Bidirectional Attention Flow for Machine Comprehension

Minjoon Seo[1] Aniruddha Kembhavi[2] Ali Farhadi[1],[2] Hananneh Hajishirzi [1]

[1]University of Washington,

[2]Allen Institute for Artificial Intelligence

ICLR,2017
Presenter: Arshdeep Sekhon

Machine Comprehension

1. Machine comprehension: answering a query about a given context paragraph
2. requires modeling complex interactions between the context and the query.
3. Attention Mechanisms very successful in QA and Machine Comprehension

Related Work

Related Work:

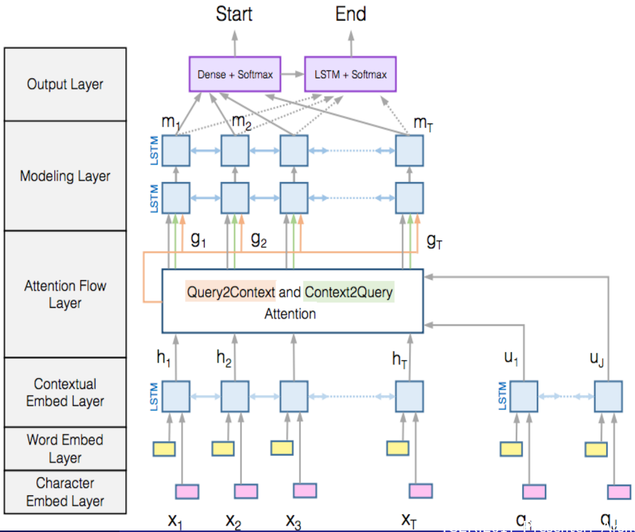
A Summarize the Context into a fixed vector

B Dynamic attention weights: attention weights are a function of the attention weights at the previous timestep

C Unidirectional: The attention weights are over the context

Overview of the Model

A hierarchical multistage model:

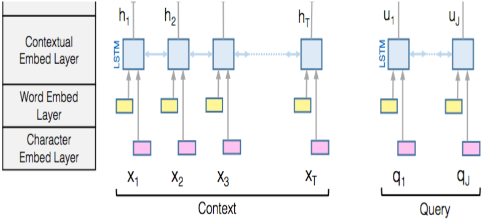


Character and Word Embedding Layer

Notations:
Context Paragraph:$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$
Query:$\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_J\}$

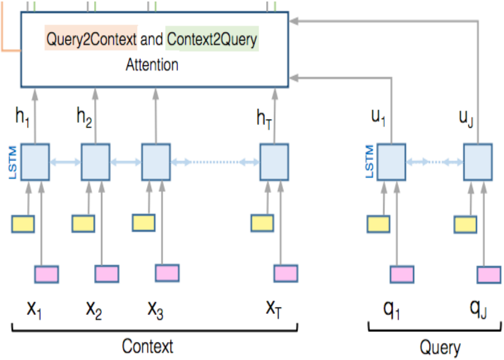
1. Character level: CNN
2. Use GloVe to get word embeddings for each word in the context as well as the query
3. Concatenate the two representations for word and character-level word representations
4. Pass to a highway network

Contextual Embedding Layer



1. Output of these word representation layers: $\boldsymbol{X} \in \mathbb{R}^{d\times T}$ for paragraph $\boldsymbol{Q} \in \mathbb{R}^{d\times J}$ for query
2. $\boldsymbol{X}$ and $\boldsymbol{Q}$ are input to a bidirectional LSTM network with hidden size d
3. Final representation for context:$\boldsymbol{H} \in \mathbb{R}^{2d\times T}$
4. Final representation for query: $\boldsymbol{U} \in \mathbb{R}^{2d\times J}$

Attention Modeling: Attention Flow Layer

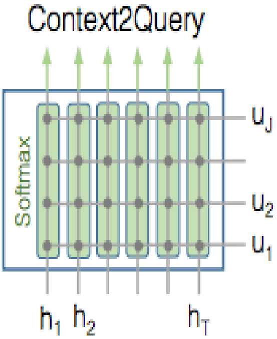


1. Fuse information between query and context
2. Previous Layer Query $\boldsymbol{U}$
3. Context Representation $\boldsymbol{H}$
4. Objective: Model Interactions between Query and Context
5. **Bidirectional** Attention:
    1. Context to Query
    2. Query to Context

Context to Query

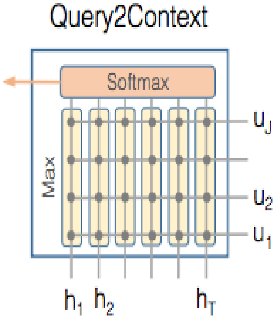
1. Need a Similarity Matrix $\boldsymbol{S}$
2. $\boldsymbol{S}_{tj} = \alpha(\boldsymbol{H}_{:t}, \boldsymbol{U}_{:j}) \in \mathbb{R}$
3. $\boldsymbol{S} \in \mathbb{R}^{T \times J}$
4. α is a trainable function that measures similarity
5. In this paper, $\alpha = \boldsymbol{w}_s^T[\boldsymbol{h}, \boldsymbol{u}, \boldsymbol{h} \circ \boldsymbol{u}]$
6. $\boldsymbol{h} \circ \boldsymbol{u}$ is element wise multiplication

Context to Query



1. Reweigh Query words for each context word
2. Which query word is important for a particular context word?
3. Easy to do using $\boldsymbol{S}$
4. for a context word c_t: Softmax($\boldsymbol{S}_{t:}$)
5. $\tilde{\boldsymbol{U}}_{:t} = \sum_j \boldsymbol{a}_{tj} \boldsymbol{U}_{:t}$
6. $\tilde{\boldsymbol{U}}$: Attended Query for all context vectors

Query to Context

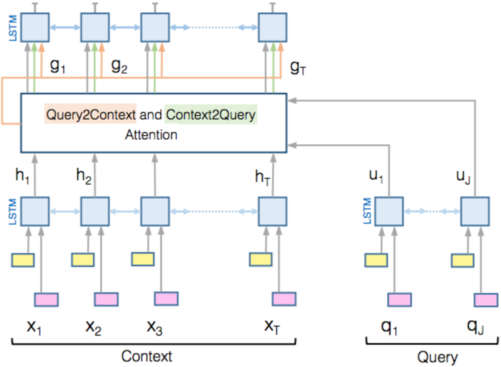


1. which context words have the closest similarity to one of the query words?
2. $\tilde{\boldsymbol{H}}_{:t}$ is output of Contextual Embedding
3. Choose the one with the highest value of $\boldsymbol{S}_{tj}$ for a particular t
4. $\boldsymbol{b} = Softmax(max_{col}(\boldsymbol{S})) \in \mathbb{R}^T$
5. $\tilde{\boldsymbol{h}} = \sum_t \boldsymbol{b}_t \boldsymbol{H}_{:t} \in \mathbb{R}^{2d}$
6. Replicate $\tilde{\boldsymbol{h}}$ for T

Attention Layer

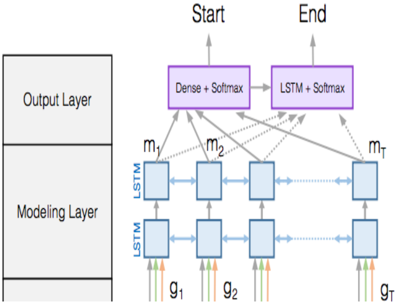
1. Combine the generated:
    1. $\boldsymbol{H}_{:t}$: Context Representation
    2. $\tilde{\boldsymbol{U}}_{:t}$: Context to Query [reweighted query for each context]
    3. $\tilde{\boldsymbol{H}}_{:t}$ Query to Context
2. $\boldsymbol{G}_{:t} = \beta(\boldsymbol{H}_{:t}, \tilde{\boldsymbol{U}}_{:t}, \tilde{\boldsymbol{H}}_{:t})$
3. β could be an MLP or any other trainable function
4. In this paper; $[\boldsymbol{h}; \tilde{\boldsymbol{u}}; \boldsymbol{h} \circ \tilde{\boldsymbol{u}}; \boldsymbol{h} \circ \tilde{\boldsymbol{h}}] \in \mathbb{R}^{8d \times T}$

Modeling Layer



1. input is $\boldsymbol{G} \in \mathbb{R}^{8d \times T}$
2. Modeling Layer is a bidirectional LSTM
3. Output $\boldsymbol{M} \in \mathbb{R}^{2d \times T}$
4. Similar to the Contextual Embedding Layer, but now query Aware
5. $\boldsymbol{M}$ is expected to contain contextual information about the word with respect to the entire context paragraph and the query

Output Layer



- Task is QA: Find the subphrase to answer the question
- Need a start and end pointer
- Start Pointer

$$\mathbf{p}^1 = softmax(\mathbf{w}_{p^1}^T[\mathbf{G}; \mathbf{M}]) \tag{1}$$

- End Pointer

$$\mathbf{p}^2 = softmax(\mathbf{w}_{p^2}^T[\mathbf{G}; \mathbf{M}^2]) \tag{2}$$

Loss Function

1. Training Loss: sum of the negative log probabilities of the true start and end indices by the predicted distributions

$$\boldsymbol{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log(\boldsymbol{p}^1_{y_i}) + \log(\boldsymbol{p}^2_{y_i}) \tag{3}$$

The Model

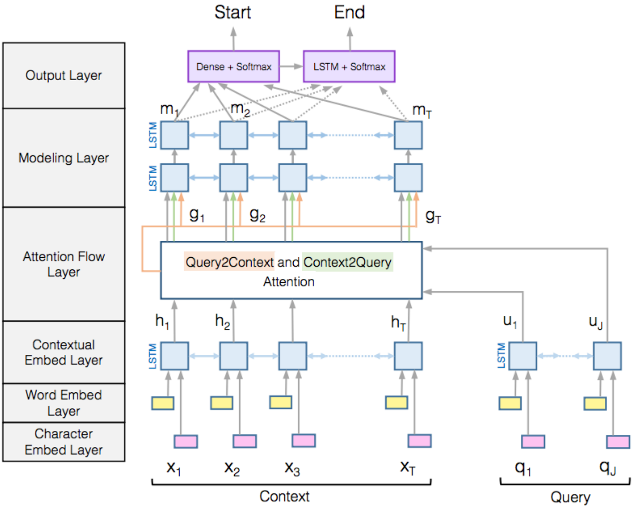
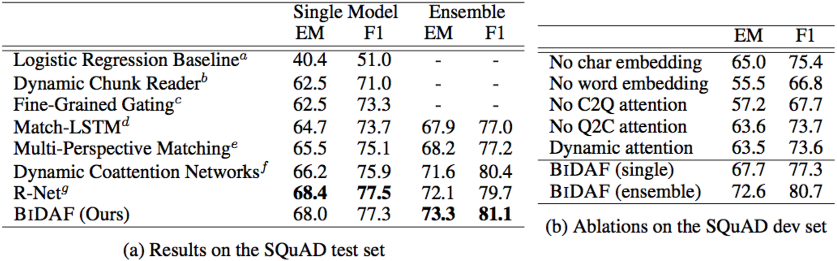


Figure: *

Results on QA



	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**

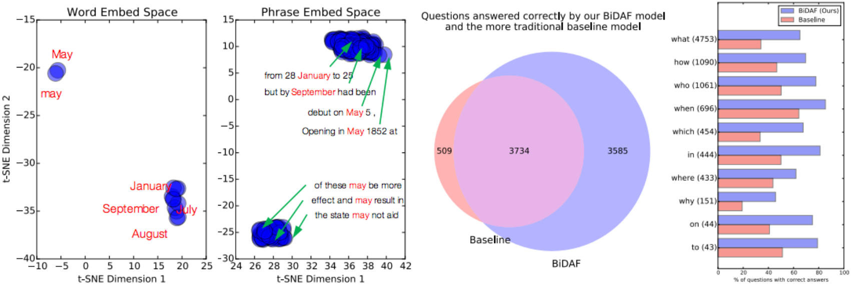
(a) Results on the SQuAD test set

	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

Figure: *

Visualization of embedding space

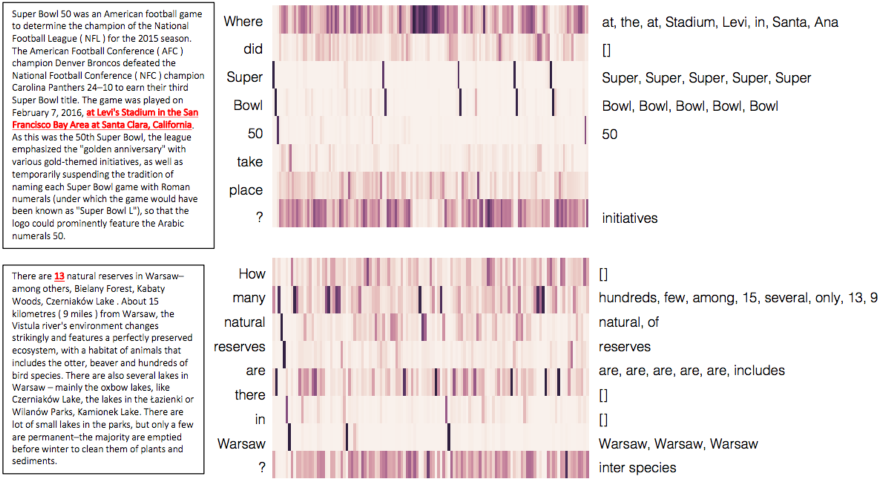


(a) (b) (c)

Figure: *

t-SNE: dimension reduction technique that visualizes high-dimensional data by giving each datapoint a location in a two or three-dimensional map

Attention Maps



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. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

There are 13 natural reserves in Warsaw– among others, Bielany Forest, Kabaty Woods, Czerniaków Lake . About 15 kilometres (9 miles) from Warsaw, the Vistula river's environment changes strikingly and features a perfectly preserved ecosystem, with a habitat of animals that includes the otter, beaver and hundreds of bird species. There are also several lakes in Warsaw – mainly the oxbow lakes, like Czerniaków Lake, the lakes in the Łazienki or Wilanów Parks, Kamionek Lake. There are lot of small lakes in the parks, but only a few are permanent–the majority are emptied before winter to clean them of plants and sediments.

Where
did
Super
Bowl
50
take
place
?
at, the, at, Stadium, Levi, in, Santa, Ana
[]
Super, Super, Super, Super, Super
Bowl, Bowl, Bowl, Bowl, Bowl
50
initiatives
How
many
natural
reserves
are
there
in
Warsaw
?
[]
hundreds, few, among, 15, several, only, 13, 9
natural, of
reserves
are, are, are, are, are, includes
[]
[]
Warsaw, Warsaw, Warsaw
inter species

Figure: *

Key points of the model

1. Bidirectional Attention Flow: C2Q and Q2C
2. Multistage hierarchical process
3. No early summary