentive Reader

2. Passage 에 대한 vector 생성

- 주어진 문장에 있는 단어들을 사전에 학습된 Glove 에서 가져옴
- One layer BiLSTM 에 넣음.
- 각각의 포지션에서 BiLSTM의 hidden state를 concat해서 문장 내 단어 개수 만큼의 passage vector 생성!



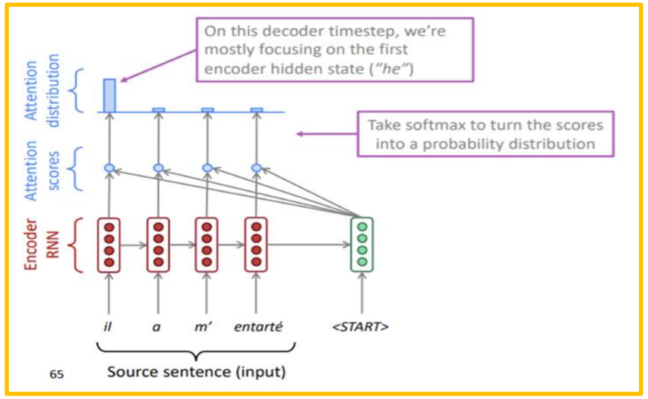
Unit 03 | QA model

1. Stanford Attentive Reader

3. Attention

- αi : i 개의 p벡터와 한 개의 q벡터를 이용하여 어텐션을 적용한 후 소프트맥스를 취함.
- os : αi와 pi벡터를 곱하여 모두 더함.
- as : os에 linear transform을 취함.

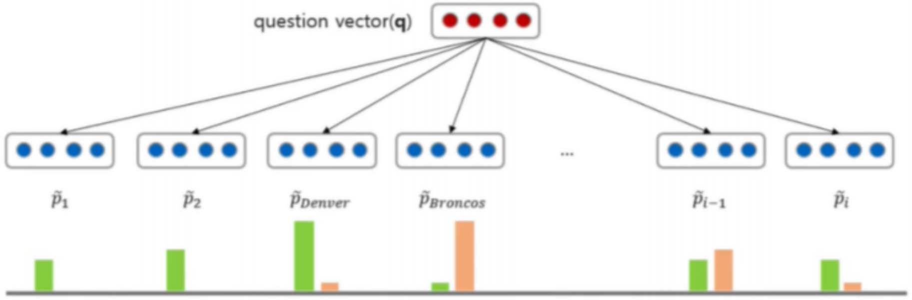
→ 이렇게 start token 과 end token 을 구한다!



Question: Which team won Super Bowl 50?

Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.



Obj_func

- $J = L_{Cross-Entropy}(a_s, \hat{a}_S) + L_{Cross-Entropy}(a_E, \hat{a}_E)$

Attention-1

Attention for predicting start token

- $\alpha_i = softmax_i(q^T W_S \tilde{p}_i)$
- $o_S = \Sigma_i \alpha_i \tilde{p}_i$
- $\hat{a}_S = argmax_{a_s} W_{a_s}^T o_S$

Attention-2

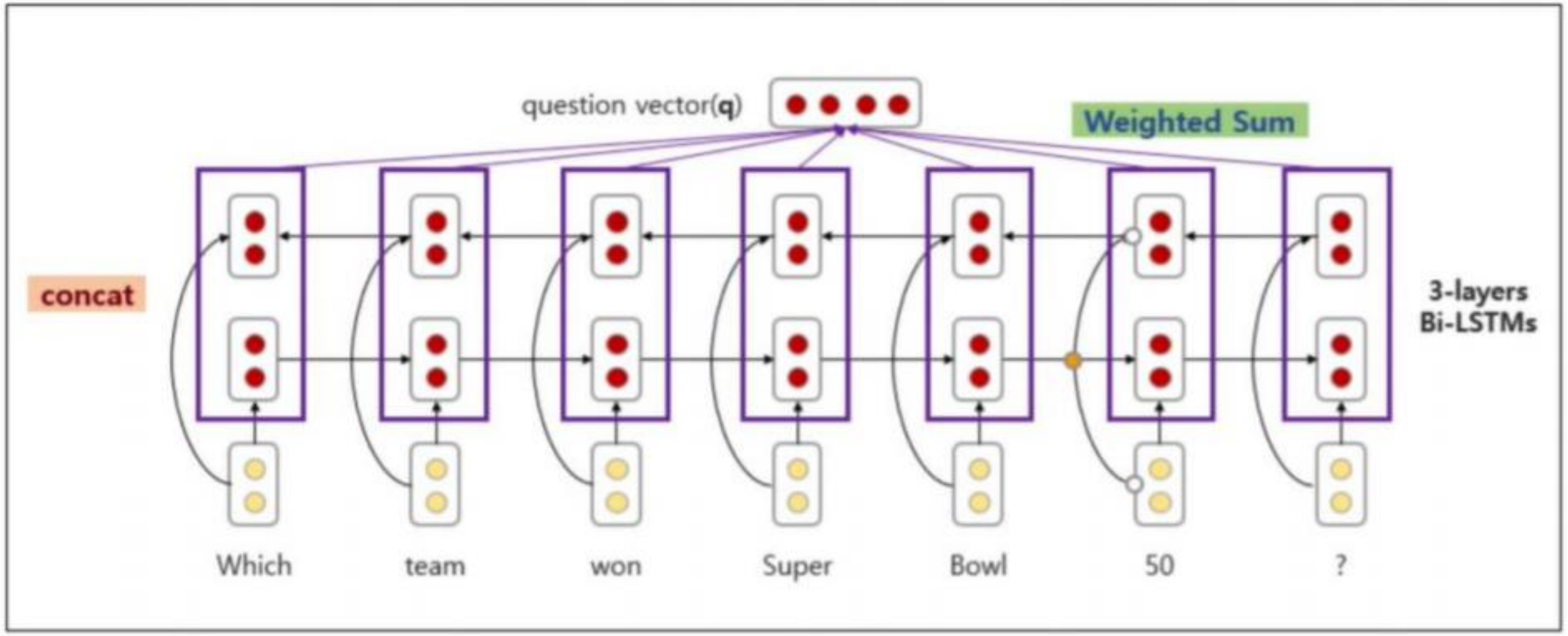
Attention for predicting end token

- $\alpha'_i = softmax_i(q^T W_E \tilde{p}_i)$
- $o_E = \Sigma_i \alpha'_i \tilde{p}_i$
- $\hat{a}_E = argmax_{a_E} W_{a_E}^T o_E$

Unit 03 | QA model

2. Stanford Attentive Reader++ (DrQA)

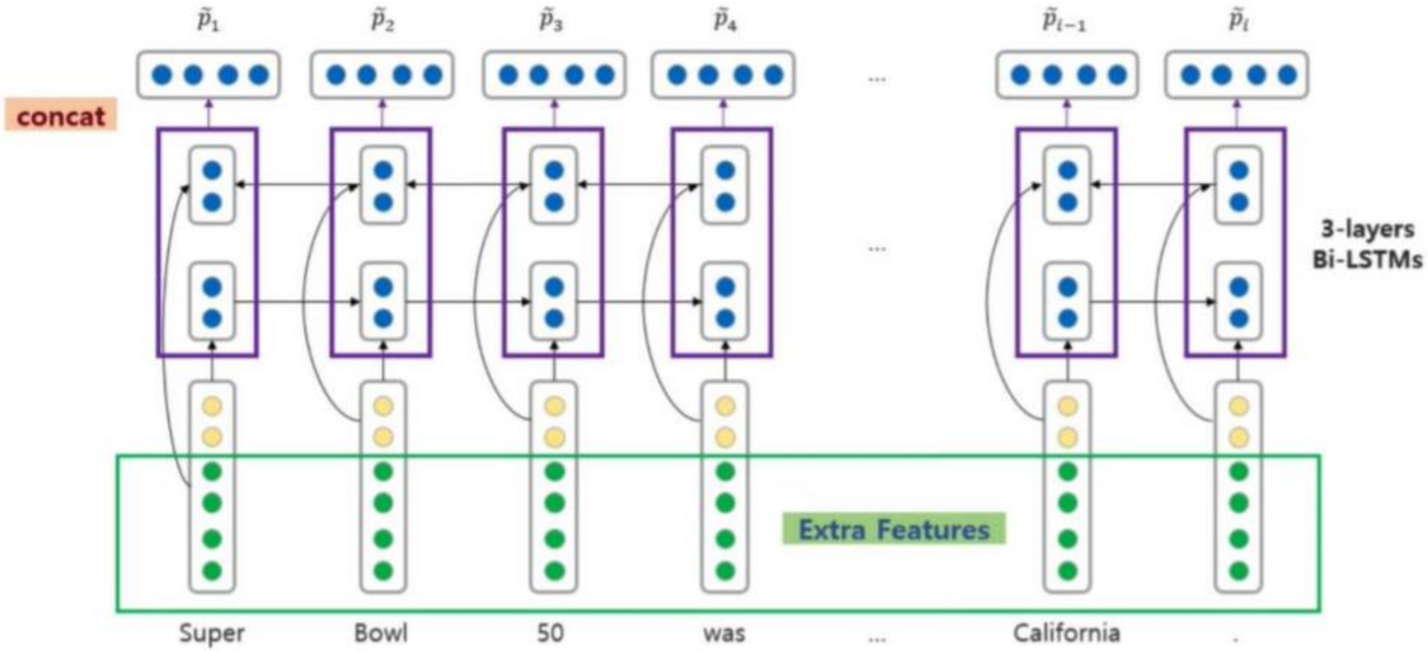
1. 3 layer BiLSTM
2. Bi-lstm의 hidden state를 포지션별로 concate 후, weighted sum을 하여 q 벡터 구성함.



Unit 03 | QA model

2. Stanford Attentive Reader++ (DrQA)

1. 3 layer BiLSTM
2. Glove 벡터만 사용한것이 아닌, 단어의 feature 도 같이 넣어줌.



Unit 03 | QA model

2. Stanford Attentive Reader++ (DrQA)

* 추가해준 feature

- Word embedding (GloVe 300d)
- Linguistic features: POS & NER tags, one-hot encoded
- Term frequency (unigram probability)
- Exact match: whether the word appears in the question
  - 3 binary features: exact, uncased, lemma
- Aligned question embedding (“car” vs “vehicle”)

Unit 03 | QA model

3. BiDAF: Bi-Directional Attention Flow for Machine Comprehensive

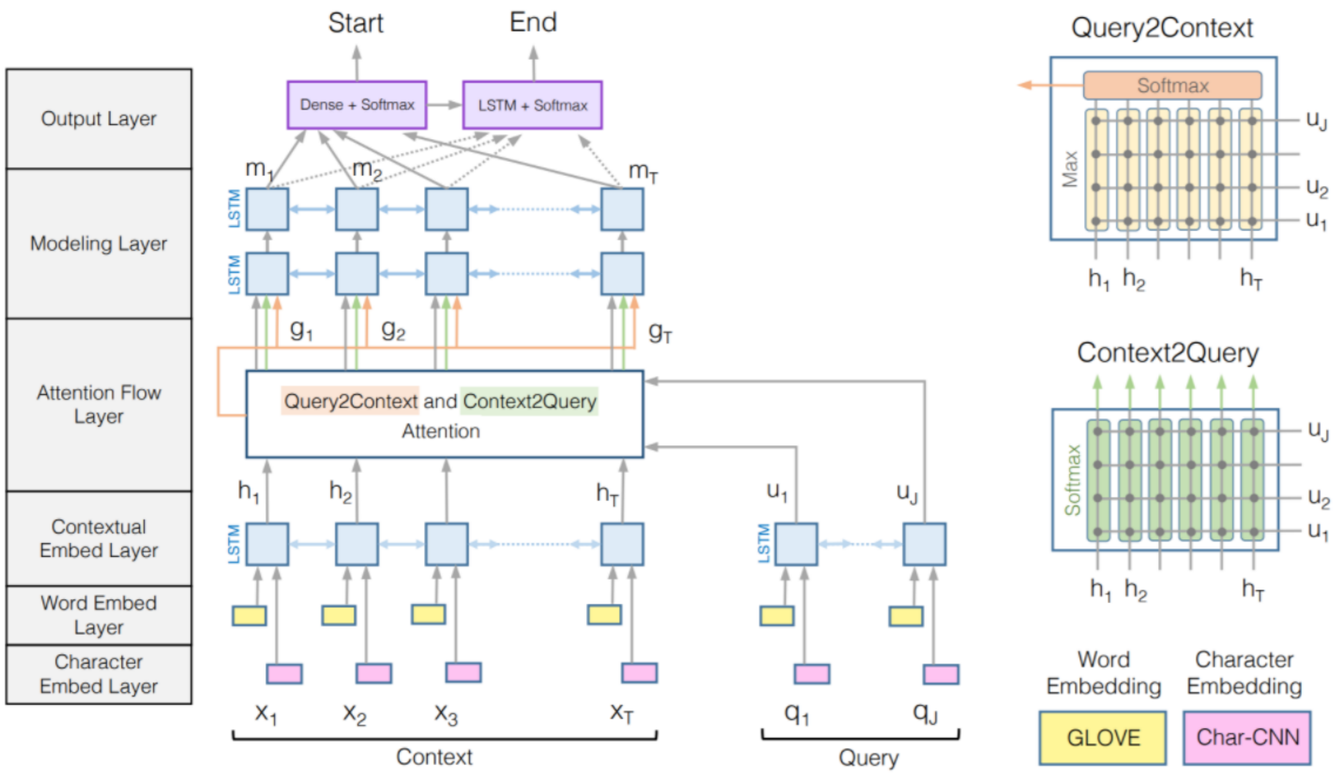


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Question에서 paragraph로 한방향으로만

진행되는 Stanford Attentive Reader와 달리,

attention이 양방향으로 적용됨 (question ↔ paragraph)

Unit 03 | QA model

3. BiDAF: Bi-Directional Attention Flow for Machine Comprehensive

1. Character Embedding Layer: CharCNN을 사용하여 각단어를 vector space에 mapping 함.
2. Word Embedding Layer: pre-trained word embedding 모델을 사용하여 각 단어를 vector space에mapping 함
3. Contextual Embedding Layer: Target word의 주변단어들을 통해 embedding 을 정제함. 처음 3개의 layer에 대해서는 query 와 context에 모두 적용됨.
4. Attention Flow Layer: Query 와 Context 를 쌍으로 묶어서 Attention 을 학습하게 된다.
5. Modeling Layer: RNN을 통해 Context를 탐색함.
6. Output Layer: Query에 대해 답을 생성함.

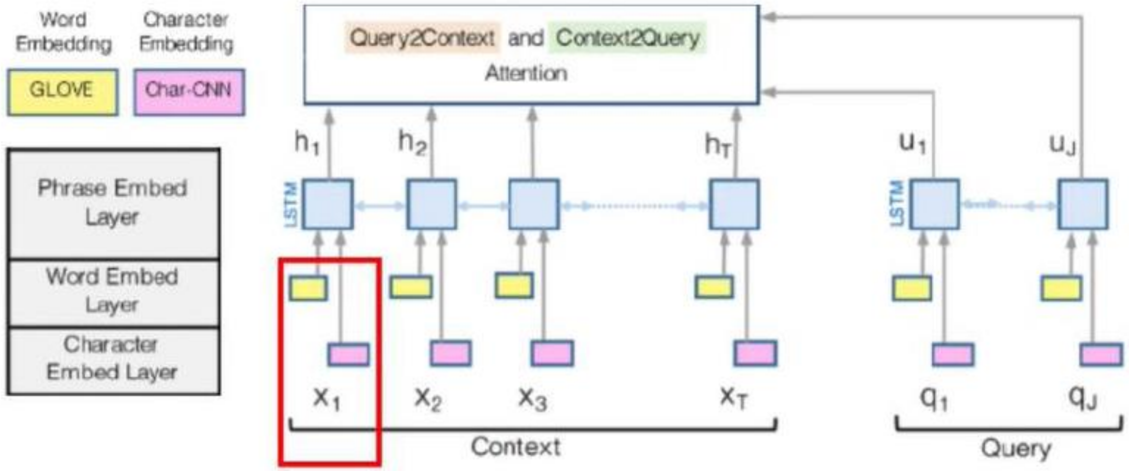
https://hwkim94.github.io/deeplearning/rnn/lstm/attention/nlp/paperreview/2018/02/23/BiDAF1.html

Unit 03 | QA model

3. BiDAF: Bi-Directional Attention Flow for Machine Comprehensive

1. Character Embedding Layer: CharCNN을 사용하여 각단어를 vector space에 mapping 함.
2. Word Embedding Layer: pre-trained word embedding 모델을 사용하여 각 단어를 vector space에mapping 함.
3. Contextual Embedding Layer: Target word의 주변단어들을 통해 embedding 을 정제함. 처음 3개의 layer에 대해서는 query 와 context에 모두 적용됨.

a. Wordvector 와 character vector를 concat함.
b. Bi-LSTM에 넣기 전, two-layer highway-network를 거쳐 d-dimension을 가진vector로 만들어 줌.
c. Bi-LSTM에 넣어 2d vector 생성함.



3. BiDAF: Bi-Directional Attention Flow for Machine Comprehensive

4. Attention Flow Layer: Query 와 Context 를 쌍으로 묶어서 Attention 을 학습하게 된다.

Query to Context : Context 의 어떤 정보다 Query 와 관련이 있는지

Context to Query : Query 의 어떤 정보가 Context 와 관련이 있는지

ex) Context : 사자가 동물원에 있다.

ex) Query : 여기엔 무슨 동물이 있나?

➔ 위의 문장에서 '동물원~여기', '사자~동물'을 연결시키는 역할

이전과는 다르게 query 와 Context 를 single feature vector로 요약하지 않고, query 와 context 를 연결시켰다.

Unit 03 | QA model

3. BiDAF: Bi-Directional Attention Flow for Machine Comprehensive

4. Attention Flow Layer: Query 와 Context 를 쌍으로 묶어서 Attention 을 학습하게 된다.

a. 양방향 attention을 가능하게 하기 위해 shared matrix인 S 사용함.

$Stj = \alpha(H{:}t, U{:}j)$ -> similarity matrix

→ $H{:}t$= context의 t번째 word vector, $U{:}j$ = query의 j번째 word vector

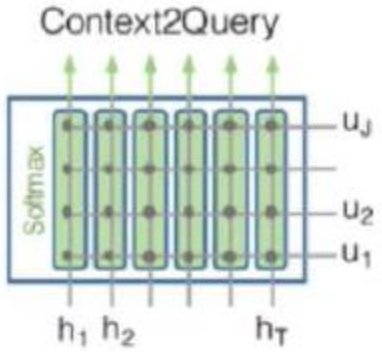
→ α = similarity를 encoding하는 역할

Stj 는 t-th context word와 j-th query word의 similarity 를 의미

Unit 03 | QA model

3. BiDAF: Bi-Directional Attention Flow for Machine Comprehensive

4. Attention Flow Layer: Query 와 Context 를 쌍으로 묶어서 Attention 을 학습하게 된다.



b. Context2Query Attention

$a_t = softmax(St:) \in RJ$

→ St: = t-th context word와 query word들의 similarity를 의미

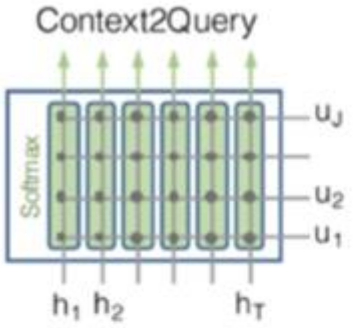
→ similarity Stj에 $softmax$를 취해줬으므로 유사성이 큰 weight만 값이 커지게 됨

$\widetilde{U}_{:t} = \sum a_{tj} U_{:j}$ (j-th query word에 t-th context와의 attention weight를 곱함)

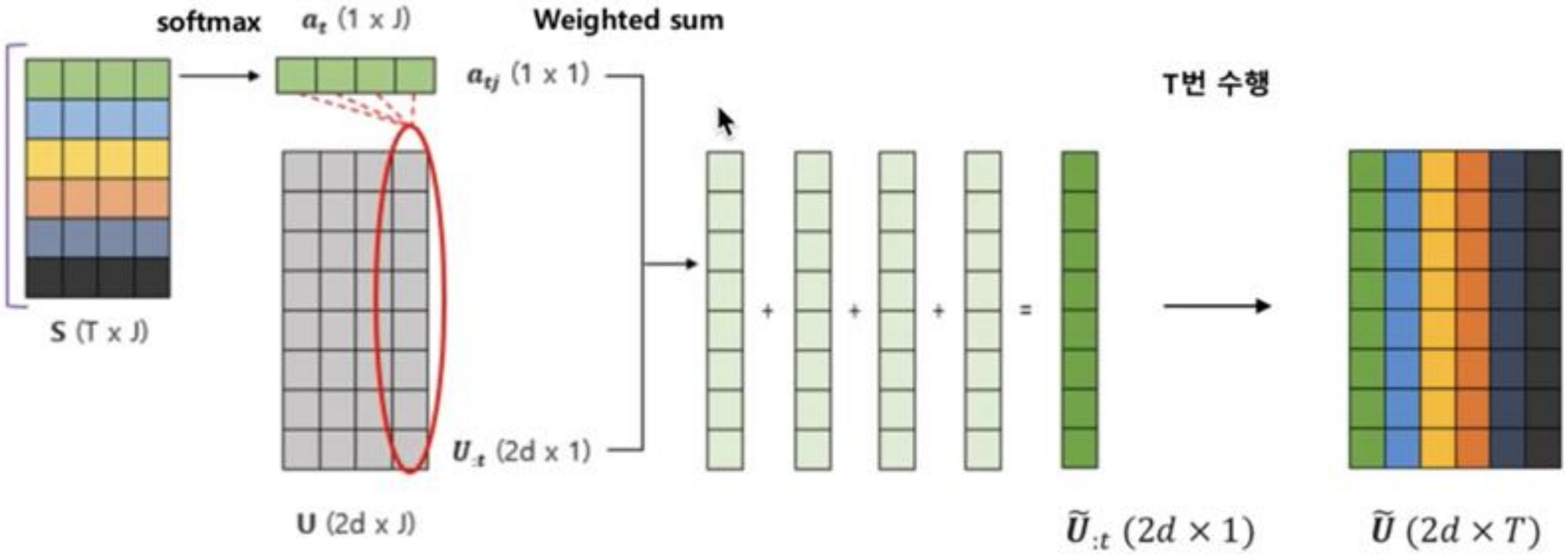
$H: t$= context의 t번째 word vector,

$U: j$ = query의 j번째 word vector

t-th context word 입장에서 유사성이 높은 query의 특성을 알게 됨



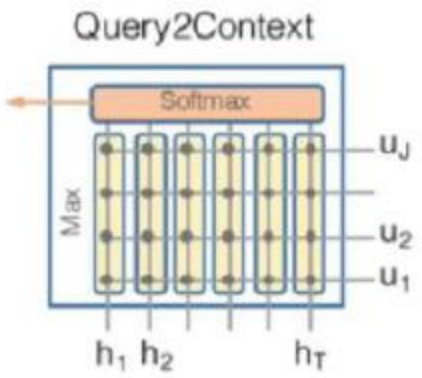
- $a_t = softmax(S_{t:}) \in \mathbb{R}^J$
- $\widetilde{U}_{:t} = \Sigma_j a_{tj} U_{:j}$



Unit 03 | QA model

3. BiDAF: Bi-Directional Attention Flow for Machine Comprehensive

4. Attention Flow Layer: Query 와 Context 를 쌍으로 묶어서 Attention 을 학습하게 된다.



$H{:}t$= context의 t번째 word vector,

$U{:}j$ = query의 j번째 word vector

c. QuerytoContext Attention

b = $softmax(maxcol(S))$

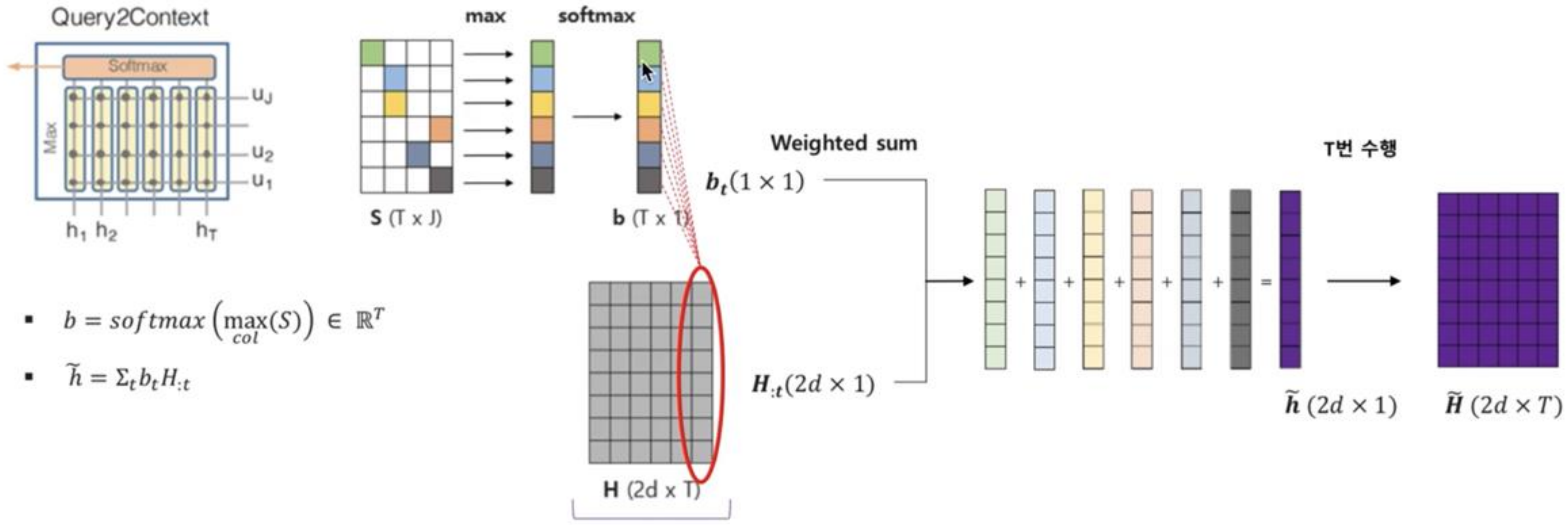
→ $maxcol$: 각 row에서 가장 큰 값을 뽑아냄

→ 각 context word마다 가장 query words중에서 similarity가 큰 항만 뽑히게 됨.

→ 여기에 softmax를 취해주므로 상대적으로 더 query와 상관 있는 context word만 남게 된다.

$\tilde{h} = \sum b_t H_{:t}$ (weight bt를 곱해서 값이 더해지면, query와 similarity가 큰 t-th context word만 값이 더해짐)

<u>similarity가 큰 context word만 살아남게 되므로 query 입장에서 중요한 word만 살아남게 됨</u>



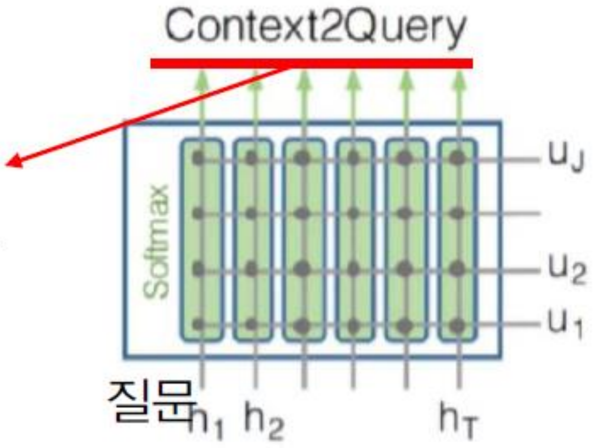
- $b = softmax\left(\max_{col}(S)\right) \in \mathbb{R}^T$
- $\tilde{h} = \Sigma_t b_t H_{:t}$

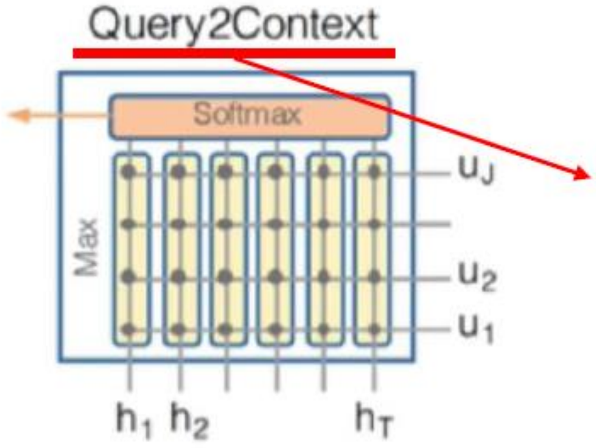
Unit 03 | QA model

3. BiDAF: Bi-Directional Attention Flow for Machine Comprehensive

4. Attention Flow Layer: Query 와 Context 를 쌍으로 묶어서 Attention 을 학습하게 된다.

질문에 있는 단어들 중에
어떤 paragraph의 단어들과
연관이 있는가
$\widetilde{U}$





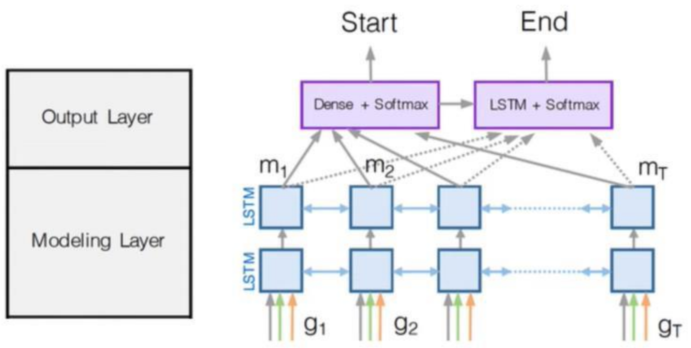
paragraph의 단어들 중
어떤 단어가 질문에 있는
단어와 가까운가
$\widetilde{H}$

Unit 03 | QA model

3. BiDAF: Bi-Directional Attention Flow for Machine Comprehensive

- Modeling Layer & Output Layer
  - Modeling Layer : 2-layer-bidirectional LSTMs
  - Output Layer :
    - ➢ $p^1 = softmax\left(w_{(p^1)}^{\mathrm{T}}\,[G; M]\right)$
    - ➢ $p^2 = softmax\left(w_{(p^2)}^{\mathrm{T}}[G; M^2]\right)$
  - Training
    - ➢ Average (sum of negative log probabilities)

$$L(\theta) = -\frac{1}{N}\sum_{i}^{N} \log\left(p_{y_i^1}^1\right) + \log\left(p_{y_i^2}^2\right)$$



$G{:}t = \beta(H{:}t, \tilde{U}{:}t, \tilde{H}{:}t)$

1.앞서구한 G를 modeling layer에서 2-layer bi-lstm을 거쳐 M을 만듦.

2. M과 G를 concat하여 linear transformation을 하고 softmax를 취함. → token 예측

최종적으로 sum of negative log probabilities 의 average를 최소화 하는 방향으로 학습함.

Unit 04 | 출처

- https://hwkim94.github.io/deeplearning/rnn/lstm/attention/nlp/paperreview/2018/02/23/BiDAF1.html
- https://www.youtube.com/watch?v=yIdF-17HwSk&list=PLoROMvodv4rOhcuXMZkNm7j3fVwBBY42z&index=10
- https://www.youtube.com/watch?v=7u6Ys7I0z2E&list=PLetSlH8YjIfVdobI2IkAQnNTb1Bt5Ji9U&index=8
- https://hwkim94.github.io/deeplearning/rnn/lstm/attention/nlp/paperreview/2018/02/23/BiDAF1.html
- https://www.quantumdl.com/entry/10%EC%A3%BC%EC%B0%A82-Bidirectional-Attention-Flow-for-Machine-Comprehension-BiDAF

Q & A

들어주셔서 감사합니다.