



# Question Answering

Week11\_발표자 : 김나현, 임세영

# 목차

## 01 QA motivation & history

- Motivation
- MCTest Reading Comprehension
- Turn-of-the Millennium Full NLP QA

## 02 SQuAD

- SQuAD
- SQuAD evaluation, v1.1
- SQuAD 2.0
- limitations

## 03 QA Model


- Stanford Attentive Reader Model
- Stanford Attentive Reader Model++
- BiDAF



# 01 QA motivation & history



# 01 QA motivation & history



전체

이미지

동영상

뉴스

쇼핑

더보기


도구

검색결과 약 11,400,000개 (0.70초)


도움말: 한국어 검색결과만 검색합니다. 환경설정에서 검색 언어를 지정할 수 있습니다.

아이유 / 첫 번째 앨범

## Growing Up




### 함께 찾은 검색어




Real+

2011년,  
아이유




Last Fantasy

2011년,  
아이유




Modern Times

2013년,  
아이유




Smash Hits

2015년,  
아이유




Palette

2017년,  
아이유



Modern Times - Epilogue

2013년,  
아이유



Smash Hits 2: The Stories Behind the Songs

2018년,  
아이유

# 01 QA motivation & history

## #1 Motivation

### < Two parts of question answering >

1. 질문에 대한 **정답이 있을 것 같은 문서들** 찾기 정보 검색

- traditional information retrieval (IR) / web search

2. **찾은 문서들 내에서** 정답 찾기 정보 추출

- Machine Reading Comprehension

# 01 QA motivation & history

## #2 MCTest Reading Comprehension

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

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?

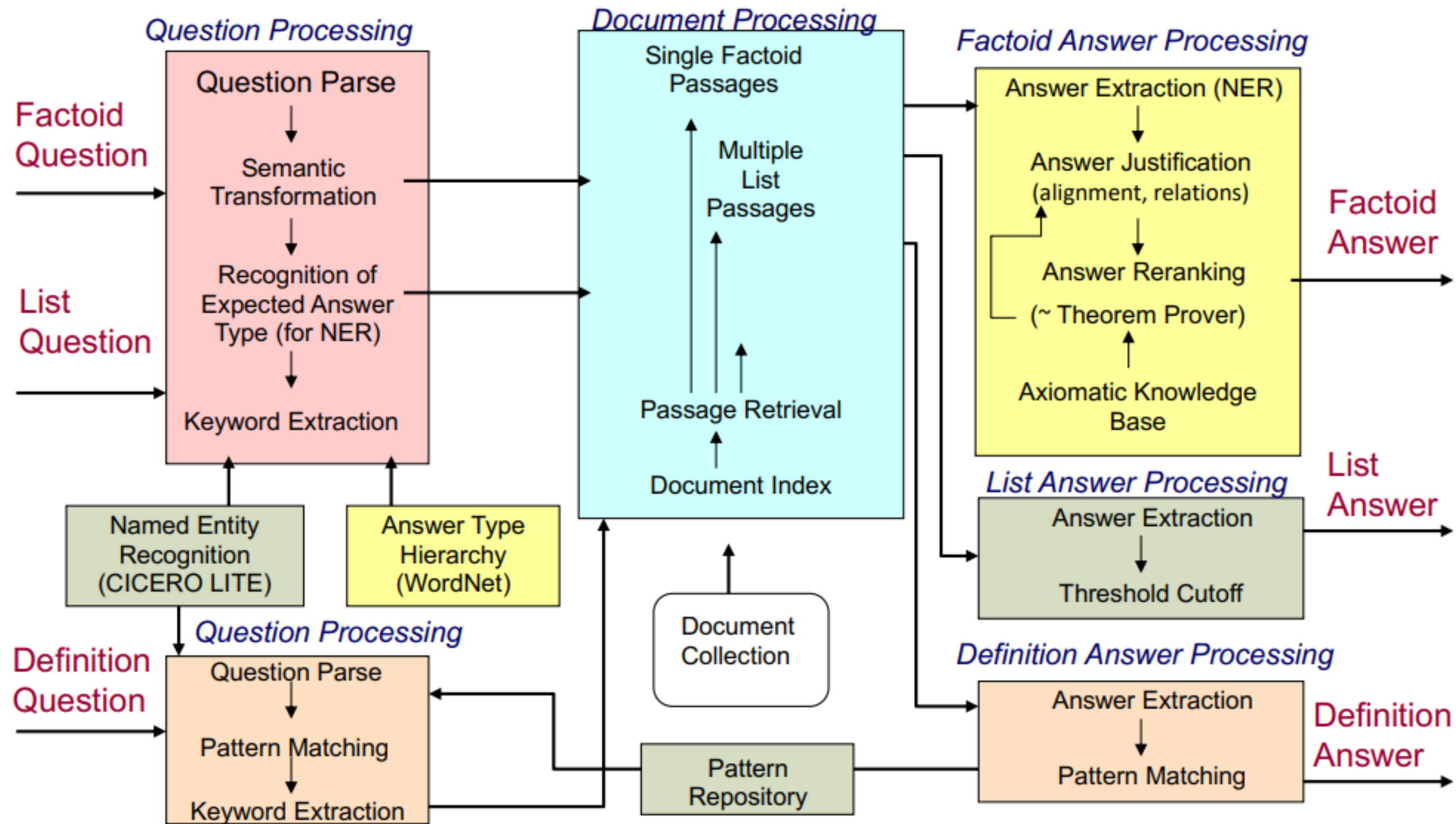
A

To visit some friends



# 01 QA motivation & history

## #3 Turn-of-the Millennium Full NLP QA



02 SQuAD





# 02 Stanford Question Answering Dataset (SQuAD)

## #1 SQuAD

**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
- 15 A.k.a. extractive question answering

# 02 Stanford Question Answering Dataset (SQuAD)

## #1 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

**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

# 02 Stanford Question Answering Dataset (SQuAD)

## #2 SQuAD evaluation, v1.1

Systems are scored on two metrics

### 1) Exact match

: 3개의 answers 중에 span이 존재하는지에 따라 0/1 accuracy 획득

### 2) F1

: 다른 답변에 대해 단어수준에서 일치

system과 gold answers 를 bag of words 로 취급

# 02 Stanford Question Answering Dataset (SQuAD)

## #2 SQuAD evaluation, v1.1

Score is (macro-)average of per-question F1 scores

- F1 measure is seen as more reliable and taken as primary
- It's less based on choosing exactly the same span that humans chose, which is susceptible to various effects, including line breaks
- Both metrics ignore punctuation and articles (a, an, the only)

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

$$\text{harmonic mean F1} = \frac{2PR}{P+R}$$

		실제 정답	
		True	False
분류 결과	True	True Positive	False Positive
	False	False Negative	True Negative

# 02 Stanford Question Answering Dataset (SQuAD)

## #2 SQuAD evaluation, v1.1

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) Google AI Language <a href="https://arxiv.org/abs/1810.04805">https://arxiv.org/abs/1810.04805</a>	87.433	93.160
2 Oct 05, 2018	BERT (single model) Google AI Language <a href="https://arxiv.org/abs/1810.04805">https://arxiv.org/abs/1810.04805</a>	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) Microsoft Research Asia	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) Microsoft Research Asia	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) Google Brain & CMU	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) Microsoft Research Asia	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) Google Brain & CMU	83.877	89.737
5 Sep 09, 2018	nlnet (single model) Microsoft Research Asia	83.468	90.133

# 02 Stanford Question Answering Dataset (SQuAD)

## #3 SQuAD 2.0

SQuAD 1.0의 결함 :

정확한 답이 context 내에 존재한다고 가정 -> 문단에서 답을 찾아야함 -> ranking task 가 됨  
정답이 정말 context에 존재하는지 X, 답에 근접해 보이는 span을 찾을 뿐



SQuAD 2.0:

기존의 응답 가능한 질문들에 응답 불가능한 질문들을 추가

- dev/test question: 반은 답이 있고 반은 답이 없음
- training data : 1/3 no answer

응답 불가능한 질문들은 응답 가능 질문들과 같은 본문 paragraphs 에서 제작됨  
질문 제작에 온라인의 다수 crowd worker 사용



# 02 Stanford Question Answering Dataset (SQuAD)

## #3 SQuAD 2.0 example

Genghis Khan united the Mongol and Turkic tribes of the steppes and became Great Khan in 1206. He and his successors expanded the Mongol empire across Asia. Under the reign of Genghis' third son, Ögedei Khan, the Mongols destroyed the weakened Jin dynasty in 1234, conquering most of northern China. Ögedei offered his nephew Kublai a position in Xingzhou, Hebei. Kublai was unable to read Chinese but had several Han Chinese teachers attached to him since his early years by his mother Sorghaghtani. He sought the counsel of Chinese Buddhist and Confucian advisers. Möngke Khan succeeded Ögedei's son, Güyük, as Great Khan in 1251. He

**When did Genghis Khan kill Great Khan?**

*Gold Answers:* <No Answer>

*Prediction:* 1234 [from Microsoft nlnet]

# 02 Stanford Question Answering Dataset (SQuAD)

## #3 SQuAD 2.0 example

The Yuan dynasty is considered both a successor to the Mongol Empire and an imperial Chinese dynasty. It was the khanate ruled by the successors of Möngke Khan after the division of the Mongol Empire. In official Chinese histories, the Yuan dynasty bore the Mandate of Heaven, following the Song dynasty and preceding the Ming dynasty. The dynasty was established by Kublai Khan, yet he placed his grandfather Genghis Khan on the imperial records as the official founder of the

**What dynasty came before the Yuan?**

*Gold Answers:* ① Song dynasty ② Mongol Empire  
③ the Song dynasty

*Prediction:* Ming dynasty [BERT (single model) (Google AI)]

# 02 Stanford Question Answering Dataset (SQuAD)

## #4 SQuAD limitations

### 1. Only span-based answers (no yes/no, counting, implicit why)

- passage에서 온 span만 answer로 예측할 수 있다.

ex : 중고등학교 영어 독해문제

### 2. Question were constructed looking at the passages

- passages 내에서만 정답을 찾으려 하는 질문 구성
- 여러 문서들을 비교하여 진짜 정답을 찾지 X
- 실제 question-answer 보다 lexical & syntactic 한 matching, 문법 구조를 갖고 있음.

### 3. 동일 지시어(coreference) 문제를 제외하고는 Multi-fact 문제, 문장 추론 문제가 거의 없음

Nevertheless, SQuAD is well-targeted, well-structured, clean dataset

- QA 문제를 푸는데 있어 가장 많이 사용되고, 경쟁하고 있는 데이터셋
- 실제 시스템을 개발하기 위해 유용한 starting point

# 02 Stanford Question Answering Dataset (SQuAD)

## #4 SQuAD limitations

1. Only span-based answers (no yes/no, counting, implicit why)

ex : 중고등학교 영어 독해문제

2. Question were constructed looking at the passages

- 실제 question-answer 보다 lexical & syntactic 한 matching, 문법 구조를 갖고 있음.

3. 동일 지시어(coreference) 문제를 제외하고는 Multi-fact 문제, 문장 추론 문제가 거의 없음

Nevertheless, SQuAD is well-targeted, well-structured, clean dataset

- QA 문제를 푸는데 있어 가장 많이 사용되고, 경쟁하고 있는 데이터셋
- 실제 시스템을 개발하기 위해 유용한 starting point

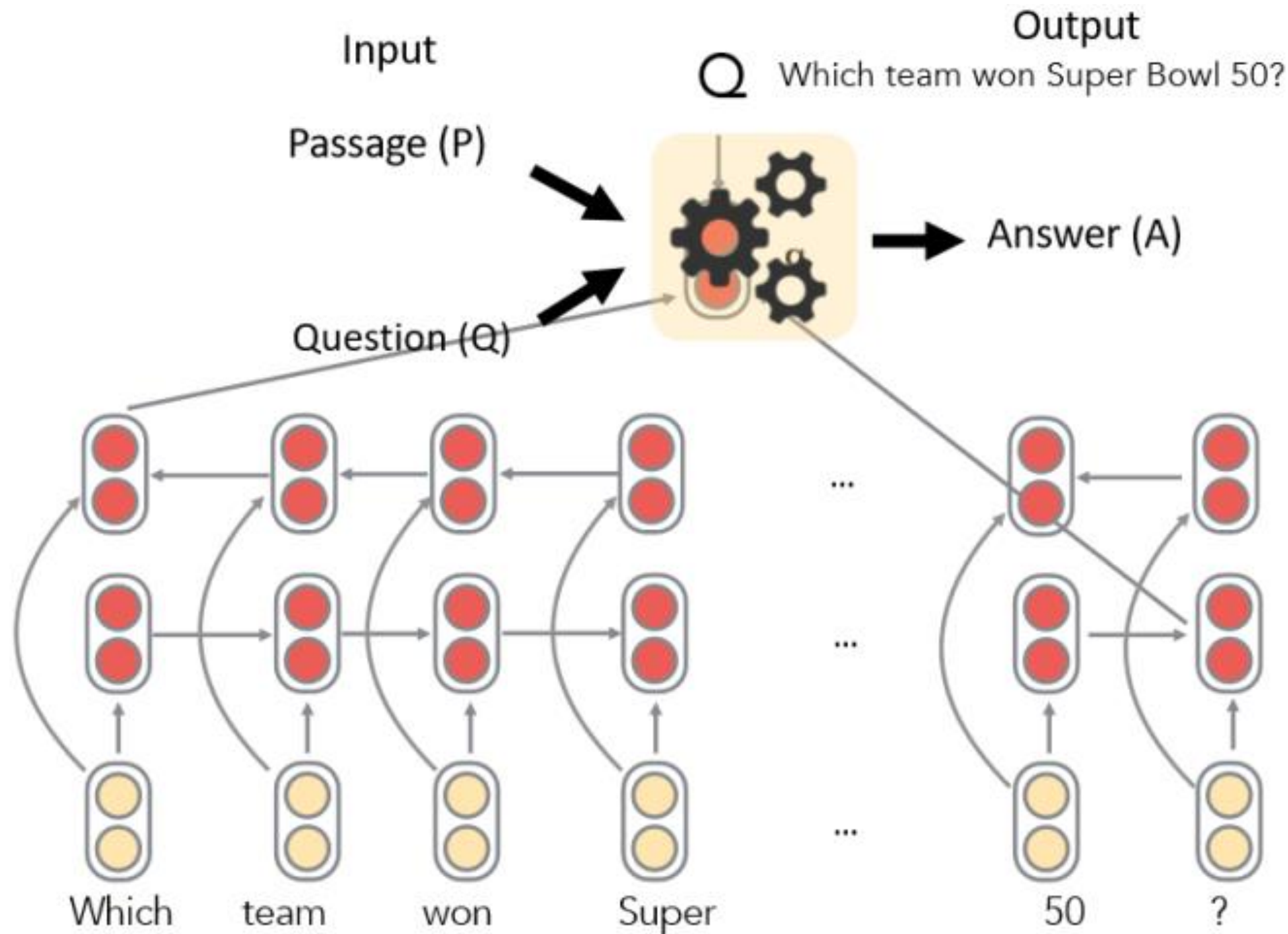
# QA Model





# The Stanford Attentive Reader

: 독해력 및 QA에 매우 성공적이고, 최소화된 아키텍처  
1-layer-bidirectional-LSTMs + GloVe300d

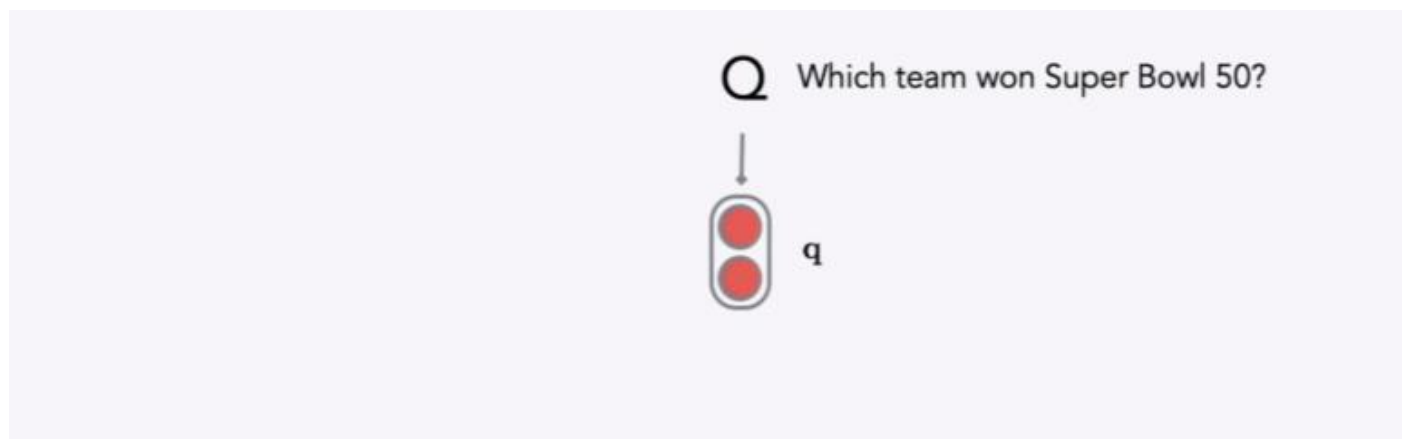




# The Stanford Attentive Reader

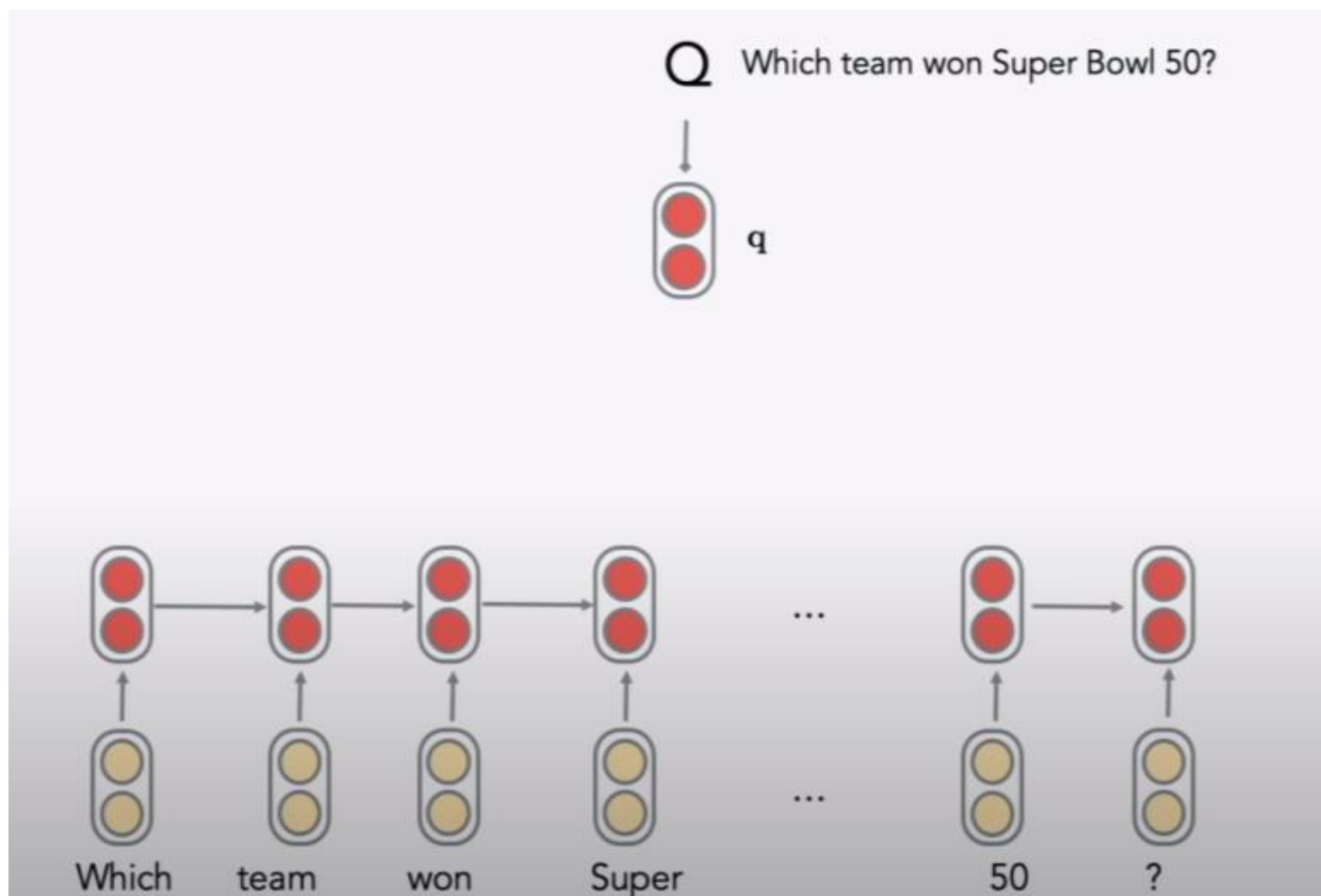
-Question Encoding

## 1) 벡터 q로 질문 표현

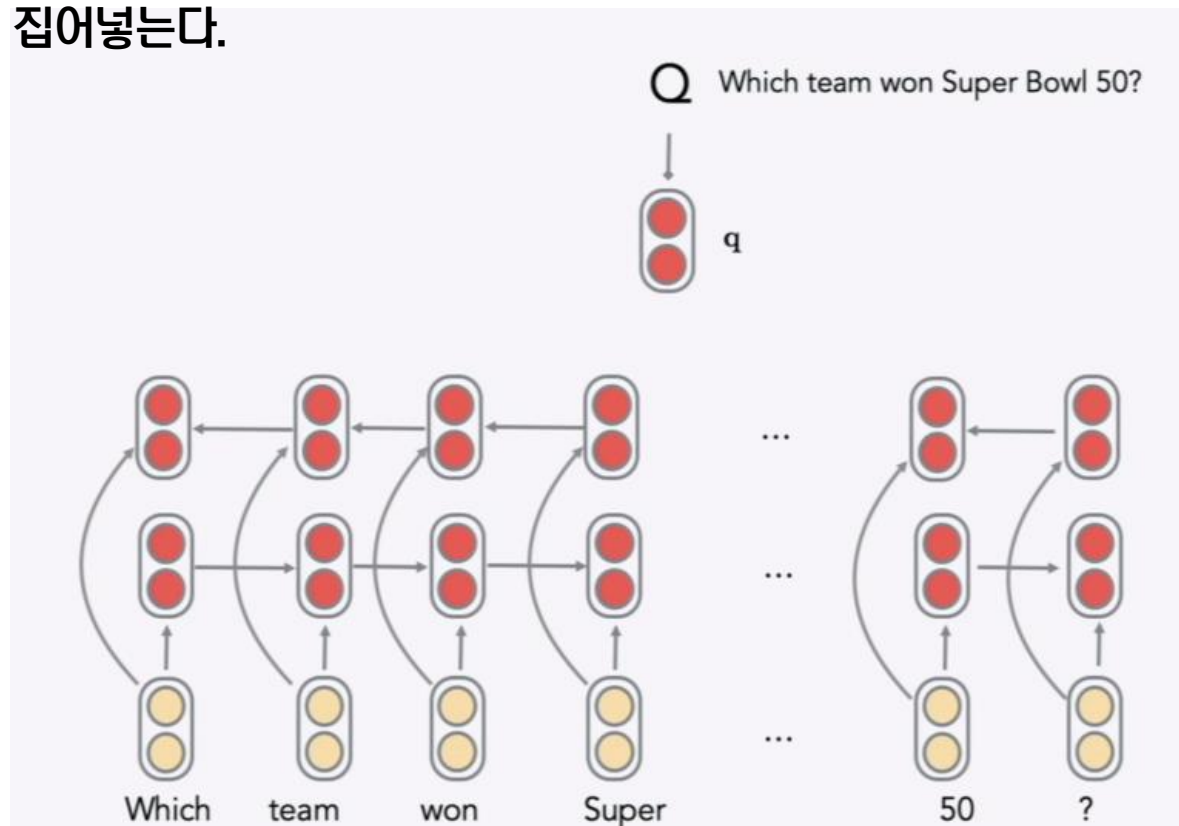


## 2) question의 각 단어 word embedding

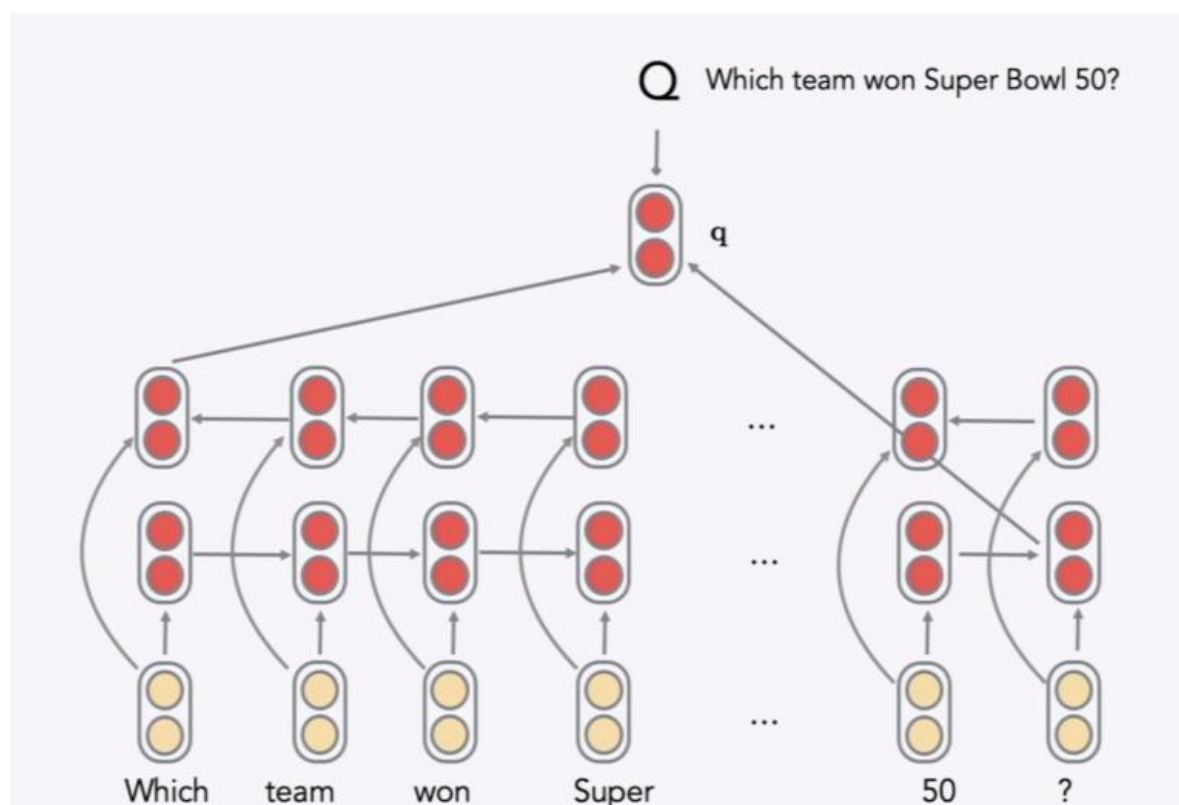
- 사전에 학습된 glove 300 dimension 이용



## 3) 1-layer-bidirectional-LSTMs model에 집어넣는다.



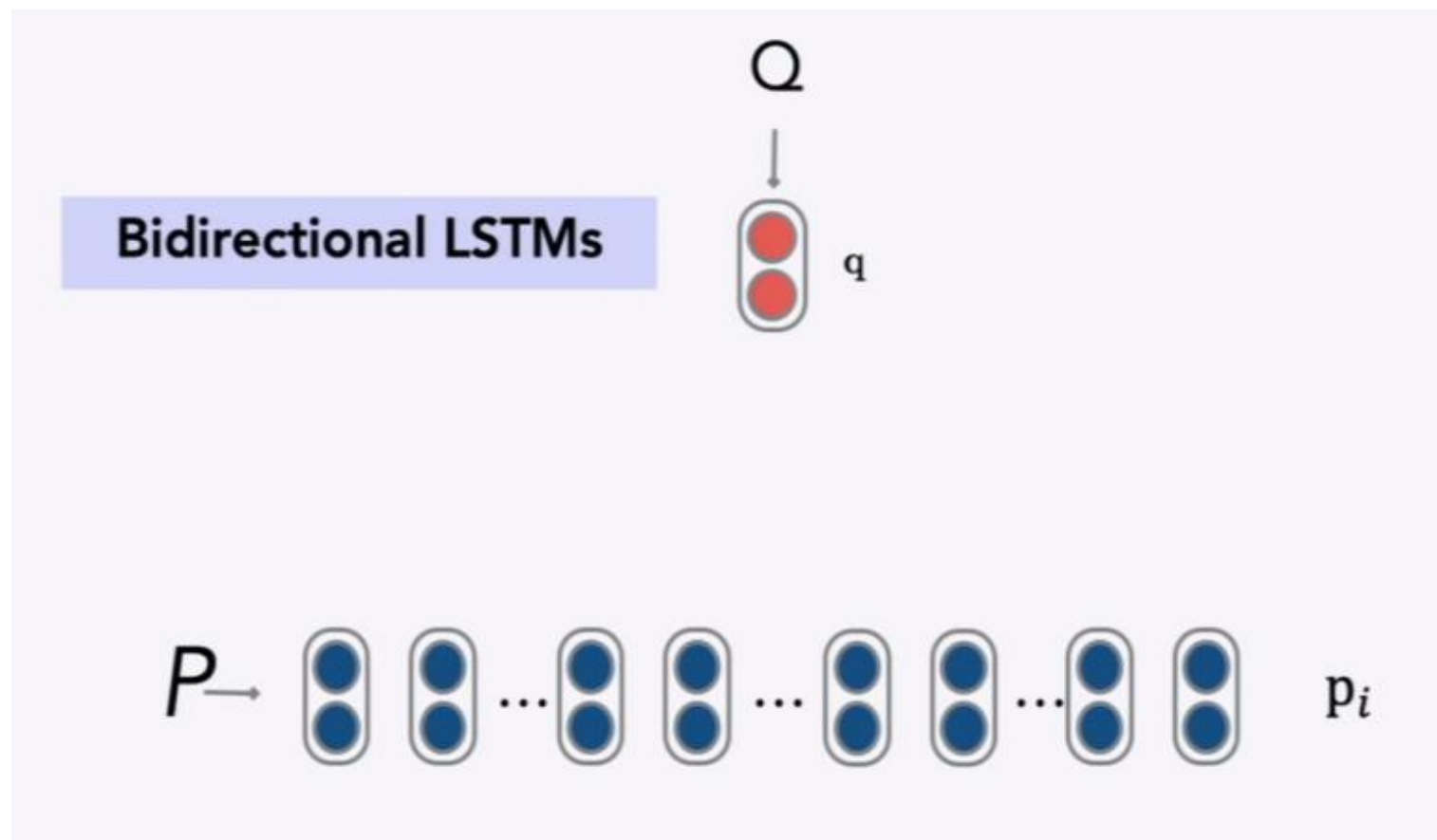
## 4) 각 방향의 마지막 hidden state 를 concat



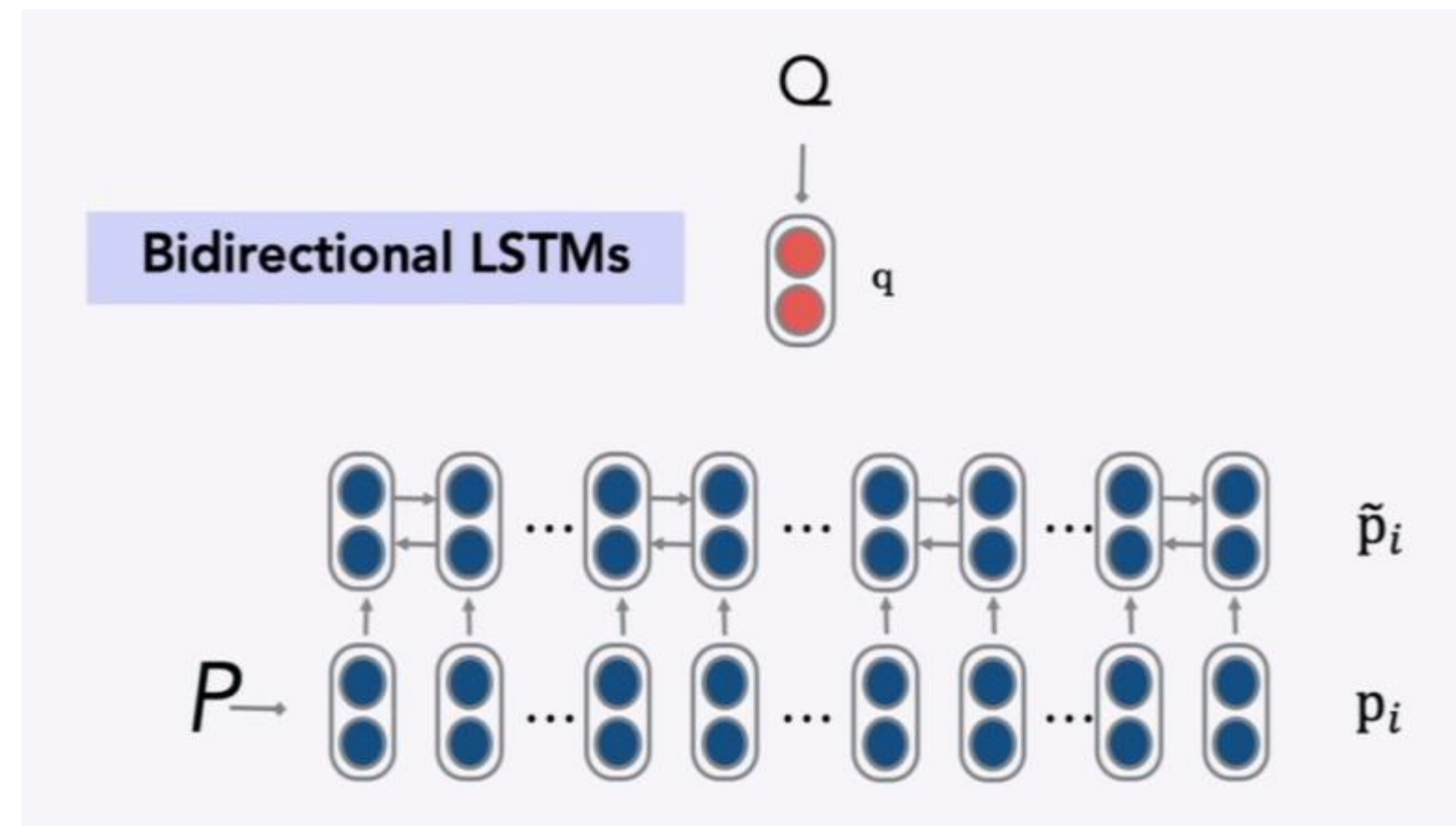
# The Stanford Attentive Reader

-Passage Encoding

1) passage 내 모든 단어 word embedding

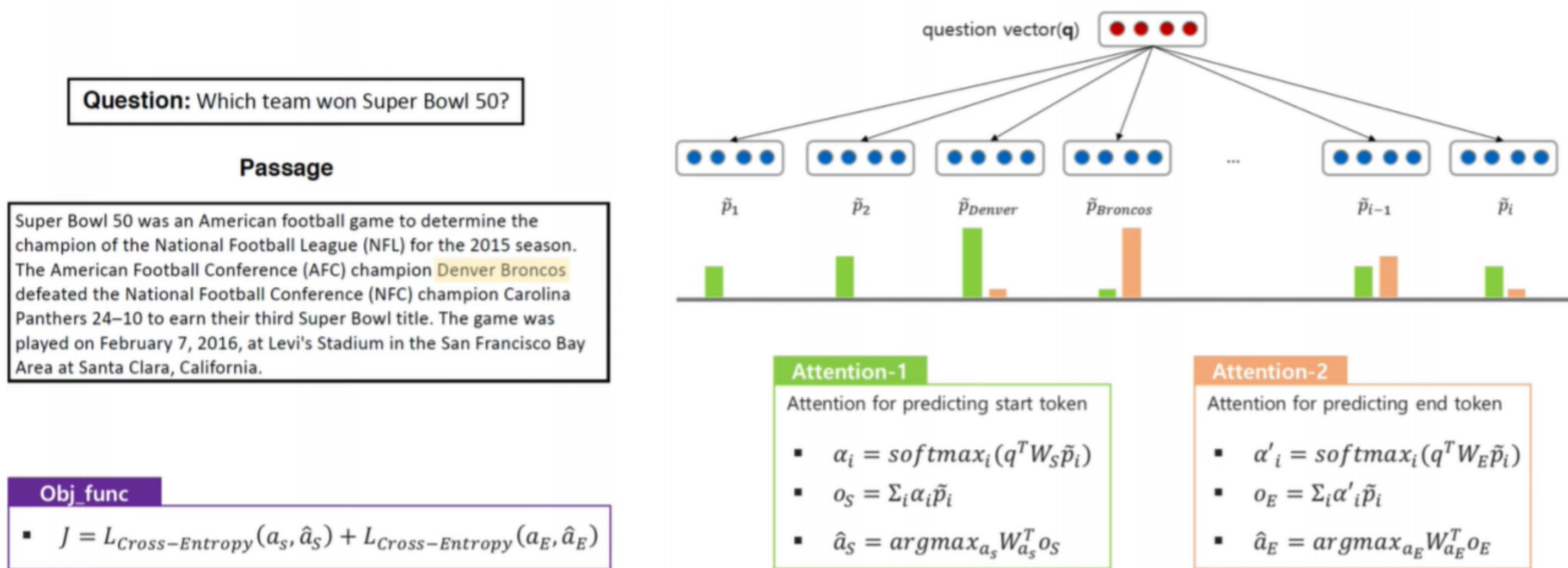


2) 양방향 LSTM 실행  
(passage에서 답을 찾음)



# The Stanford Attentive Reader

-Attention Encoding



알파 i: text내 모든 단어의 확률분포

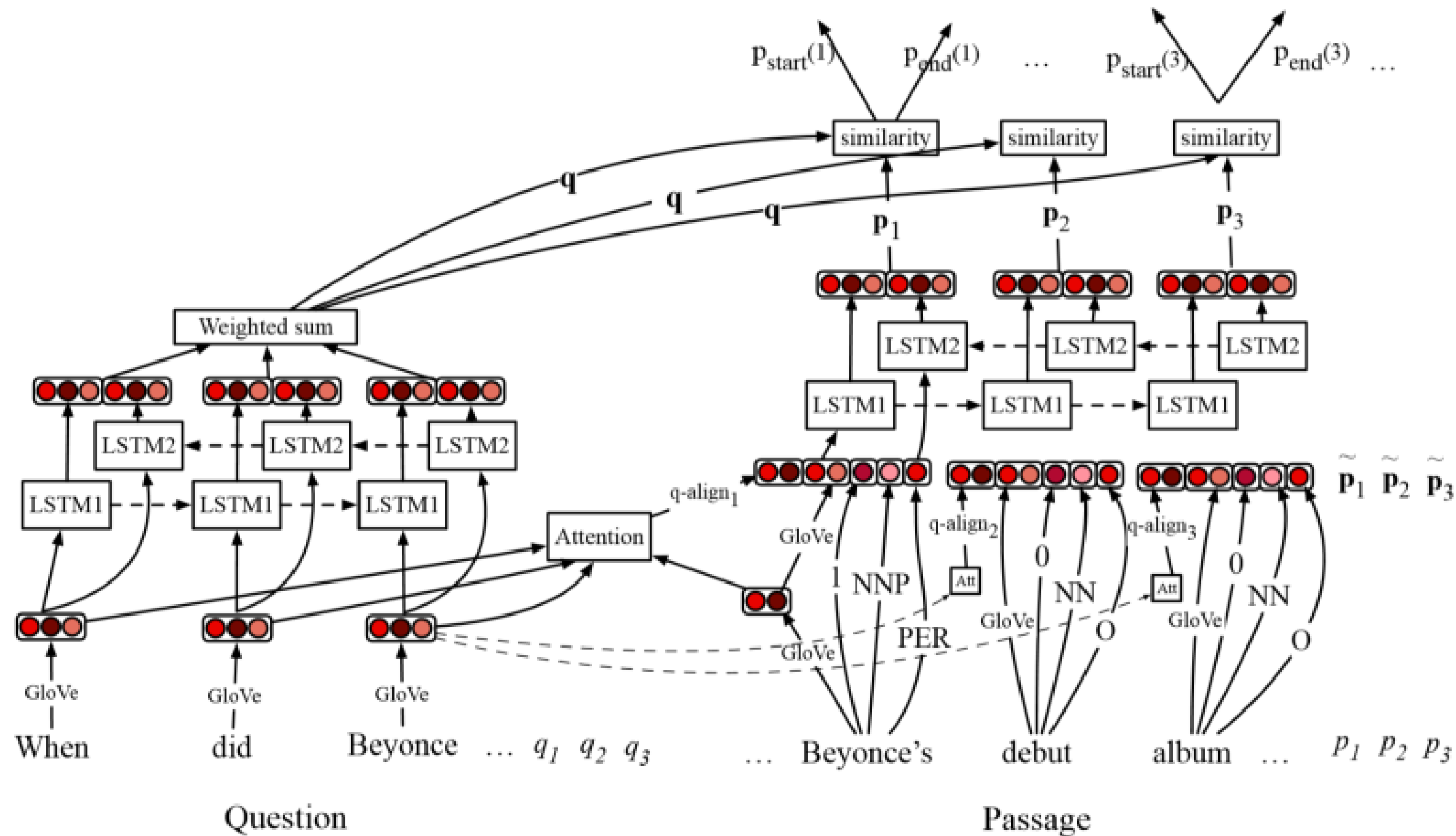
Os: 알파i와 passage vector pi를 곱한뒤 전부 더한 output vector

→ 다시한번 linear transformation 시켜준 start token 에 대해 예측 수행

최종 목적함수 J = start token과 end token 의 loss 값의 합  
→ J로 모델 학습

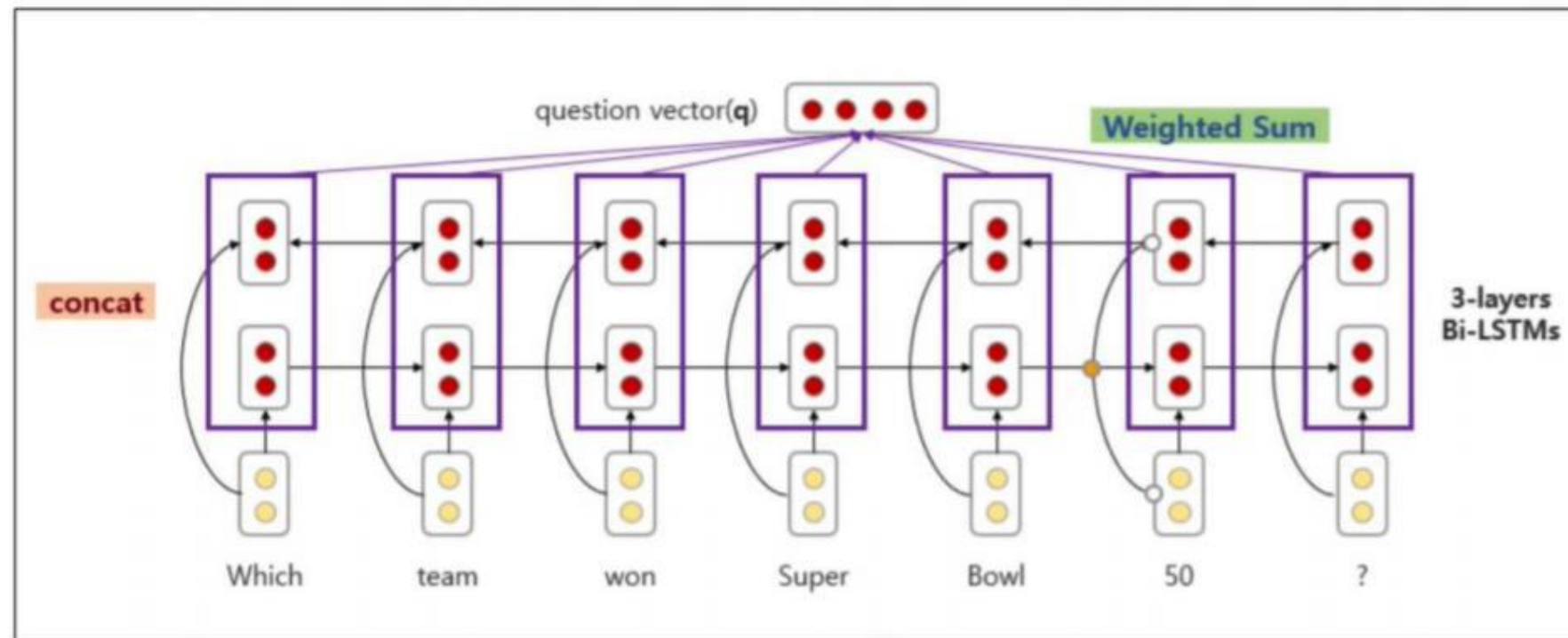
# Stanford Attentive Reader++

마찬가지로 question, passage, attention encoding 로 구성  
1 layer → 3 layer



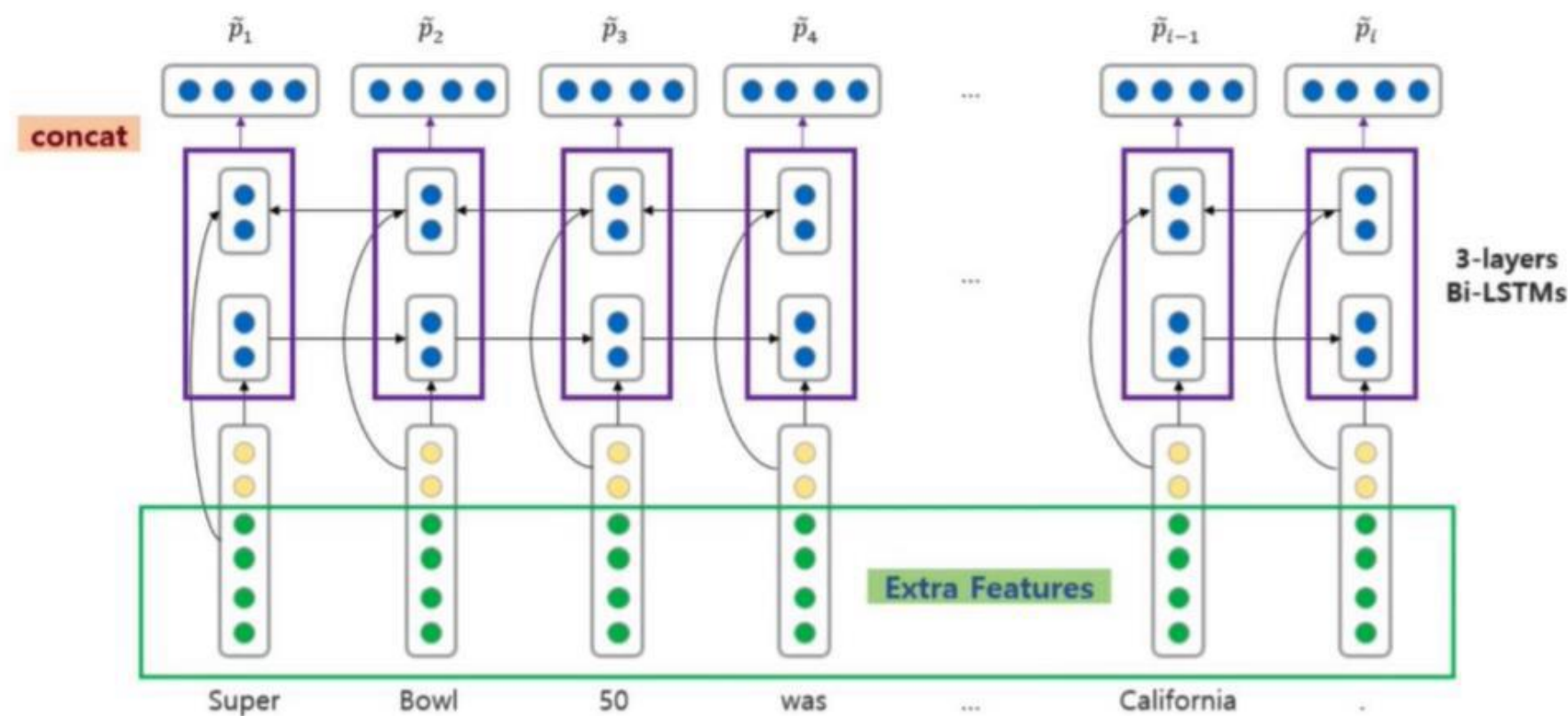


# Stanford Attentive Reader++



## Question Encoding

- Stanford Attentive Reader++에서는 one layer BiLSTM 이 아닌 3 layer BiLSTM을 사용
- BiLSTM state를 포지션별로 concat 후 weighted sum을 하여 구성



## Passage Encoding

- extra feature 추가

# Stanford Attentive Reader++

- $\mathbf{p}_i$ : Vector representation of each token in passage  
Made from concatenation of
  - 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”)

$$f_{align}(p_i) = \sum_j a_{i,j} \mathbf{E}(q_j) \quad q_{i,j} = \frac{\exp(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q_j)))}{\sum_{j'} \exp(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q'_j)))}$$

Where  $\alpha$  is a simple one layer FFNN



# BiDAF

<https://www.quantumdl.com/entry/10%EC%A3%BC%EC%B0%A82-Bidirectional-Attention-Flow-for-Machine-Comprehension-BiDAF>

## Character Embedding Layer

: CharCNN으로 각 단어를 vector space에 mapping

## Word Embedding Layer

: pre-trained word embedding 모델을 사용하여 각 단어를 vector space에 mapping

## Contextual Embedding Layer

: Target word의 주변 단어들을 통해 embedding을

## Attention Flow Layer

: Context에 대해 Query-aware feature vector를 만들기 위해 Query와 Context를 쌍으로 묶어 Attention을 학습.

## Modeling Layer

: RNN을 통해 Context를 탐색.

## Output Layer

: Query에 대해 답을 생성.

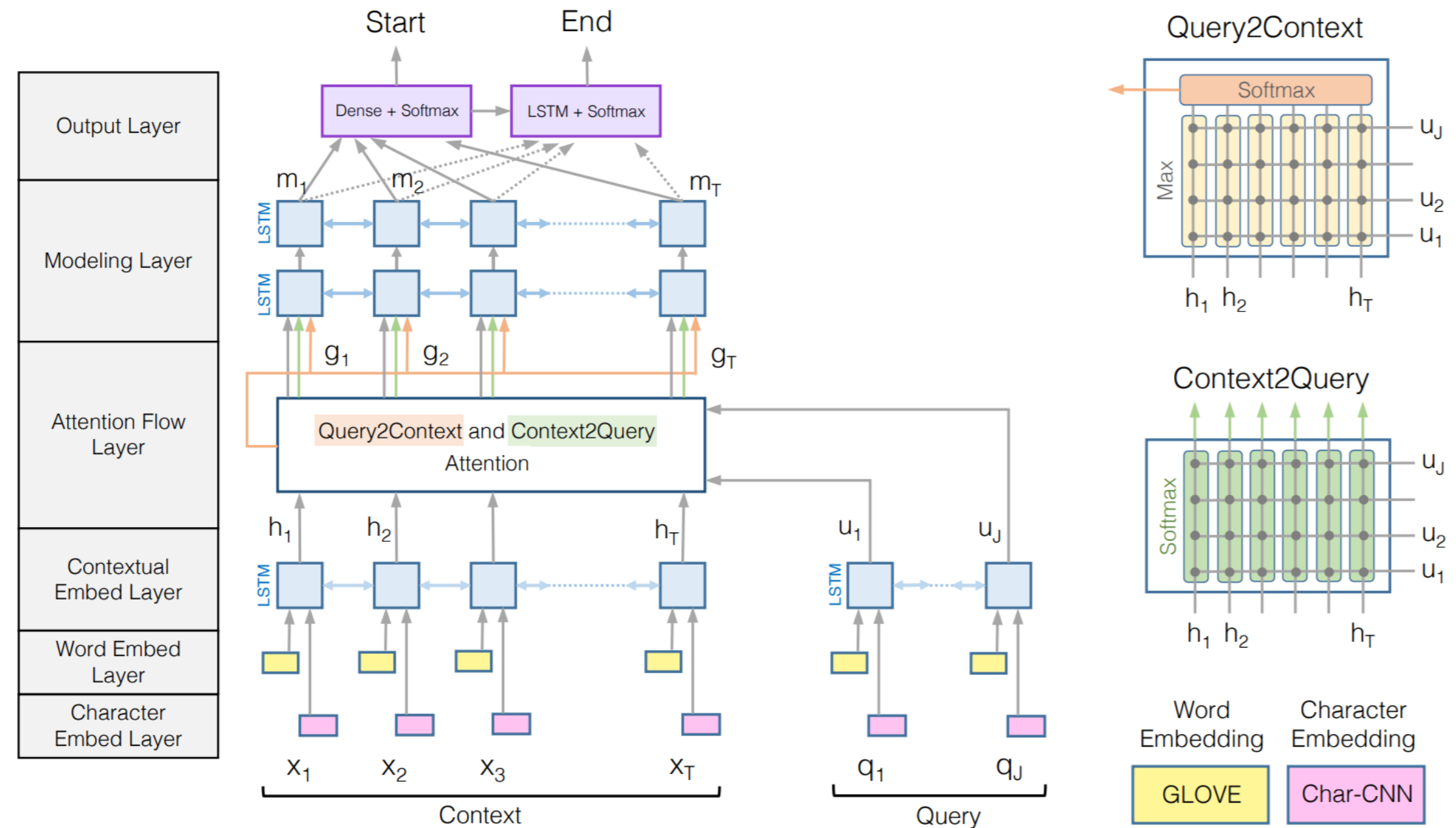


Figure 1: BiDirectional Attention Flow Model (best viewed in color)

# BiDAF

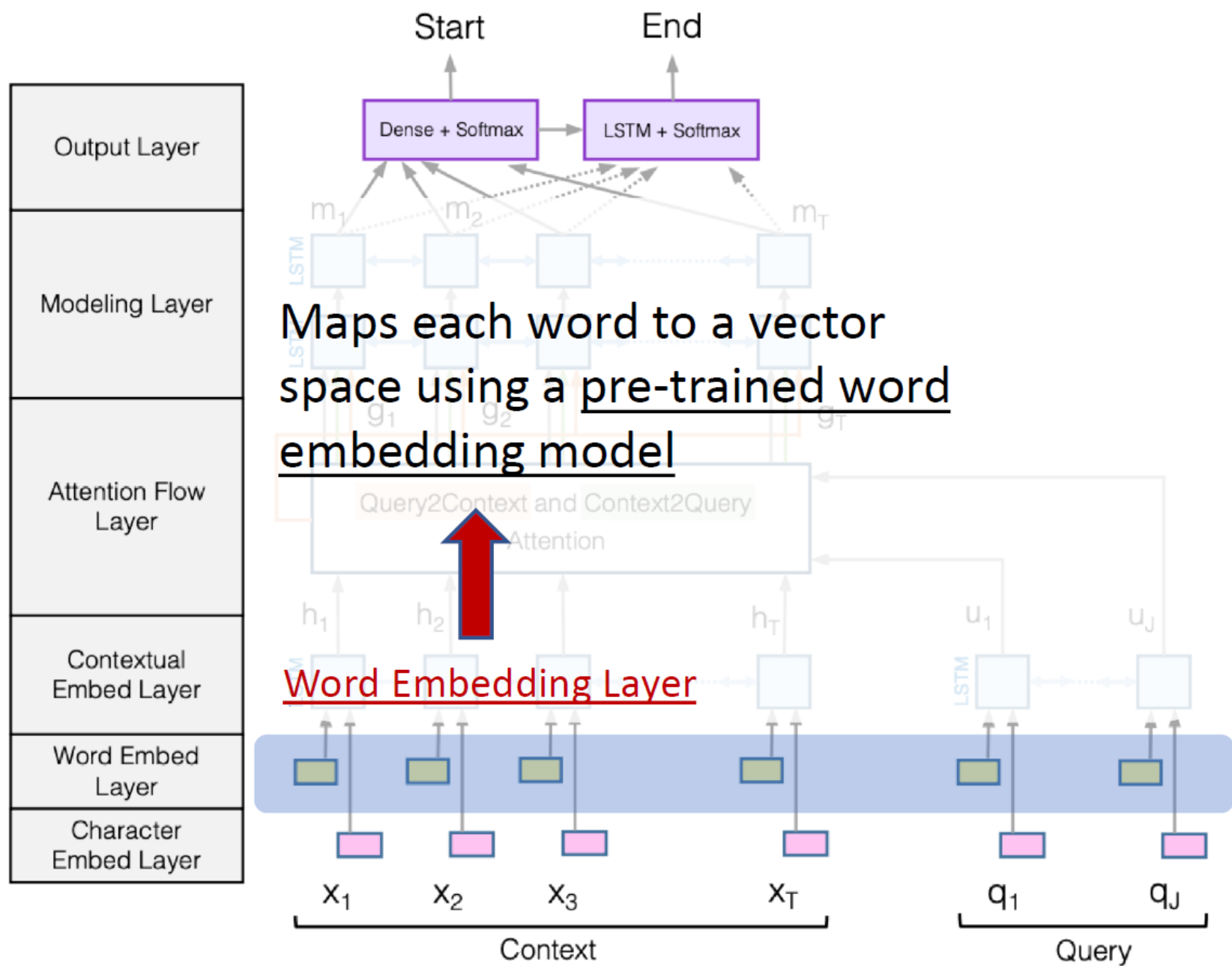


Figure 1: BiDirectional Attention Flow Model (best viewed in

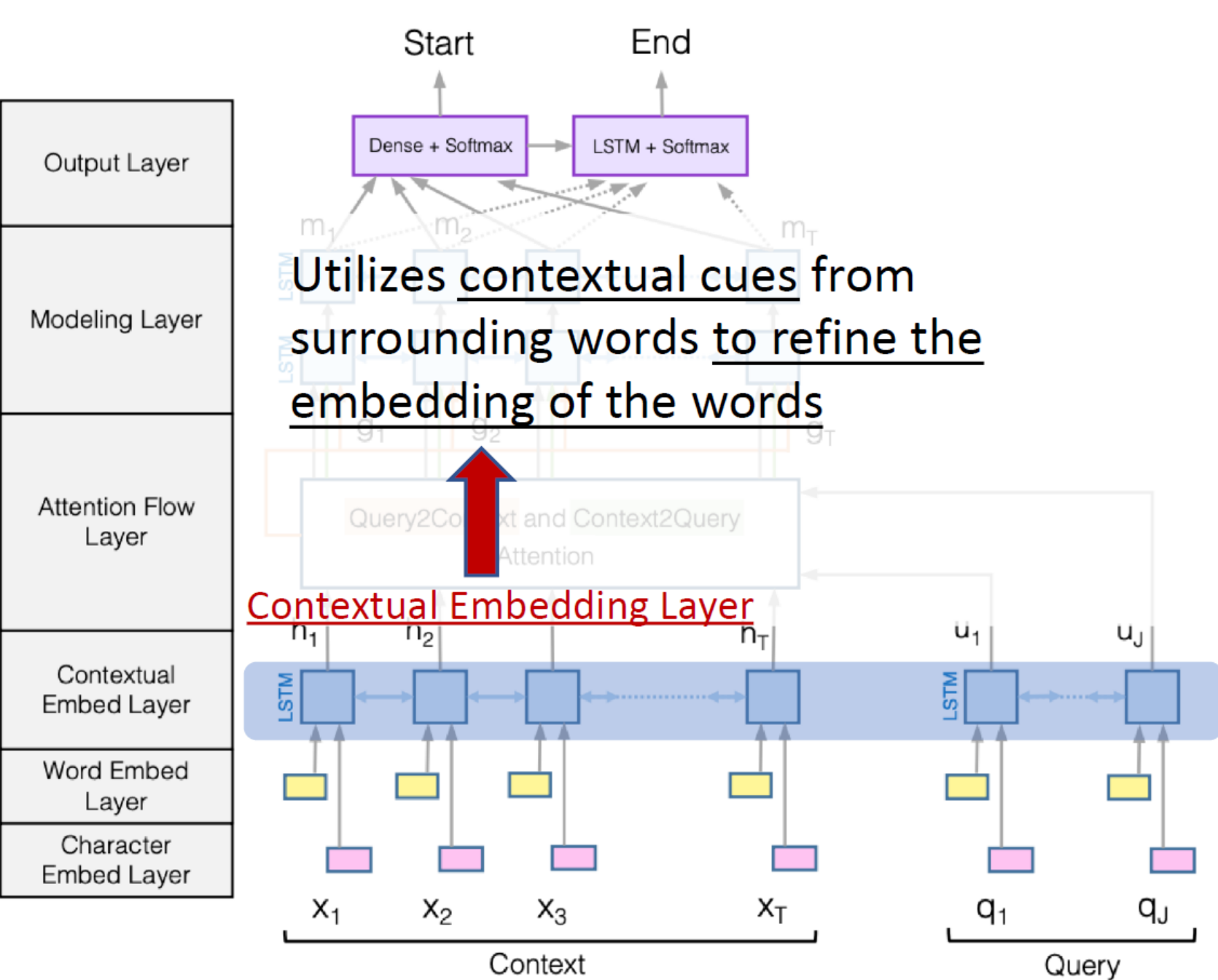
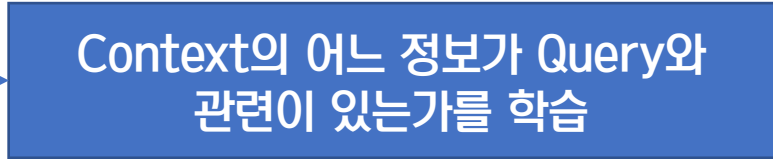


Figure 1: BiDirectional Attention Flow Model (best viewed in

EWCHA  
EUROPEAN

하나는 Query의 어느 정보가 Context와 관련이 있는가를 학습

Query와 Context에 대한 Attention이 양방향으로 발생  
→ Bidirectional Attention Flow(BiDAF)

Query와 Context간의 유사도를 파악하기  
 위해 **Similarity Matrix**라는 것도 사용한다.  
 ( t-th Context Word와 j-th Query Word간의  
 Similarity를 학습하게 된다. )

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

이전에 사용하는 Attention 방법들과는 다르게  
Attention을 Single vector(혹은 Fixed-sized  
Vector)로 요약하는 용도로 사용 X

→ 요약하면서 생기는 정보 손실에 대한 문제가  
여기서는 일어나지 않는다는 점이 특징이다.

# BiDAF

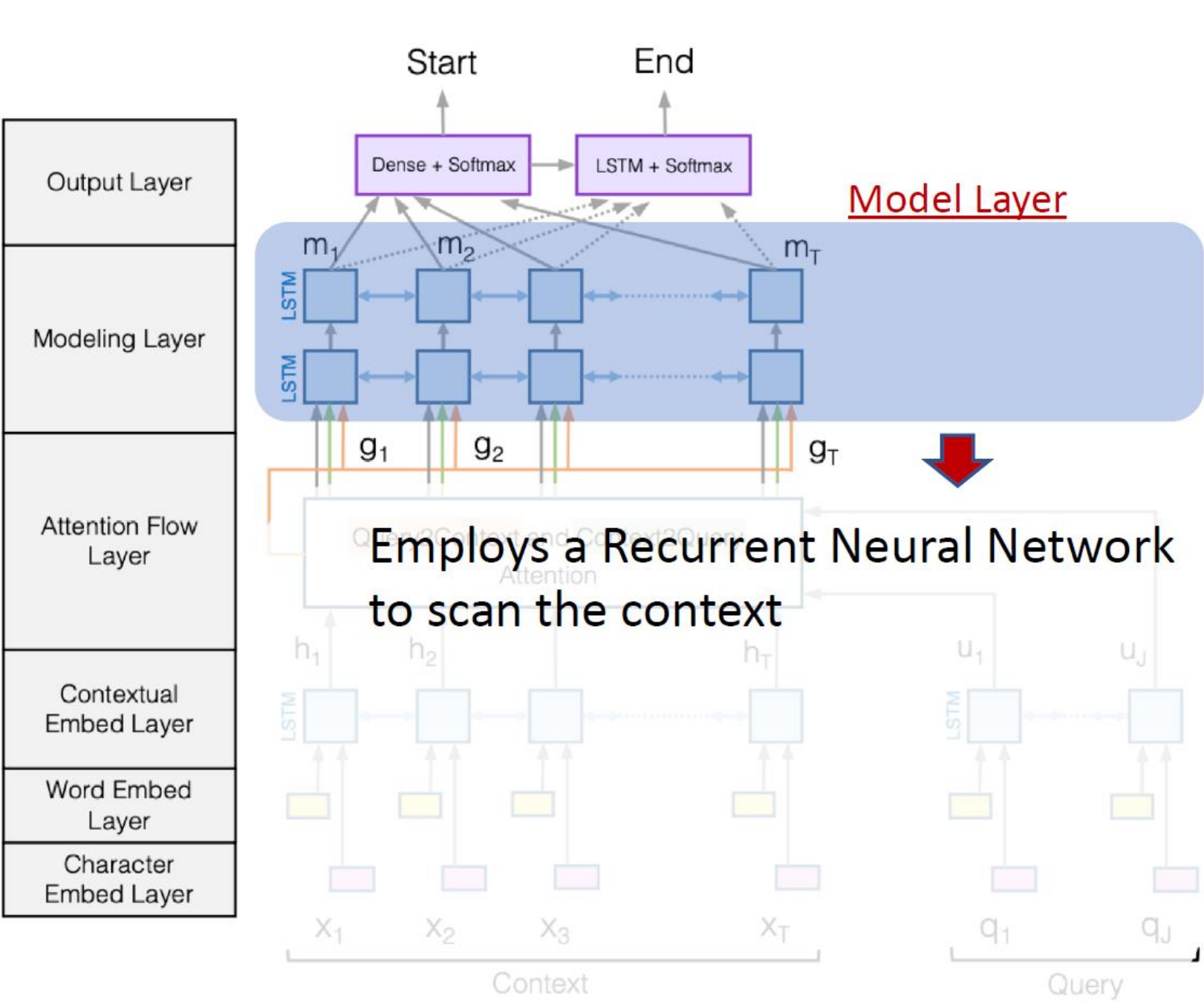


Figure 1: BiDirectional Attention Flow Model (best viewed in

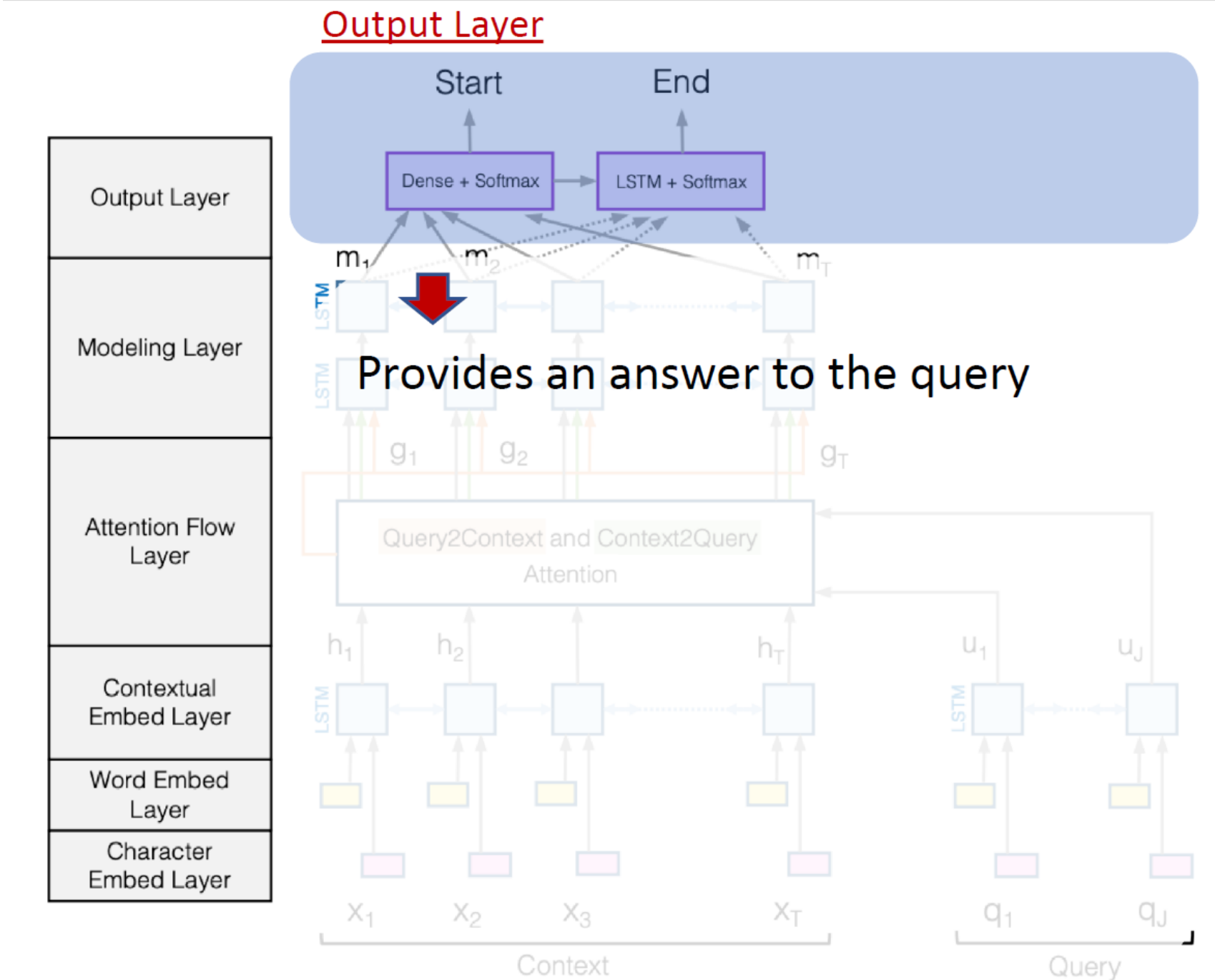


Figure 1: BiDirectional Attention Flow Model (best viewed in

# THANK YOU

