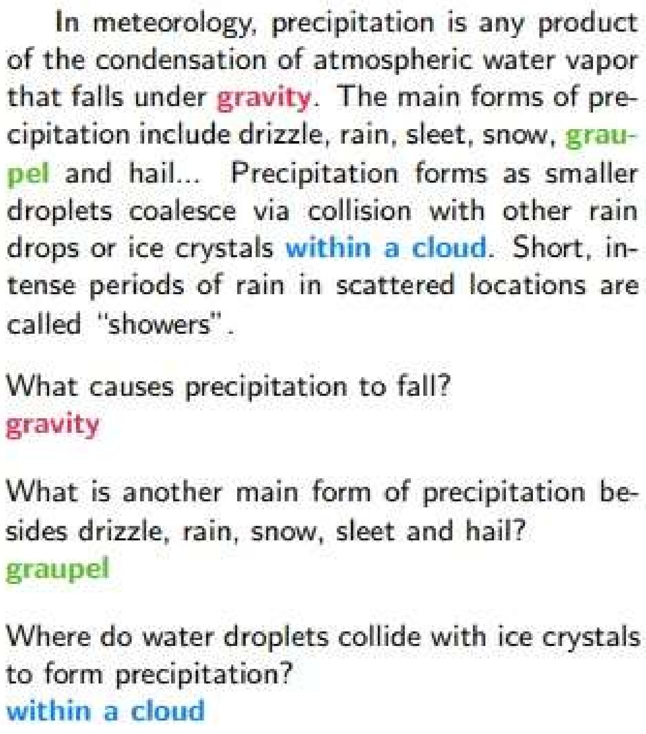
Bidirectional Attention Flow for Machine Comprehension

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, Hannaneh Hajishirzi

ICLR 2017

Bidirectional Attention Flow for Machine Comprehension

- Machine Comprehension = Question Answering
- Answer는 Context 안에 존재



Context

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".

Question1

What causes precipitation to fall?

Answer1

gravity

Question2

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

Answer2

graupel

Question3

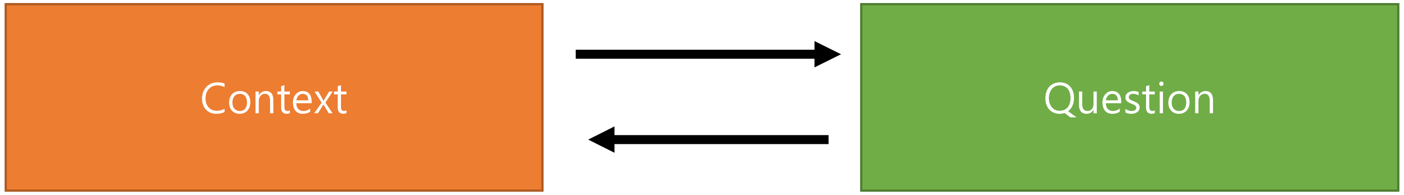
Where do water droplets collide with ice crystals to form precipitation?

Answer3

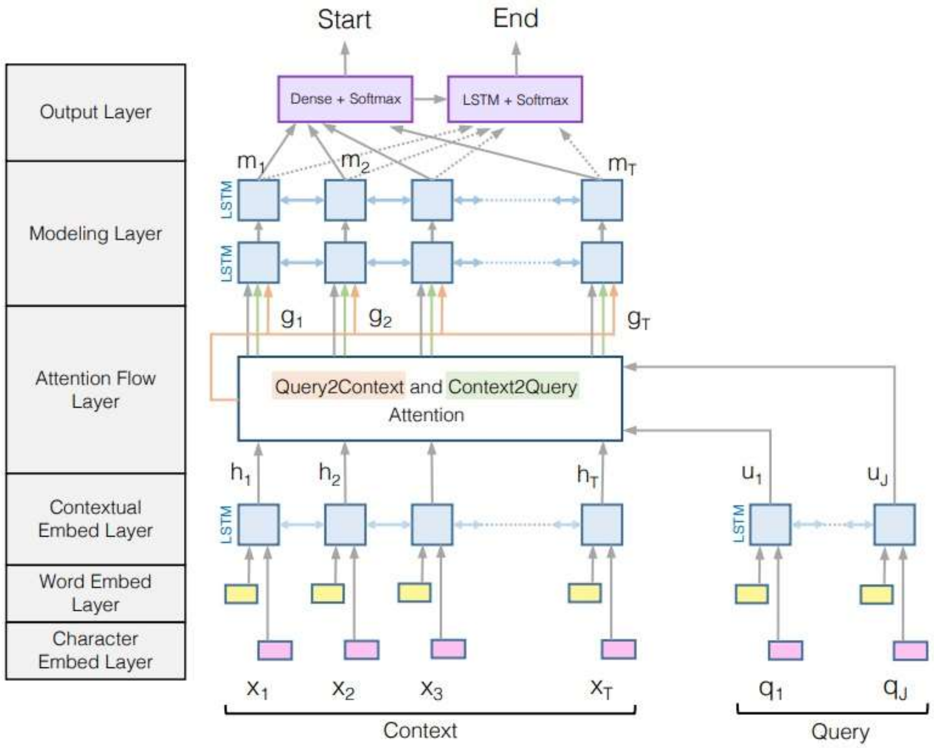
within a cloud

Bidirectional Attention Flow for Machine Comprehension

- Attention 양방향

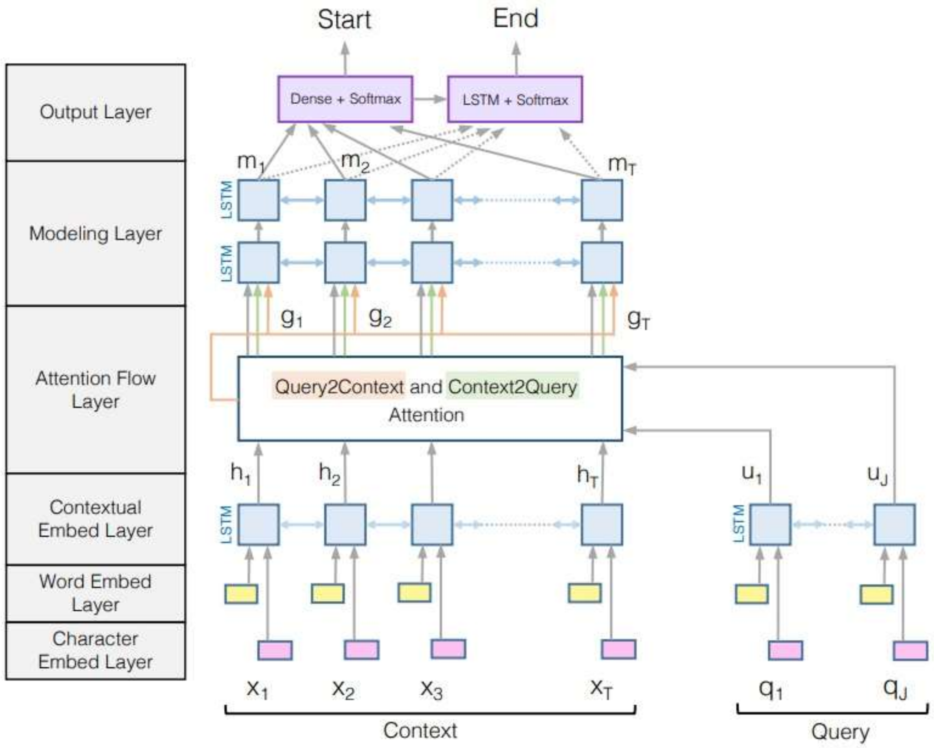


BiDAF

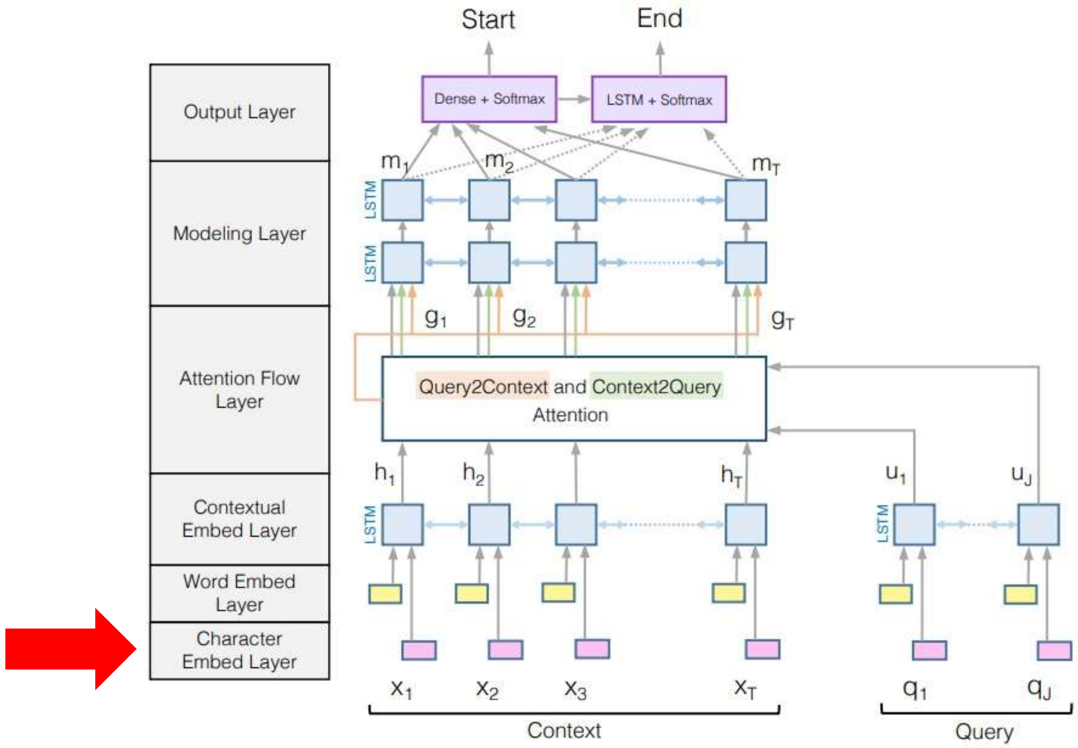


- 여러 Layer를 쌓은 계층적 구조
- Word, Character, Contextual Embed Layer
  - Context와 Question의 representation을 여러 단위에서 얻는다
- Attention Flow Layer
  - Context와 Question 사이에 정보의 교환

BiDAF

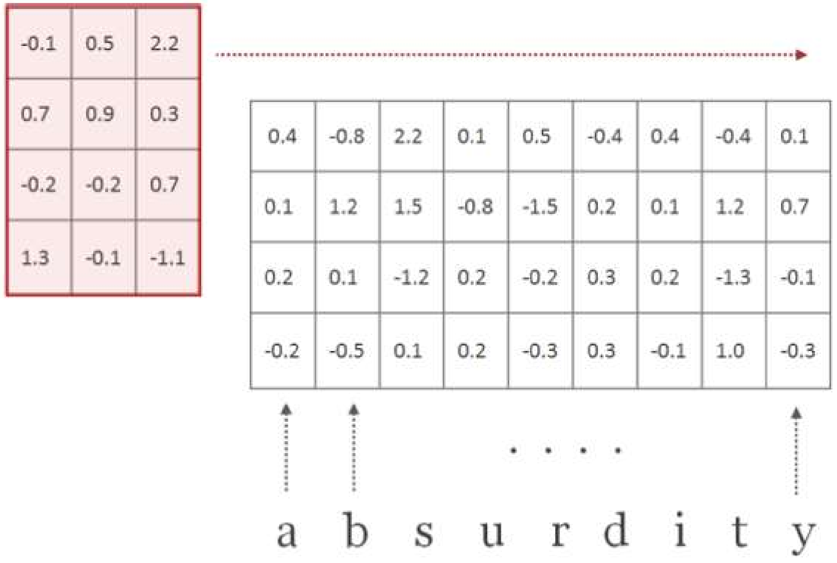


- Modeling Layer
  - Question-aware-Context word를 인코딩
- Output Layer
  - Answer 위치 예측



- Context와 Question의 각 Word를 1D-CNN을 이용하여 Vector로 표현

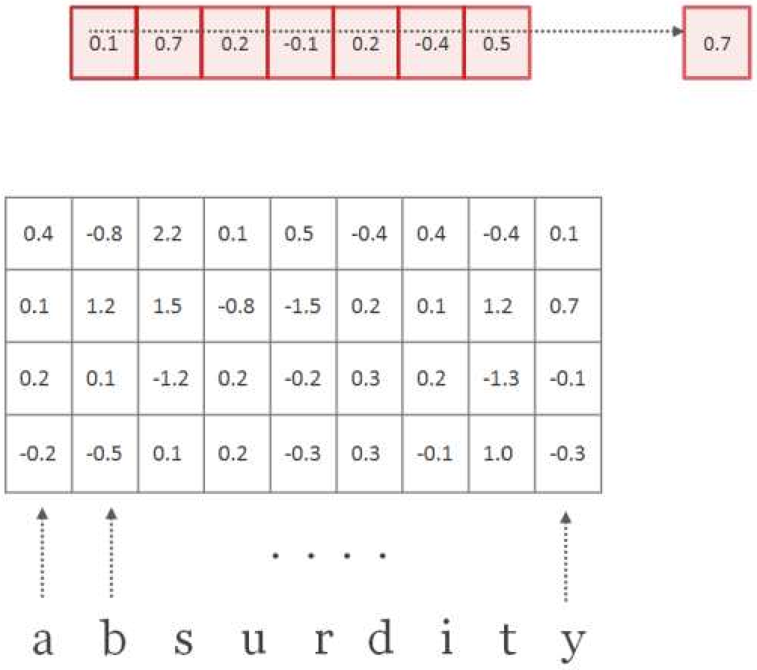
$\mathbf{H} \in \mathbb{R}^{d \times w}$: Convolutional filter matrix of width $w = 3$



-0.1	0.5	2.2
0.7	0.9	0.3
-0.2	-0.2	0.7
1.3	-0.1	-1.1

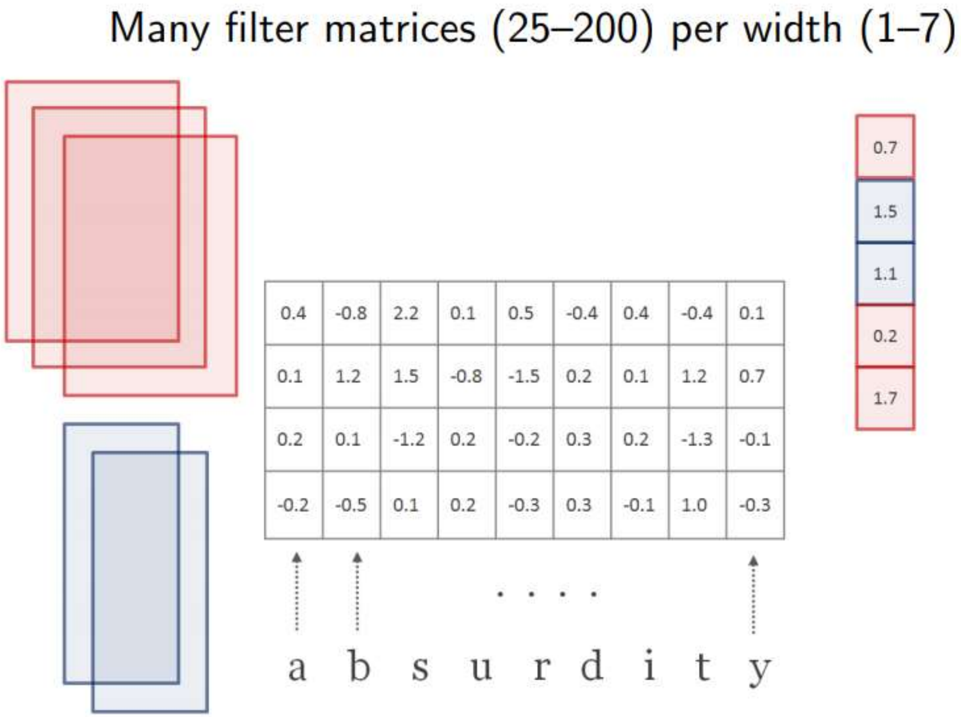
0.4	-0.8	2.2	0.1	0.5	-0.4	0.4	-0.4	0.1
0.1	1.2	1.5	-0.8	-1.5	0.2	0.1	1.2	0.7
0.2	0.1	-1.2	0.2	-0.2	0.3	0.2	-1.3	-0.1
-0.2	-0.5	0.1	0.2	-0.3	0.3	-0.1	1.0	-0.3

$$y[1] = \max_i \{\mathbf{f}[i]\}$$



0.1	0.7	0.2	-0.1	0.2	-0.4	0.5		0.7

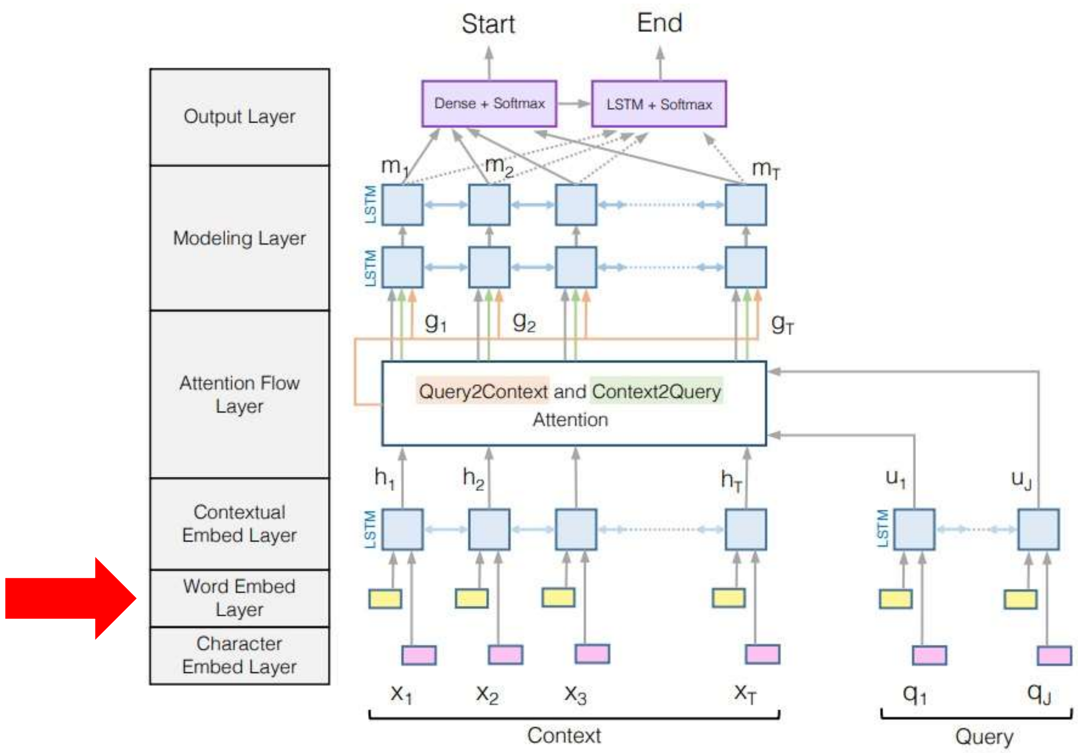
0.4	-0.8	2.2	0.1	0.5	-0.4	0.4	-0.4	0.1
0.1	1.2	1.5	-0.8	-1.5	0.2	0.1	1.2	0.7
0.2	0.1	-1.2	0.2	-0.2	0.3	0.2	-1.3	-0.1
-0.2	-0.5	0.1	0.2	-0.3	0.3	-0.1	1.0	-0.3



Many filter matrices (25–200) per width (1–7)

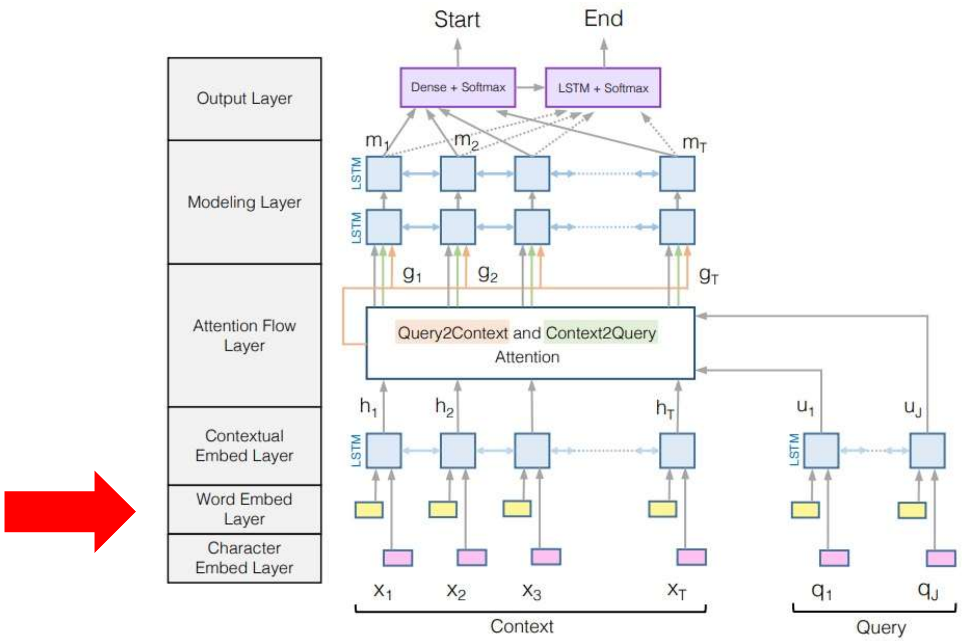
0.4	-0.8	2.2	0.1	0.5	-0.4	0.4	-0.4	0.1
0.1	1.2	1.5	-0.8	-1.5	0.2	0.1	1.2	0.7
0.2	0.1	-1.2	0.2	-0.2	0.3	0.2	-1.3	-0.1
-0.2	-0.5	0.1	0.2	-0.3	0.3	-0.1	1.0	-0.3

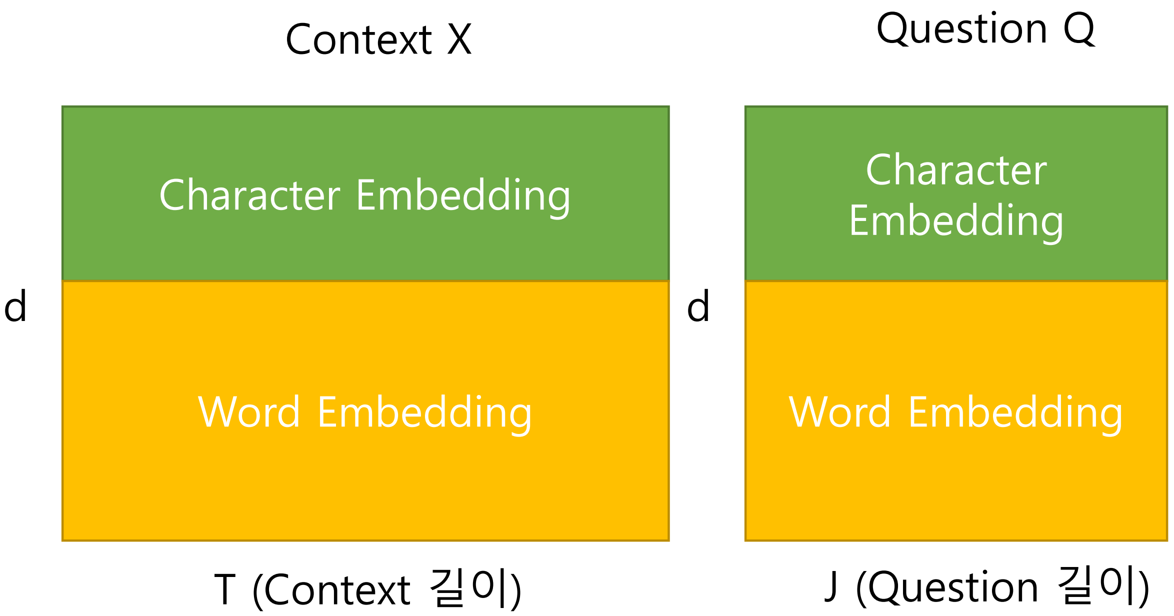
*실제로는
100 filter with
width 5



- Glove를 사용 word vector얻음

- Character와 Word Embedding을 concat해서 Embedding 생성
- Context는 $X \in R^{d \times T}$, Question은 $Q \in R^{d \times J}$
- OOV 문제를 완화




Context X

Question Q

d

Character Embedding

Word Embedding

T (Context 길이)

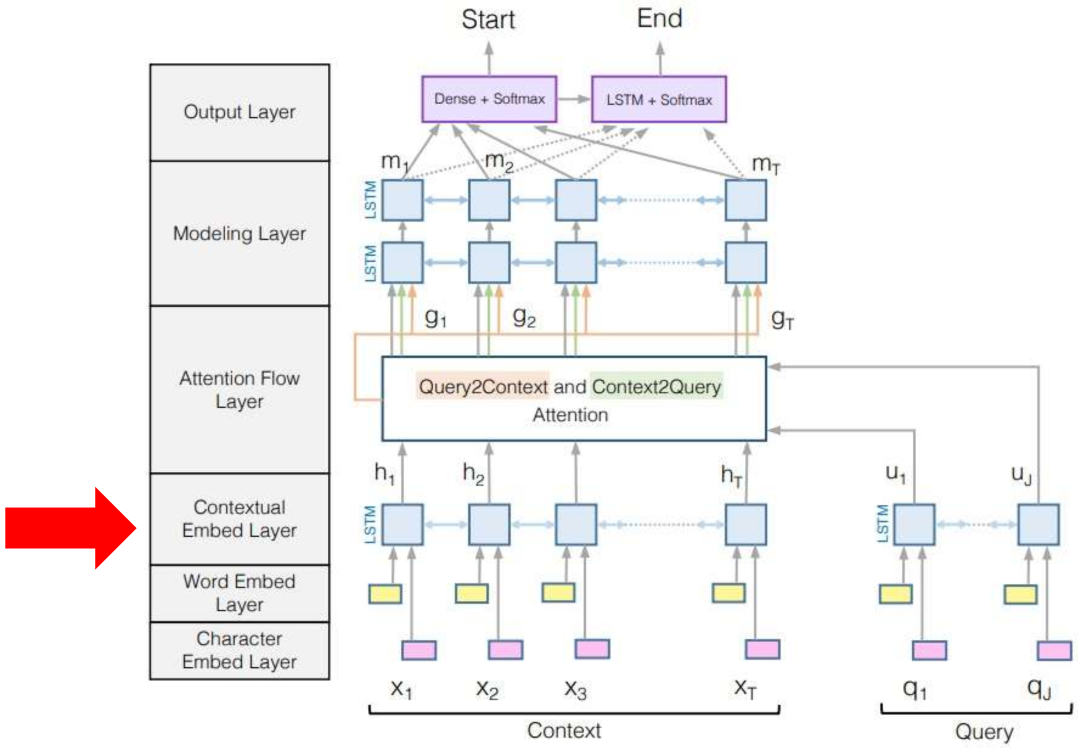
d

Character Embedding

Word Embedding

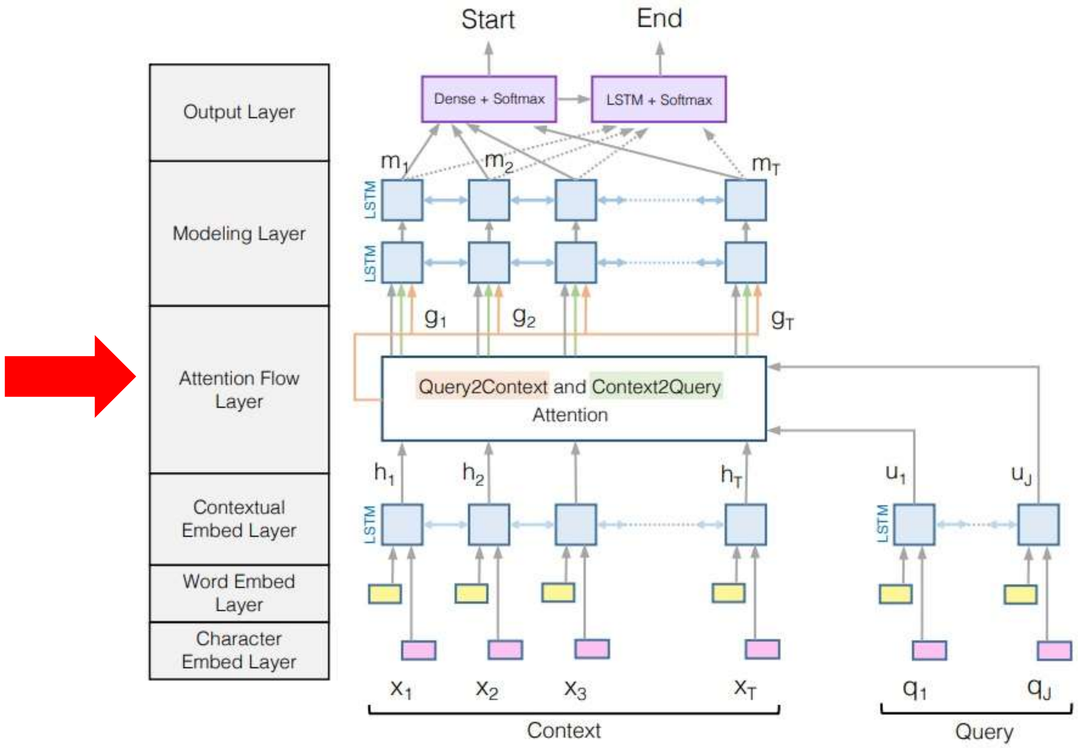
J (Question 길이)

T	Context 길이	X	Context의 Word + Char 임베딩	H	Context의 Contextual 임베딩
J	Question 길이	Q	Question의 Word + Char 임베딩	U	Question의 Contextual 임베딩



- X와 Q를 Bi-LSTM 통과시켜 Context를 고려한 Embedding을 생성
- Context Embedding $\mathrm{H} \in R^{2d \times T}$
- Question Embedding $U \in R^{2d \times J}$

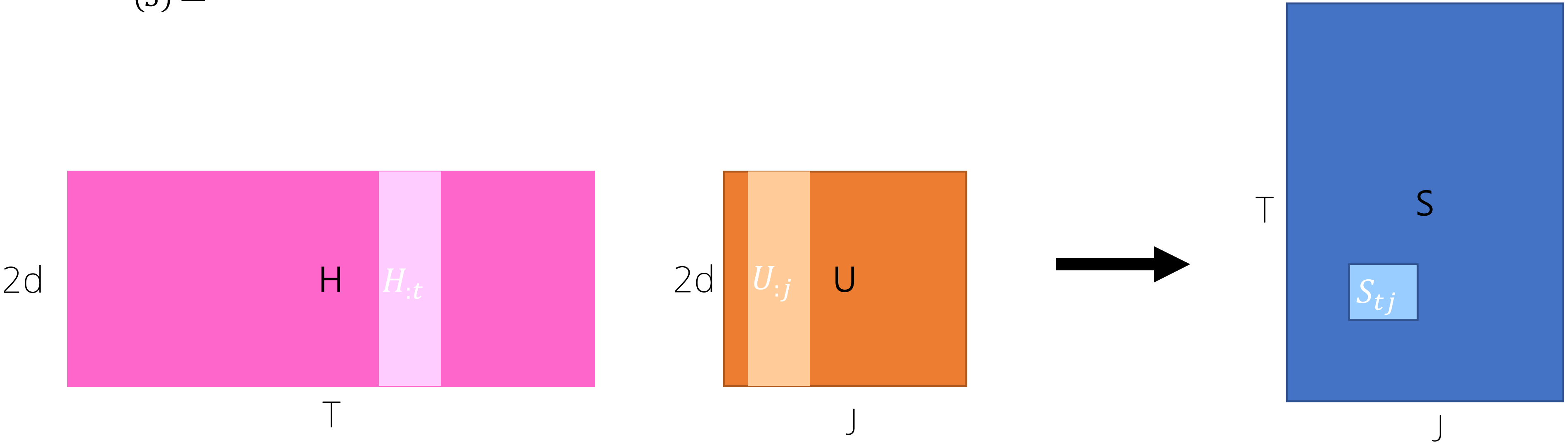
T	Context 길이	H	Context Embedding
J	Question 길이	U	Question Embedding



- Context와 Question의 정보를 서로 연결
- Question의 정보가 결합된 Context 만들기
- 먼저 similarity matrix를 만듬
- Similarity matrix로 부터 Context-to-Question attention, Question-to-Context attention을 구함

T	Context 길이	H	Context Embedding
J	Question 길이	U	Question Embedding

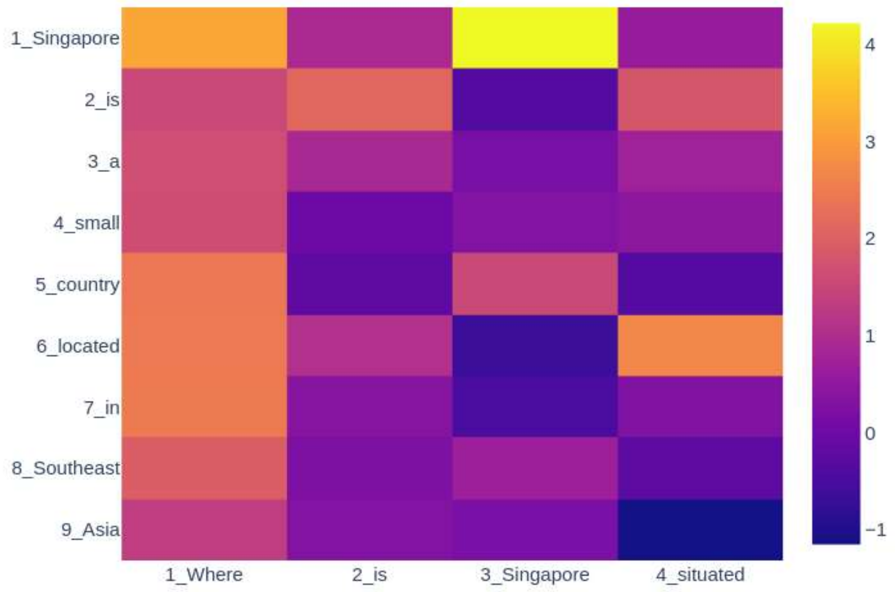
- 먼저 Similarity Matrix $S \in R^{T \times J}$ 생성
- S_{tj}는 t-th context word와 j-th context word의 유사도
- $S_{tj} = w_{(S)}^{T}[H_{:t}\,;\,U_{:j}\,;\,H_{:t} \circ U_{:j}]$
- $w_{(S)}$는 Learnable Parameter

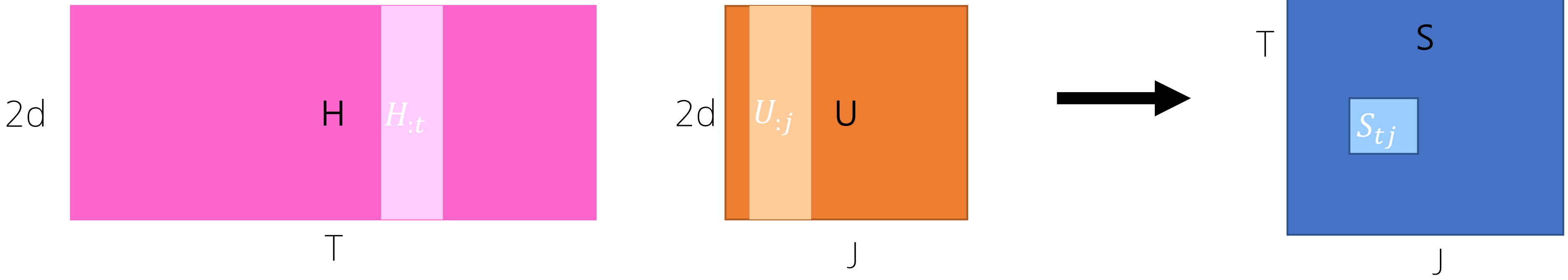


Character Embed Layer | Word Embed Layer | Contextual Embed Layer | **Attention Flow Layer** | Modeling Layer | Output Layer

T	g
J	g

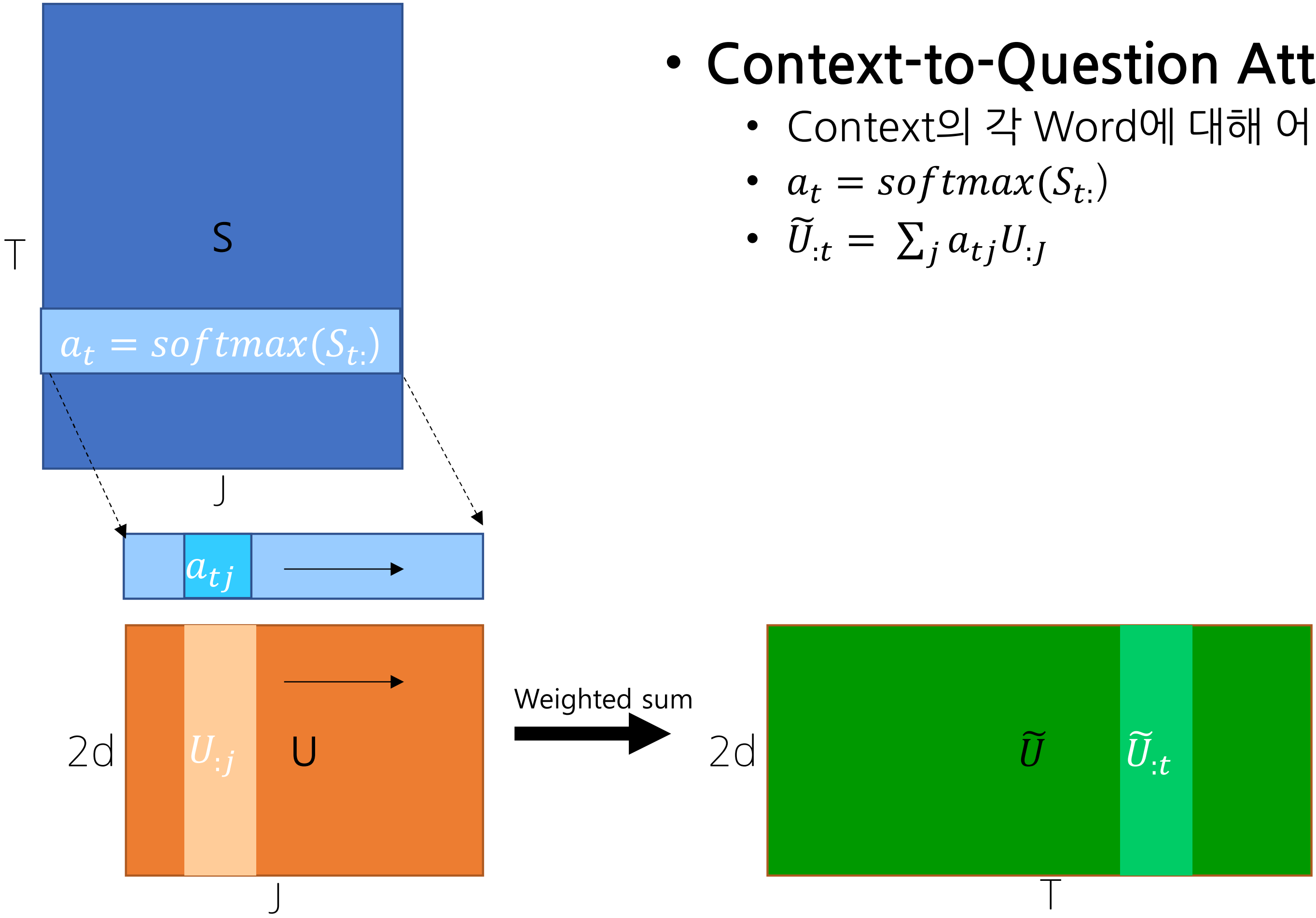
- 먼저
- S_{tj}는 사도
- S_{tj}
- $w_{(S}$



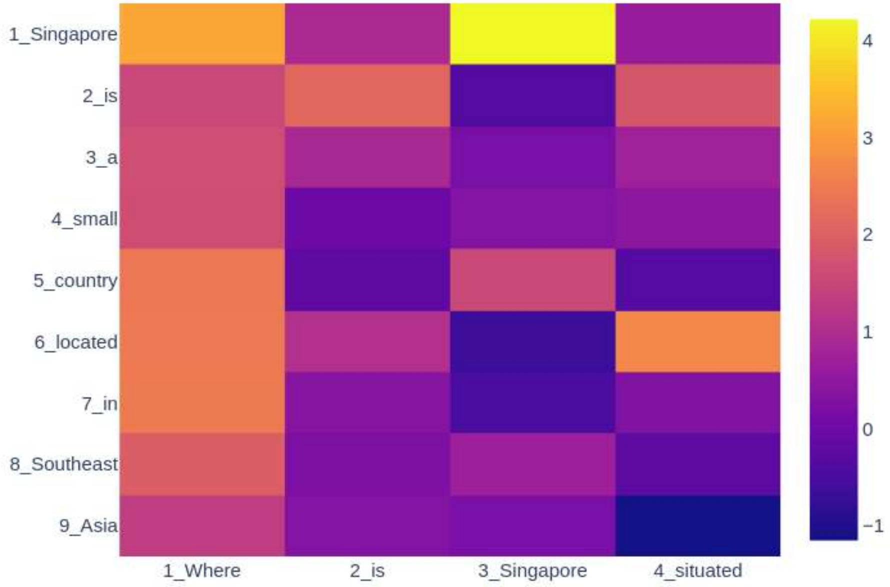


T	Context 길이	H	Context	$\tilde{U}$	Context-to-Question
J	Question 길이	U	Question		

- **Context-to-Question Attention**
  - Context의 각 Word에 대해 어떤 Question word가 관련성이 높은지
  - $a_t = softmax(S_{t:})$
  - $\tilde{U}_{:t} = \sum_j a_{tj} U_{:J}$



T	Context 길이	H	Context	$\tilde{U}$	Context-to-Question
J	Question 길이	U	Question		

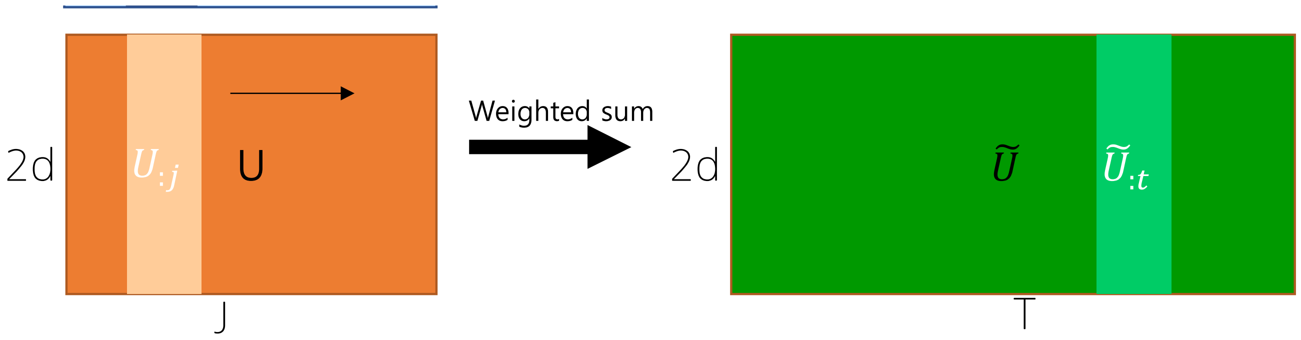


ntext-to-Question Attention

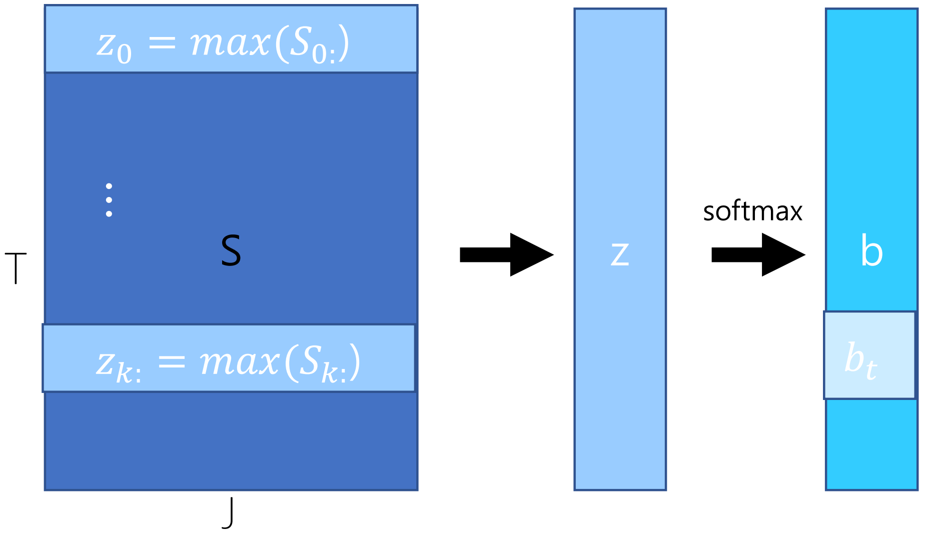
Context의 각 Word에 대해 어떤 Question word가 관련성이 높은지

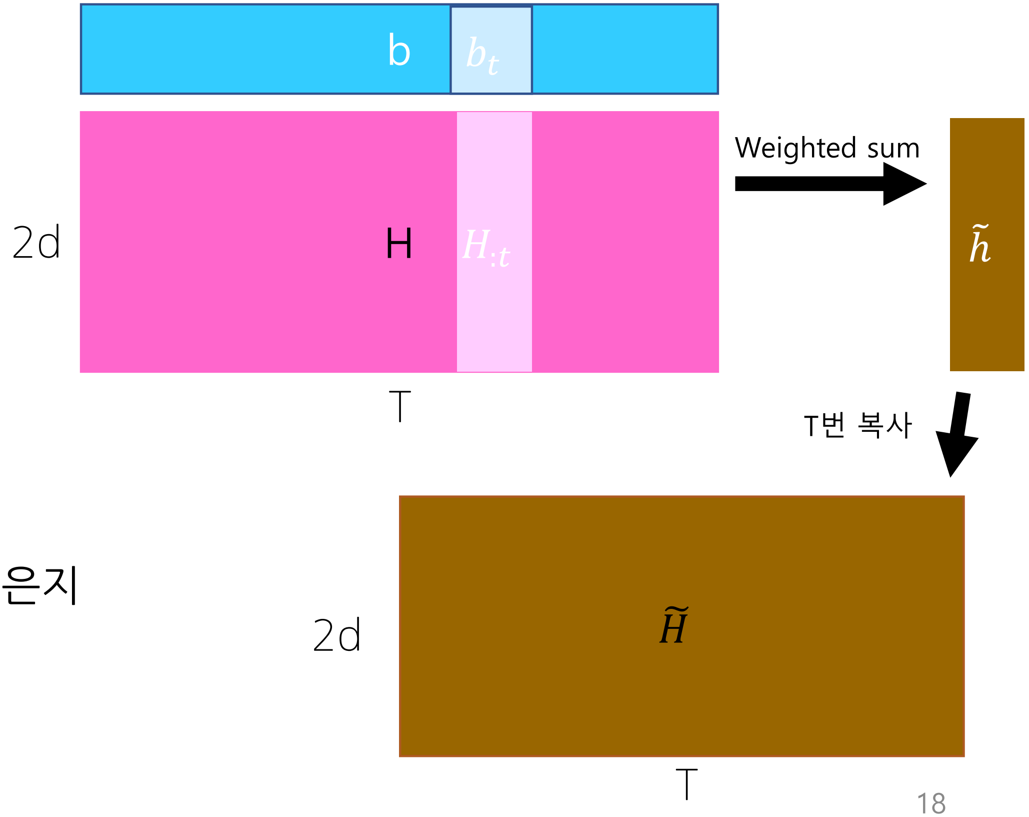
$a_t = softmax(S_{t:})$

$\tilde{U}_{:t} = \sum_j a_{tj} U_{:J}$



T	Context 길이	H	Context	$\tilde{U}$	Context-to-Question
J	Question 길이	U	Question	$\tilde{H}$	Question-to-Context

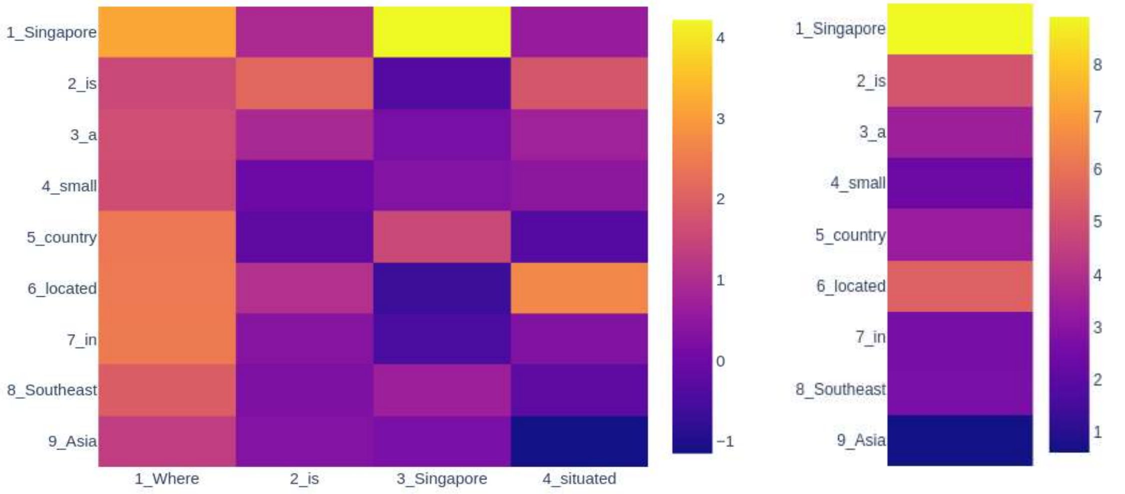




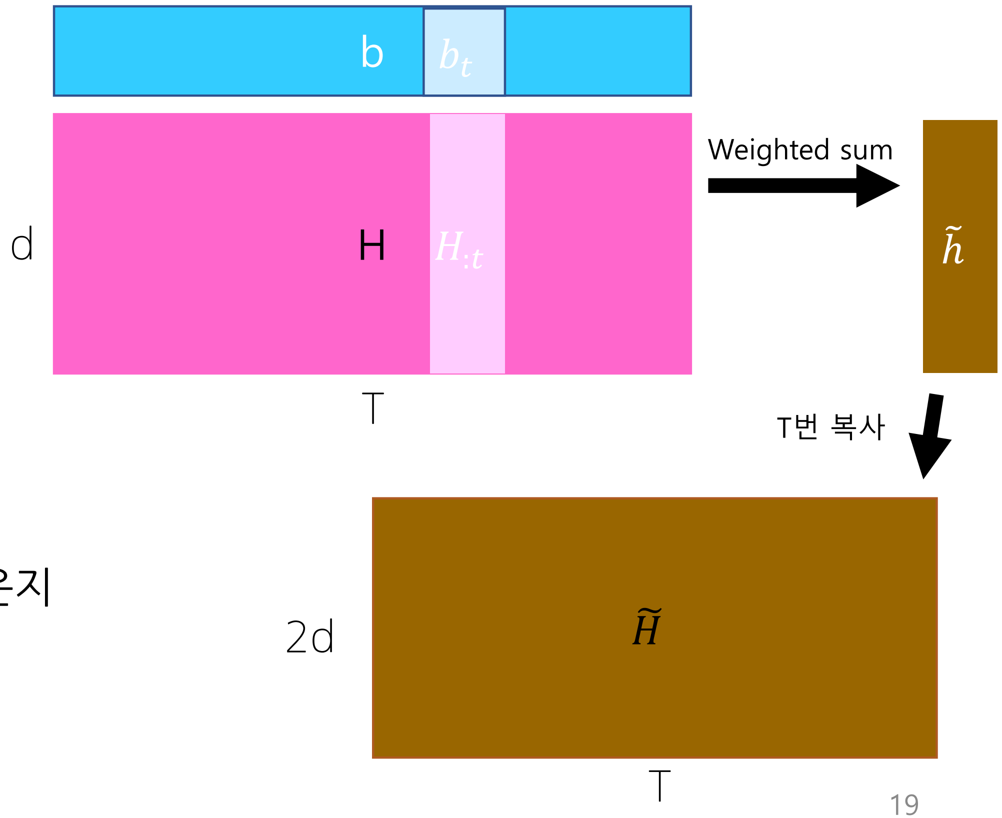
- **Question-to-Context Attention(Q2C)**
  - 어떤 Context word가 Question와 관련도가 가장 높은지
  - $b = softmax(max_{col}(S))$
  - $\tilde{h} = \sum_t b_t H_{:t}$
  - $\tilde{H}$ 는 $\tilde{h}$ 를 T번 복사

T

T	Context 길이	H	Context	$\tilde{U}$	Context-to-Question
J	Question 길이	U	Question	$\tilde{H}$	Question-to-Context



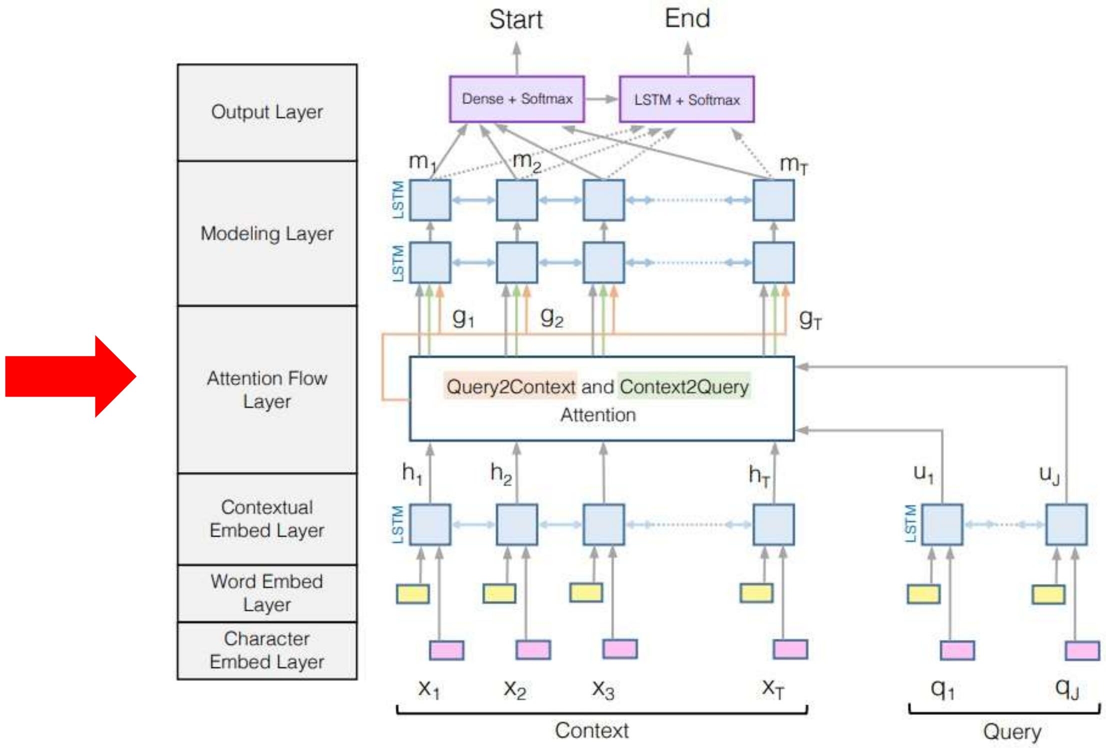
J



- **Question-to-Context Attention(Q2C)**
  - 어떤 Context word가 Question와 관련도가 가장 높은지
  - $b = softmax(max_{col}(S))$
  - $\tilde{h} = \sum_t b_t H_{:t}$
  - $\tilde{H}$ 는 $\tilde{h}$ 를 T번 복사

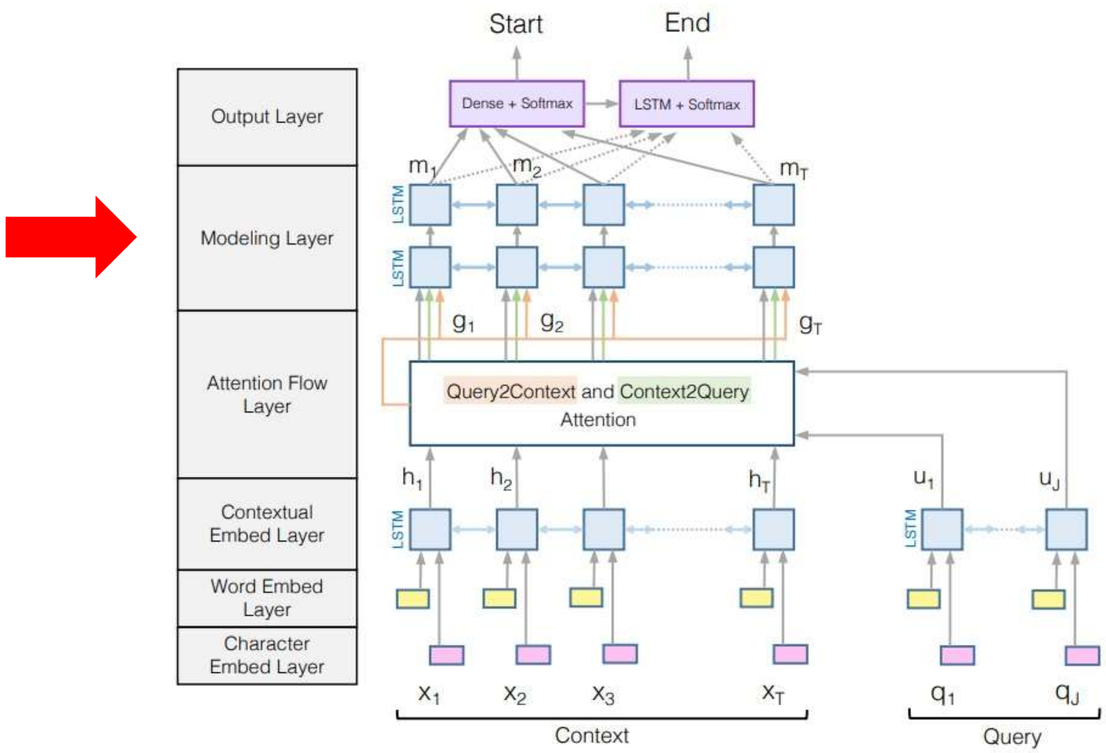
T

T	Context 길이	H	Context	$\tilde{U}$	Context-to-Question
J	Question 길이	U	Question	$\tilde{H}$	Question-to-Context



- G: Attention Flow Layer의 output
- $G_{:t} = [H_{:t}; \tilde{U}_{:t}\,; H_{:t} \circ \tilde{U}_{:t}; H_{:t} \circ \tilde{H}_{:t}]$
- G는 Question-aware representation of each context word

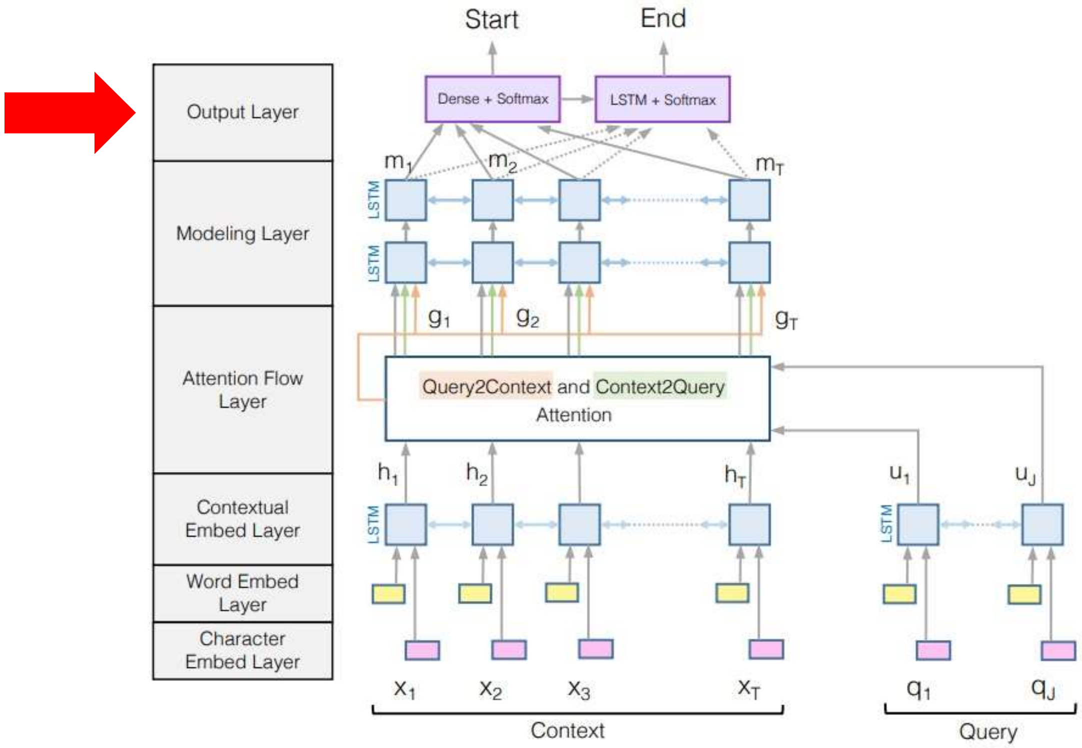
T	Context 길이	H	Context	$\tilde{U}$	Context-to-Question	G	Question-aware representation of context word
J	Question 길이	U	Question	$\tilde{H}$	Question-to-Context		



- 2-Layer Bi-LSTM
- Contextual Embed Layer와 다른 점은 Question 정보가 결합된 Context에 대해 LSTM

Character Embed Layer | Word Embed Layer | Contextual Embed Layer | Attention Flow Layer | Modeling Layer | **Output Layer**

T	Context 길이	H	Context	$\tilde{U}$	Context-to-Question	G	Question-aware representation of context word
J	Question 길이	U	Question	$\tilde{H}$	Question-to-Context		



- Context의 각 word가 answer의 start일 확률, end일 확률 계산

Question

Which NFL team won Super **Bowl?**

Context

The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24-10

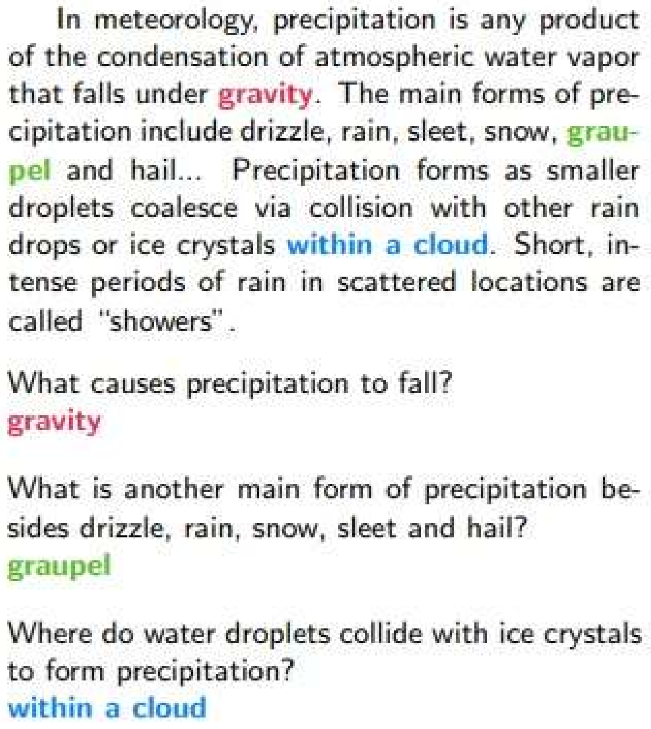
	The	American	Foot Ball	Conference	Champion	Denver	Broncos	Defeated	The National	Football	Conference	Champion	Carolina	Panthers	24-10
P(start)	0.01x	0.01x	0.01x	0.01x	0.01x	0.8	0.01x	0.01x	0.01x	0.01x	0.01x	0.01x	0.01x	0.01x	0.01x
P(end)	0.01x	0.01x	0.01x	0.01x	0.01x	0.01x	0.8	0.01x	0.01x	0.01x	0.01x	0.01x	0.01x	0.01x	0.01x

Answer

P(start) * P(end) 가 가장 높은 ‘Denver Broncos’가 answer로 선택

Experiments

- Dataset
  - SQuAD
    - Wikipedia article을 context로 사용
    - Answer는 context 안에 존재
    - Train set(90k), Dev set(10k)



Context

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".

Question1
Answer1

What causes precipitation to fall?
gravity

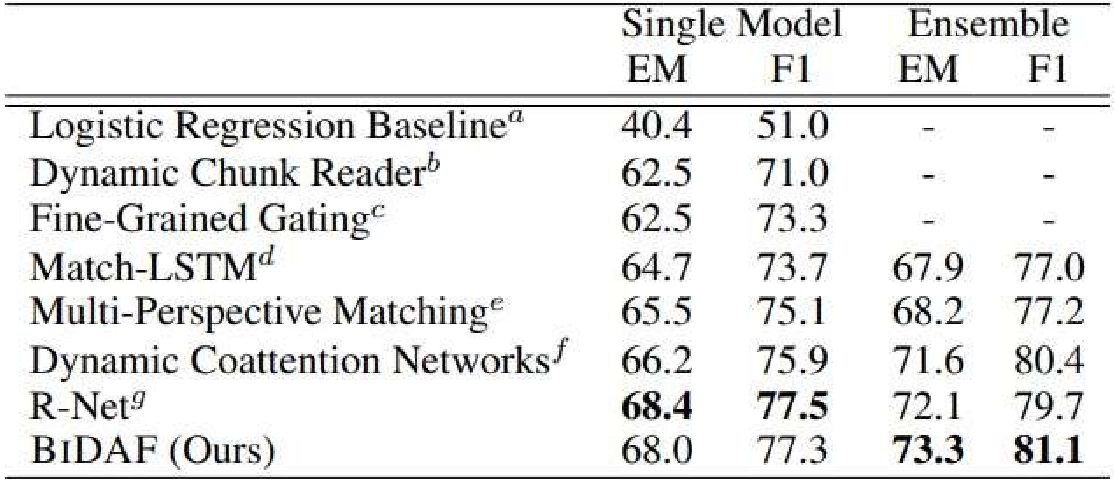
Question2
Answer2

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

Question3
Answer3

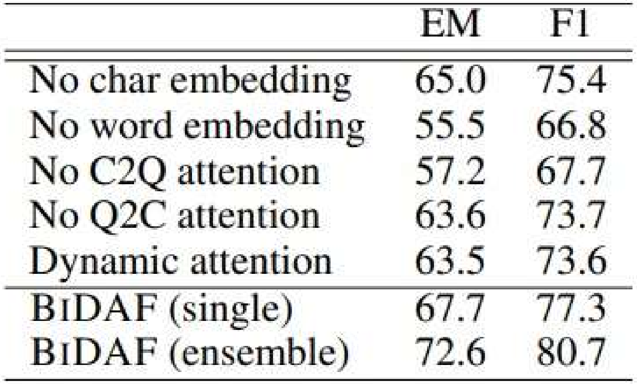
Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Experiments



	Single Model		Ensemble	
	EM	F1	EM	F1
Logistic Regression Baseline[a]	40.4	51.0	-	-
Dynamic Chunk Reader[b]	62.5	71.0	-	-
Fine-Grained Gating[c]	62.5	73.3	-	-
Match-LSTM[d]	64.7	73.7	67.9	77.0
Multi-Perspective Matching[e]	65.5	75.1	68.2	77.2
Dynamic Coattention Networks[f]	66.2	75.9	71.6	80.4
R-Net[g]	**68.4**	**77.5**	72.1	79.7
BiDAF (Ours)	68.0	77.3	**73.3**	**81.1**

(a) Results on the SQuAD test set

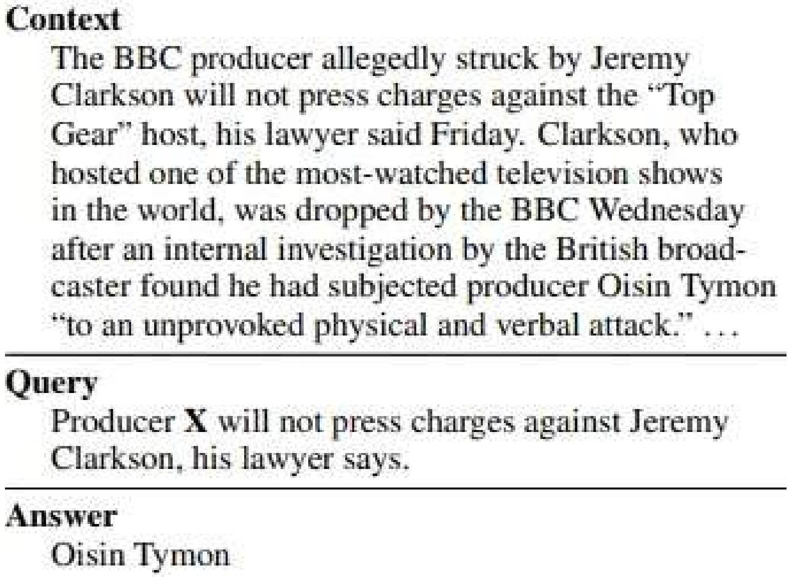


	EM	F1
No char embedding	65.0	75.4
No word embedding	55.5	66.8
No C2Q attention	57.2	67.7
No Q2C attention	63.6	73.7
Dynamic attention	63.5	73.6
BiDAF (single)	67.7	77.3
BiDAF (ensemble)	72.6	80.7

(b) Ablations on the SQuAD dev set

Experiments

- Dataset
  - CNN/DailyMail Dataset
    - 뉴스 기사를 context로 사용
    - Context를 요약한 Question의 빈칸 채우기
    - CNN(300k/4k/3k)
    - DailyMail(879k/65k/53k)



Context

The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...

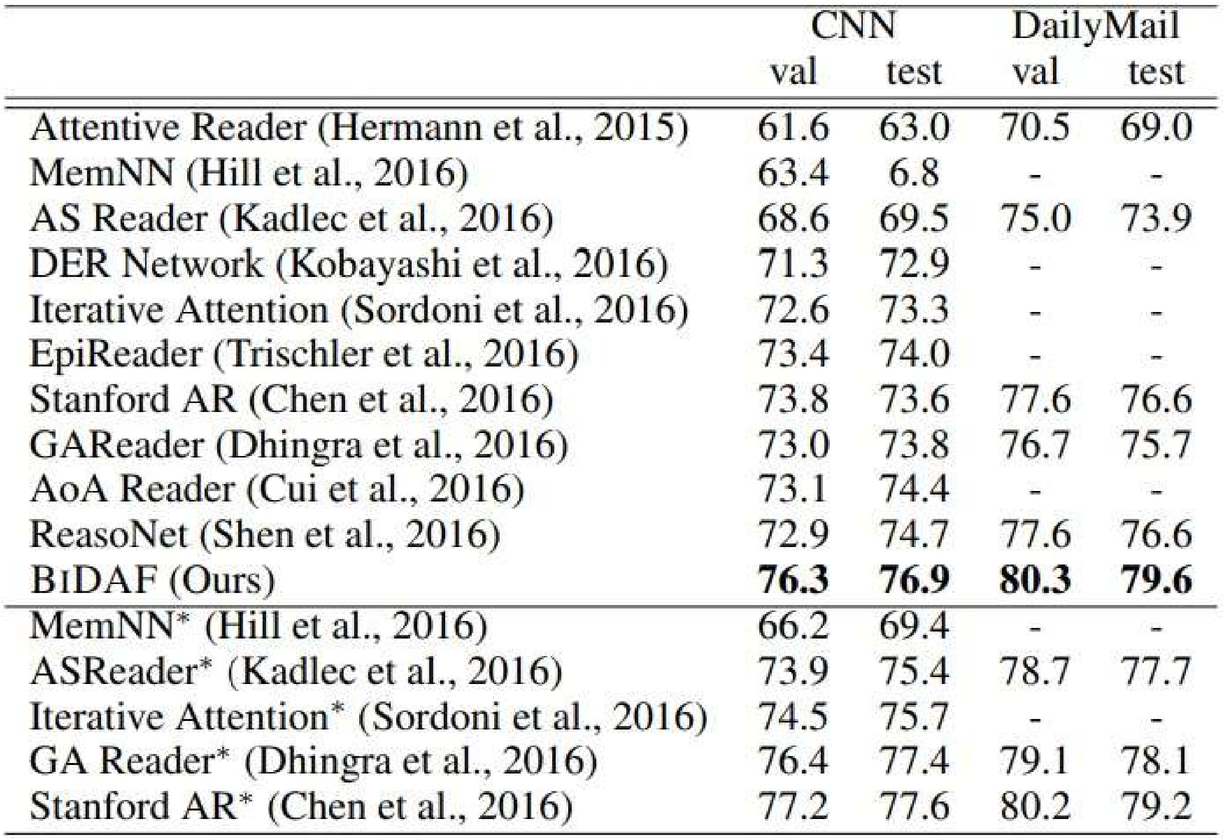
Query

Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.

Answer

Oisin Tymon

Experiments



	CNN		DailyMail	
	val	test	val	test
Attentive Reader (Hermann et al., 2015)	61.6	63.0	70.5	69.0
MemNN (Hill et al., 2016)	63.4	6.8	-	-
AS Reader (Kadlec et al., 2016)	68.6	69.5	75.0	73.9
DER Network (Kobayashi et al., 2016)	71.3	72.9	-	-
Iterative Attention (Sordoni et al., 2016)	72.6	73.3	-	-
EpiReader (Trischler et al., 2016)	73.4	74.0	-	-
Stanford AR (Chen et al., 2016)	73.8	73.6	77.6	76.6
GAReader (Dhingra et al., 2016)	73.0	73.8	76.7	75.7
AoA Reader (Cui et al., 2016)	73.1	74.4	-	-
ReasoNet (Shen et al., 2016)	72.9	74.7	77.6	76.6
BiDAF (Ours)	**76.3**	**76.9**	**80.3**	**79.6**
MemNN* (Hill et al., 2016)	66.2	69.4	-	-
ASReader* (Kadlec et al., 2016)	73.9	75.4	78.7	77.7
Iterative Attention* (Sordoni et al., 2016)	74.5	75.7	-	-
GA Reader* (Dhingra et al., 2016)	76.4	77.4	79.1	78.1
Stanford AR* (Chen et al., 2016)	77.2	77.6	80.2	79.2

* = Ensemble

- BiDAF: https://arxiv.org/abs/1611.01603
- 코드: https://github.com/allenai/bi-att-flow
- https://towardsdatascience.com/the-definitive-guide-to-bi-directional-attention-flow-d0e96e9e666b