Question answering and machine comprehension with neural attention

Minjoon Seo

PhD Student

Computer Science & Engineering

University of Washington

Two End-to-End Question Answering Systems with Neural Attention

- Bidirectional Attention Flow (BiDAF)
  - On Stanford Question Answering Dataset and CNN/DailyMail Cloze Test
- Query-Reduction Networks (QRN)
  - On bAbI QA and dialog, DSTC2 datasets

Two Question Answering Systems with Neural Attention

- Bidirectional Attention Flow (BiDAF)
  - On Stanford Question Answering Dataset and CNN/DailyMail Cloze Test
- Query-Reduction Networks (QRN)
  - On bAbI QA and dialog datasets

Question Answering Task (Stanford Question Answering Dataset, 2016)

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 $\boxed{\text{Denver Broncos}}$ 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.

Q: Which NFL team represented the AFC at Super Bowl 50?

A: Denver Broncos

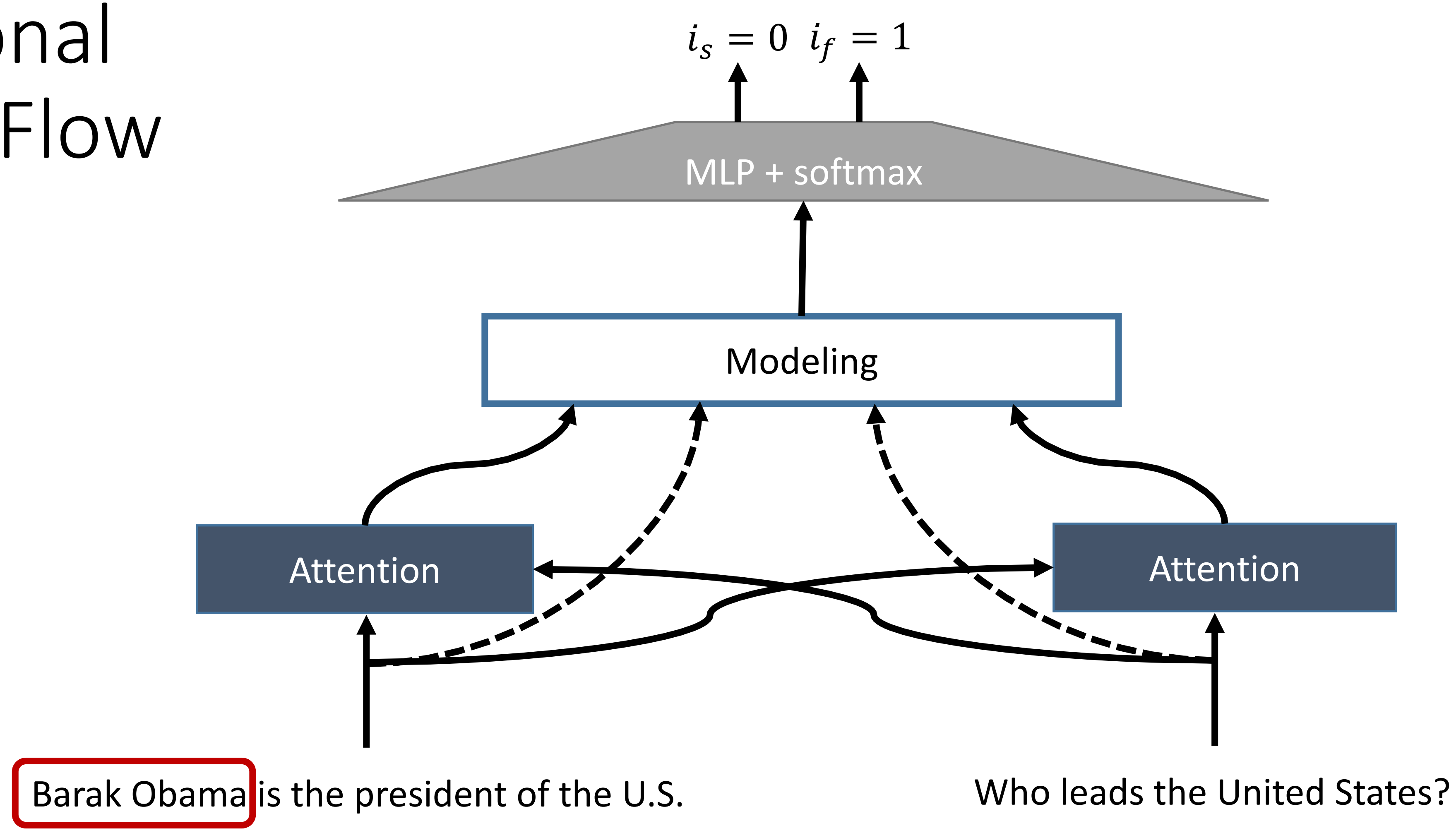
Why Neural Attention?

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.

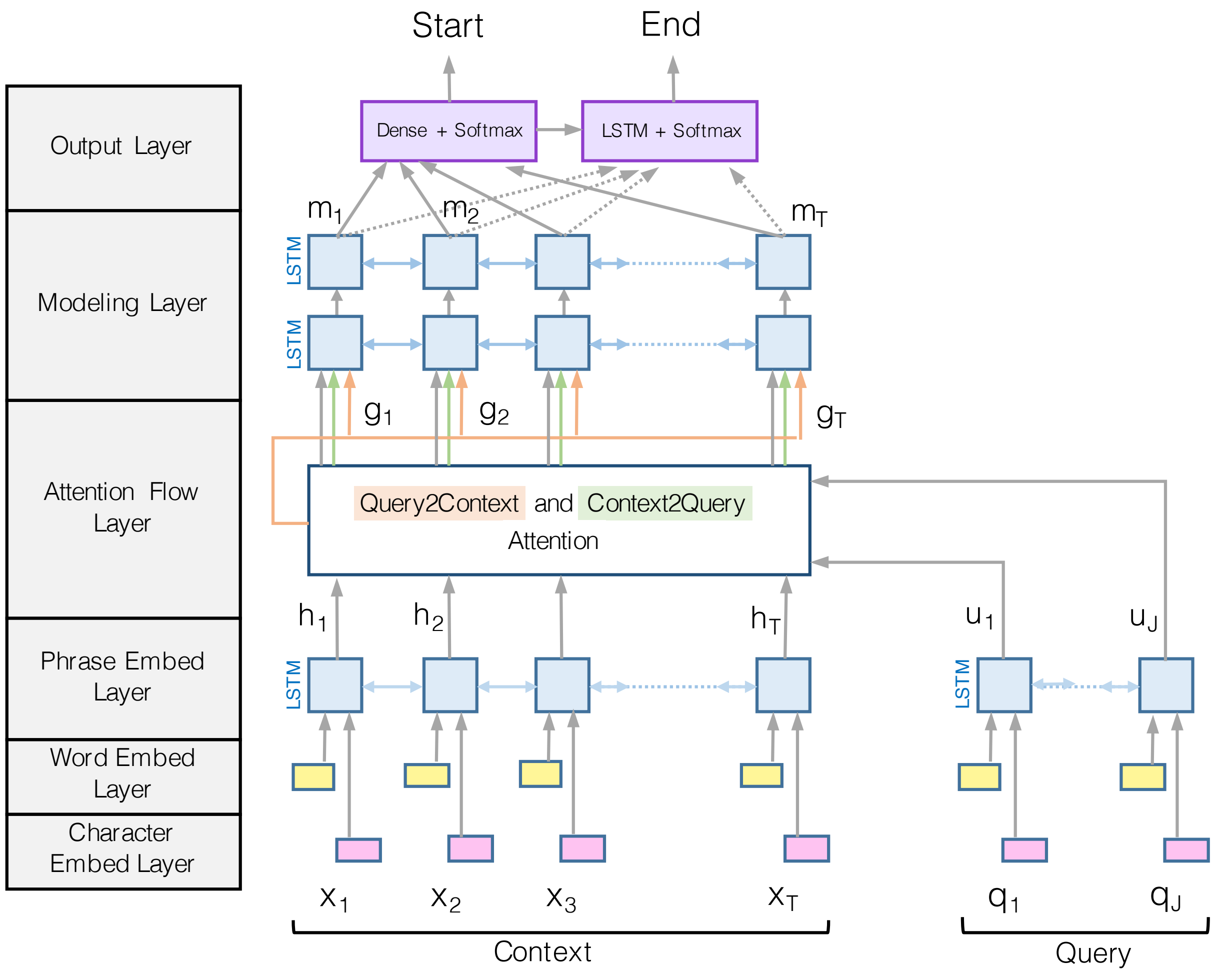
Q: Which NFL team represented the AFC at Super Bowl 50?

Allows a deep learning architecture to focus on the most relevant phrase of the context to the query in a *differentiable manner*.

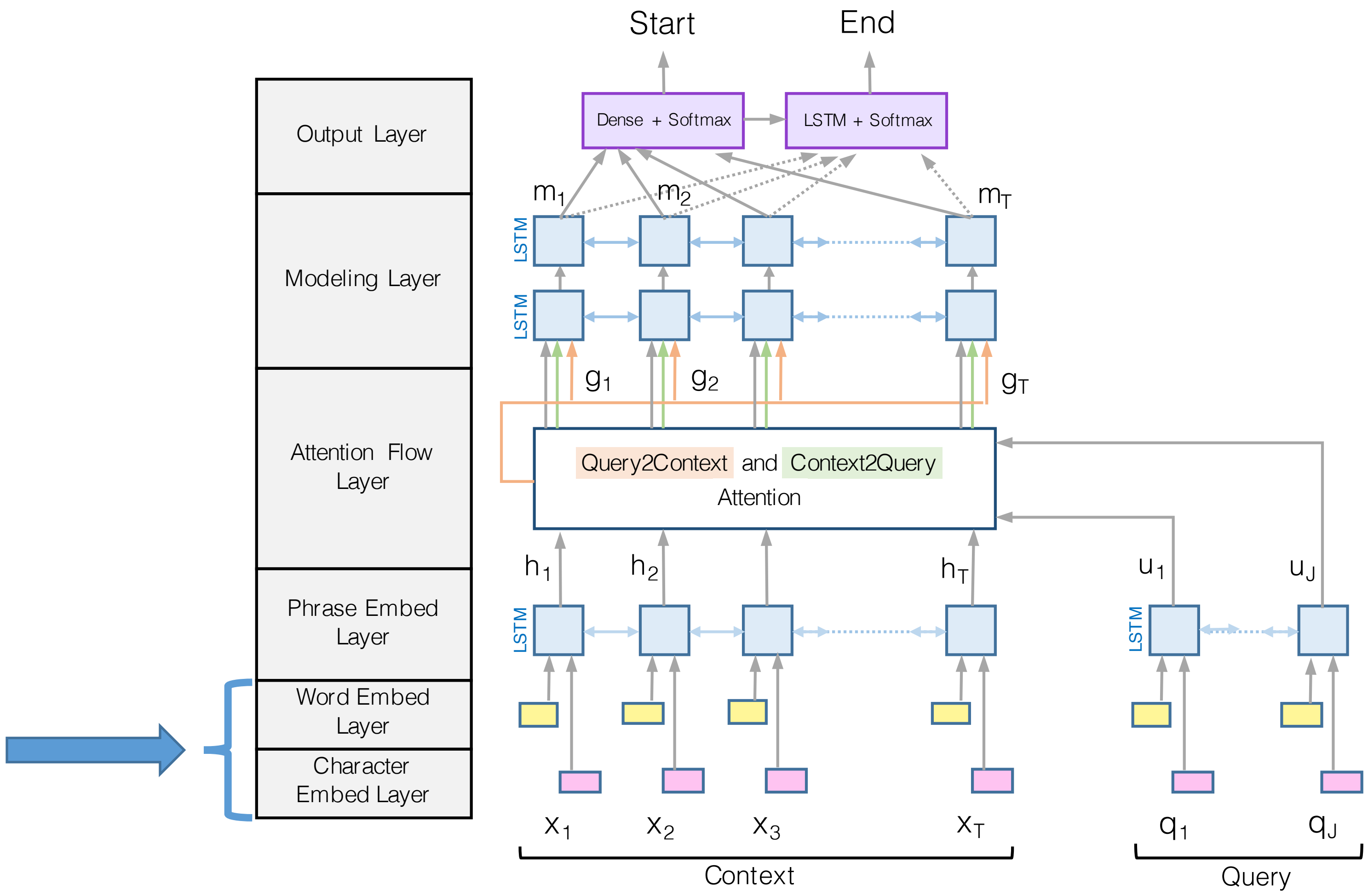
Our Model: Bi-directional Attention Flow (BiDAF)



(Bidirectional) Attention Flow



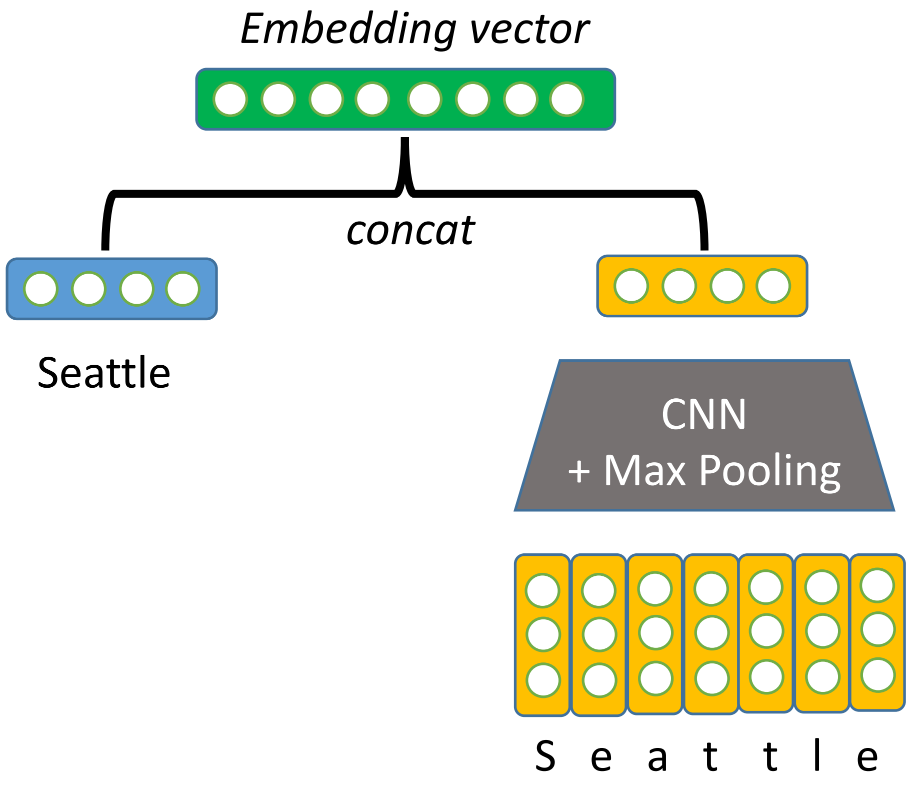
Char/Word Embedding Layers



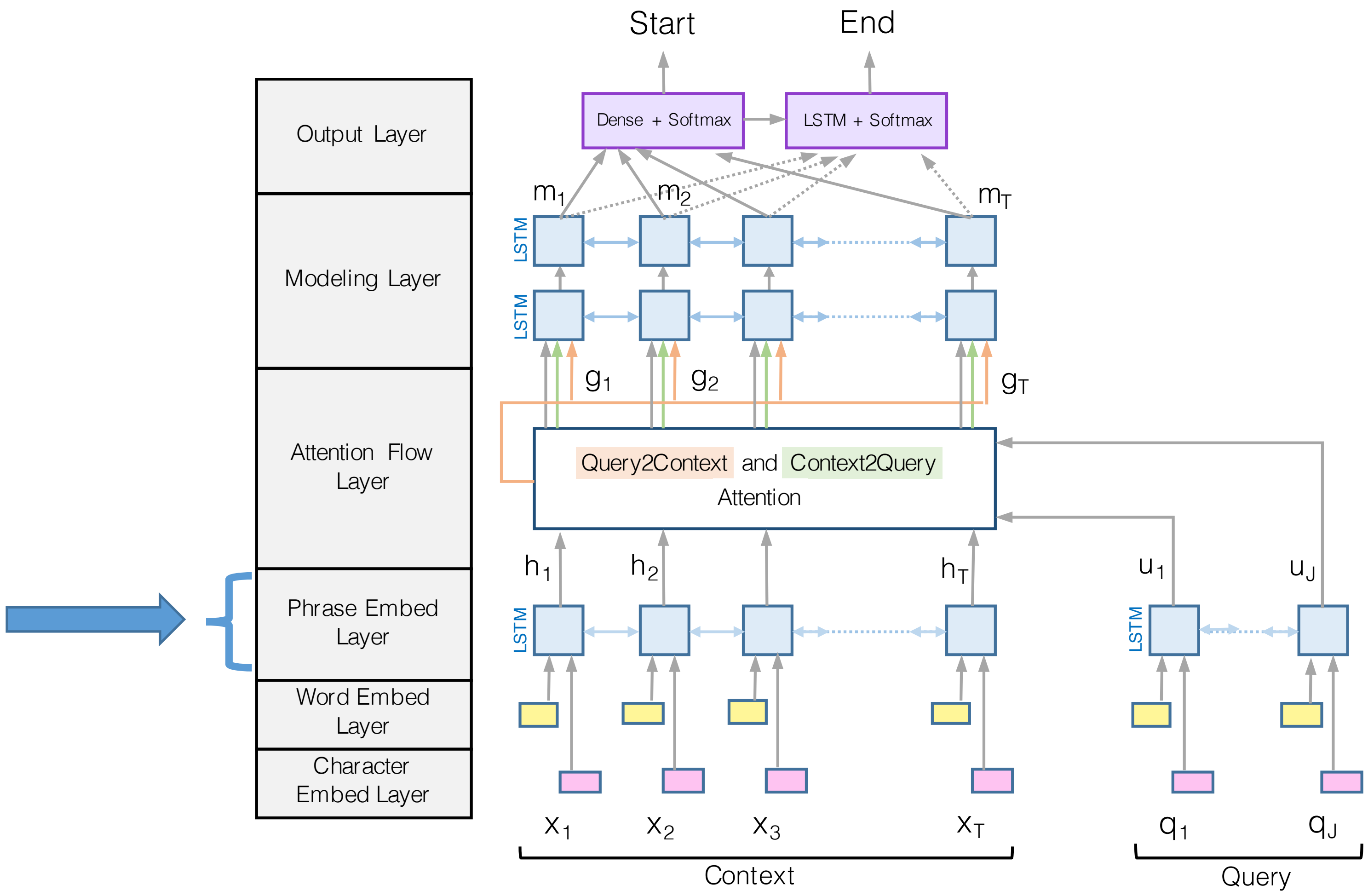
Character and Word Embedding

- Word embedding is fragile against unseen words
- Char embedding can't easily learn semantics of words
- Use both!

- Char embedding as proposed by Kim (2015)



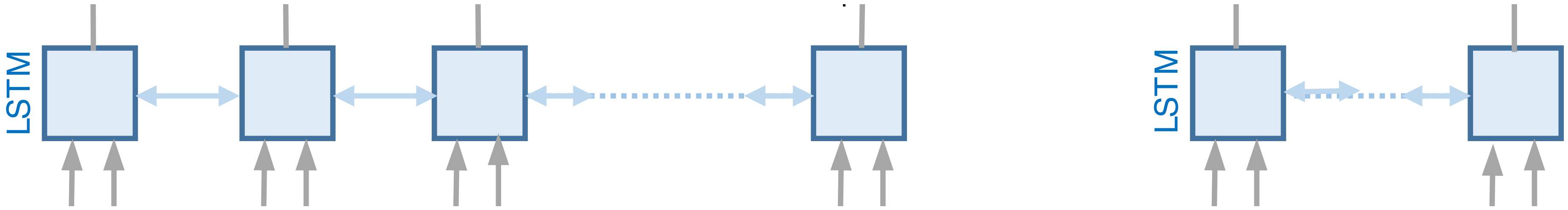
Phrase Embedding Layer



Phrase Embedding Layer

- **Inputs**: the char/word embedding of query and context words
- **Outputs**: word representations aware of their neighbors (phrase-aware words)

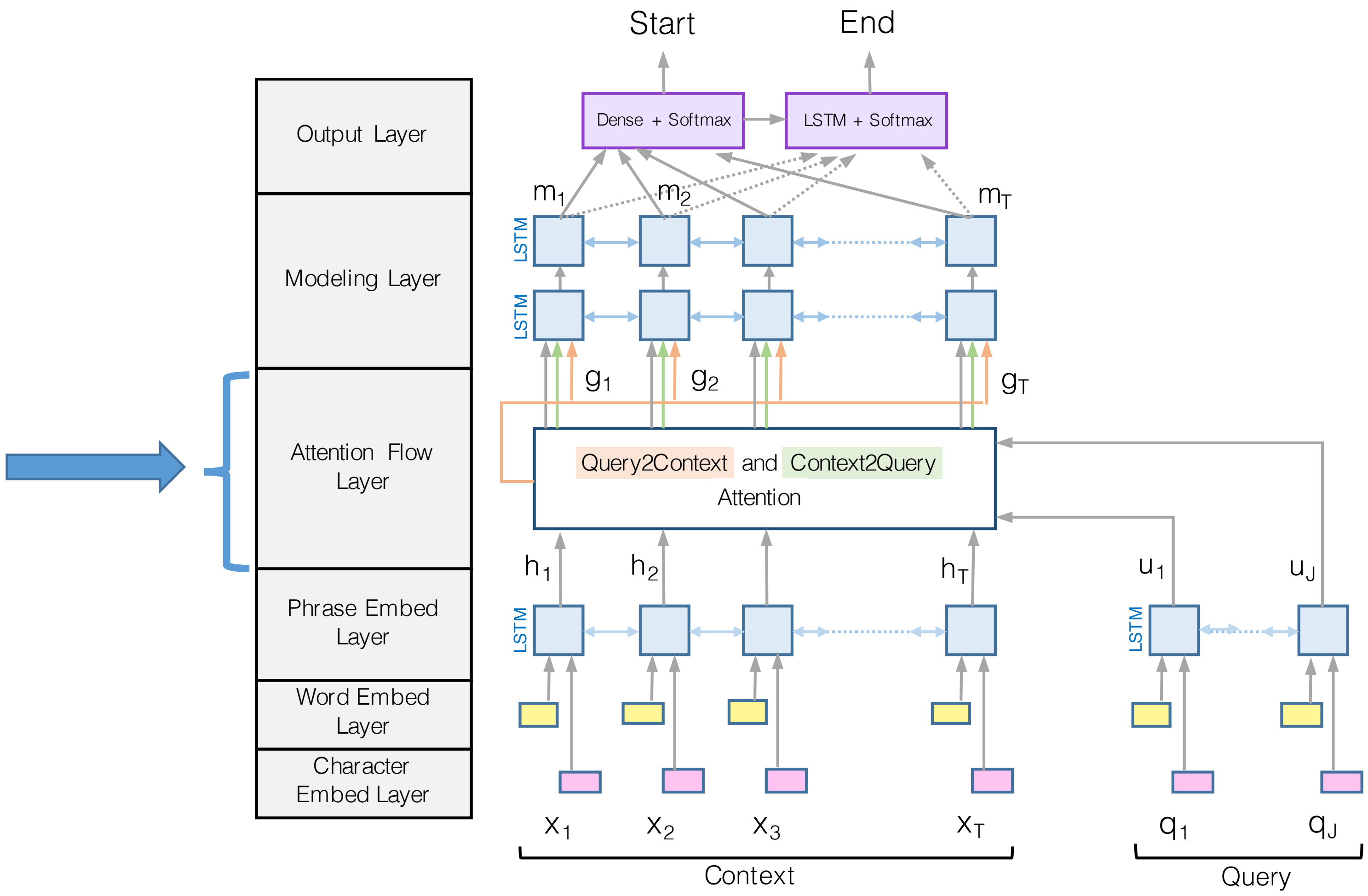
- Apply bidirectional RNN (LSTM) for both query and context



Context

Query

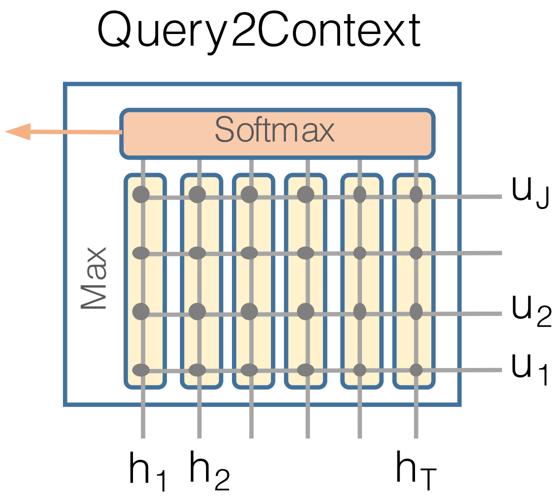
Attention Layer

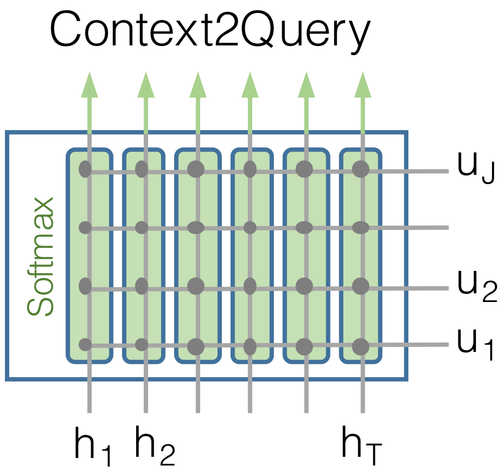


Attention Layer

- **Inputs**: phrase-aware context and query words
- **Outputs**: query-aware representations of context words

- **Context-to-query attention**: For each (phrase-aware) context word, choose the most relevant word from the (phrase-aware) query words
- **Query-to-context attention**: Choose the context word that is most relevant to any of query words.





Context-to-Query Attention (C2Q)

Q: *Who leads the United States?*

C: *Barak Obama is the president of the USA.*

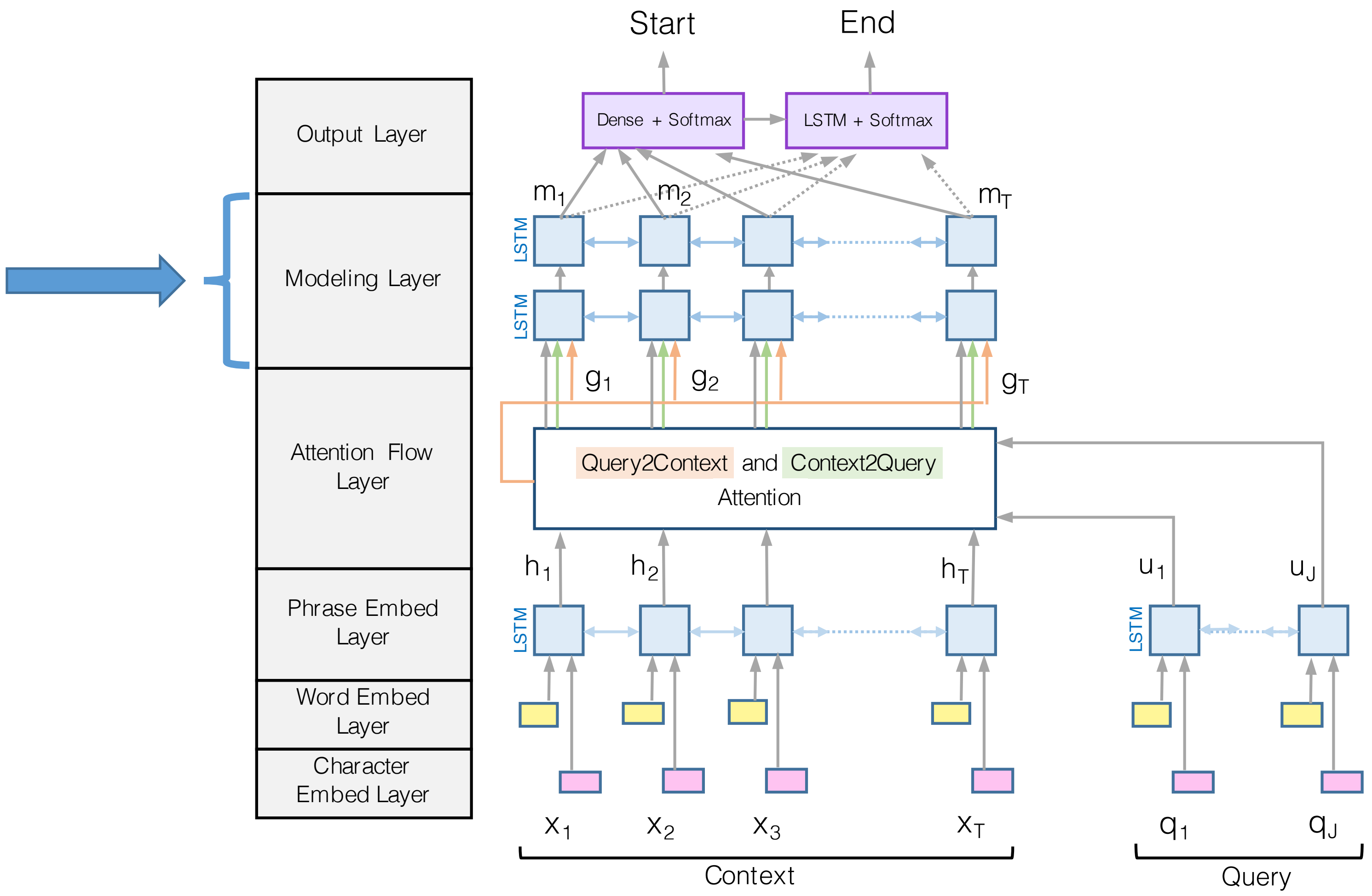
For each context word, find the most relevant query word.

Query-to-Context Attention (Q2C)

While Seattle's weather is very nice in summer, its weather is very rainy in winter, making it one of the most gloomy cities in the U.S. LA is …

Q: Which city is gloomy in winter?

Modeling Layer

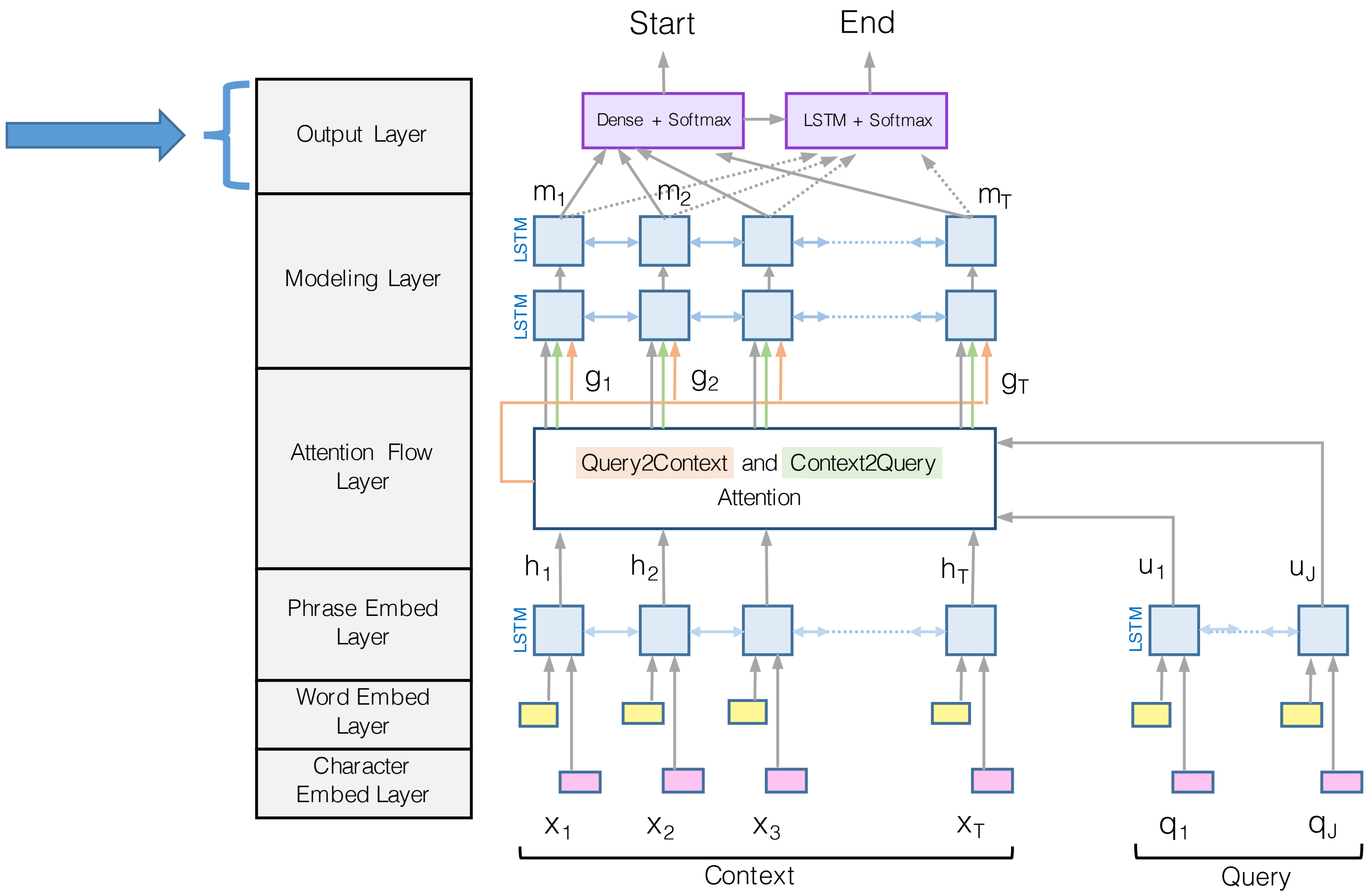


Modeling Layer

- **Attention layer**: modeling interactions between query and context
- **Modeling layer**: modeling interactions within (query-aware) context words via RNN (LSTM)

- *Division of labor*: let attention and modeling layers solely focus on their own tasks
- We experimentally show that this leads to a better result than intermixing attention and modeling

Output Layer



Training

- Minimizes the negative log probabilities of the true start index and the true end index

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

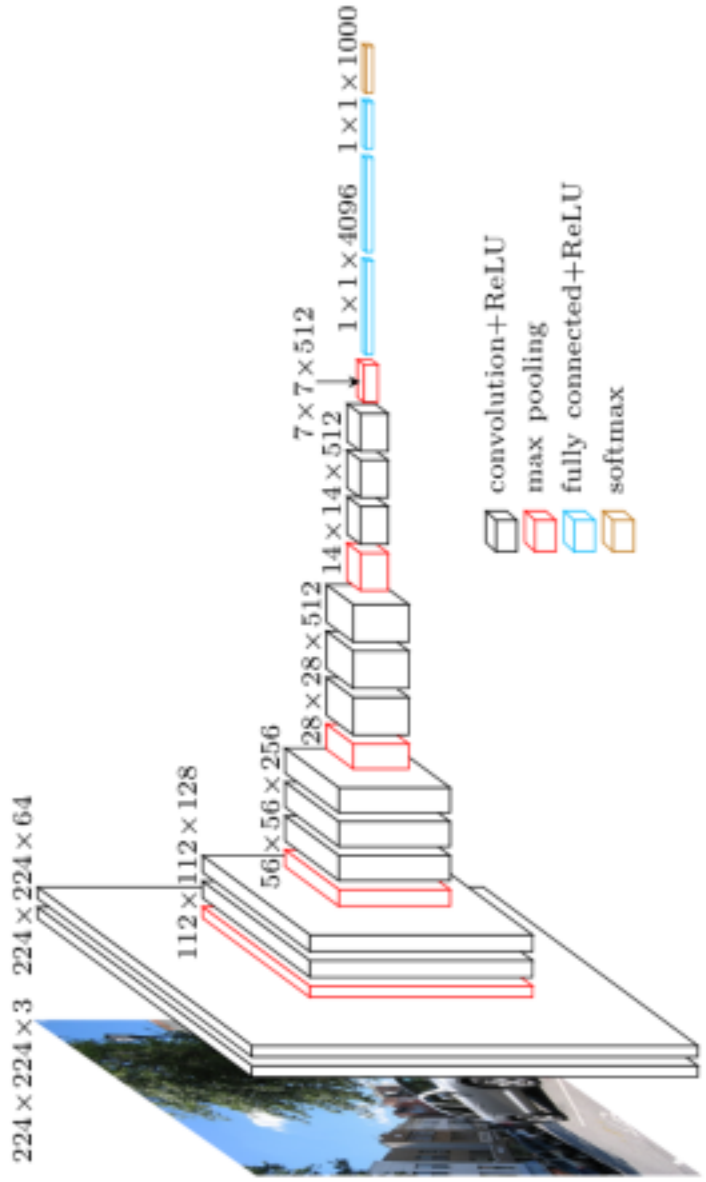
y_i^1	True start index of example i
y_i^2	True end index of example i
$\mathbf{p}^1$	Probability distribution of start index
$\mathbf{p}^2$	Probability distribution of stop index

Previous work

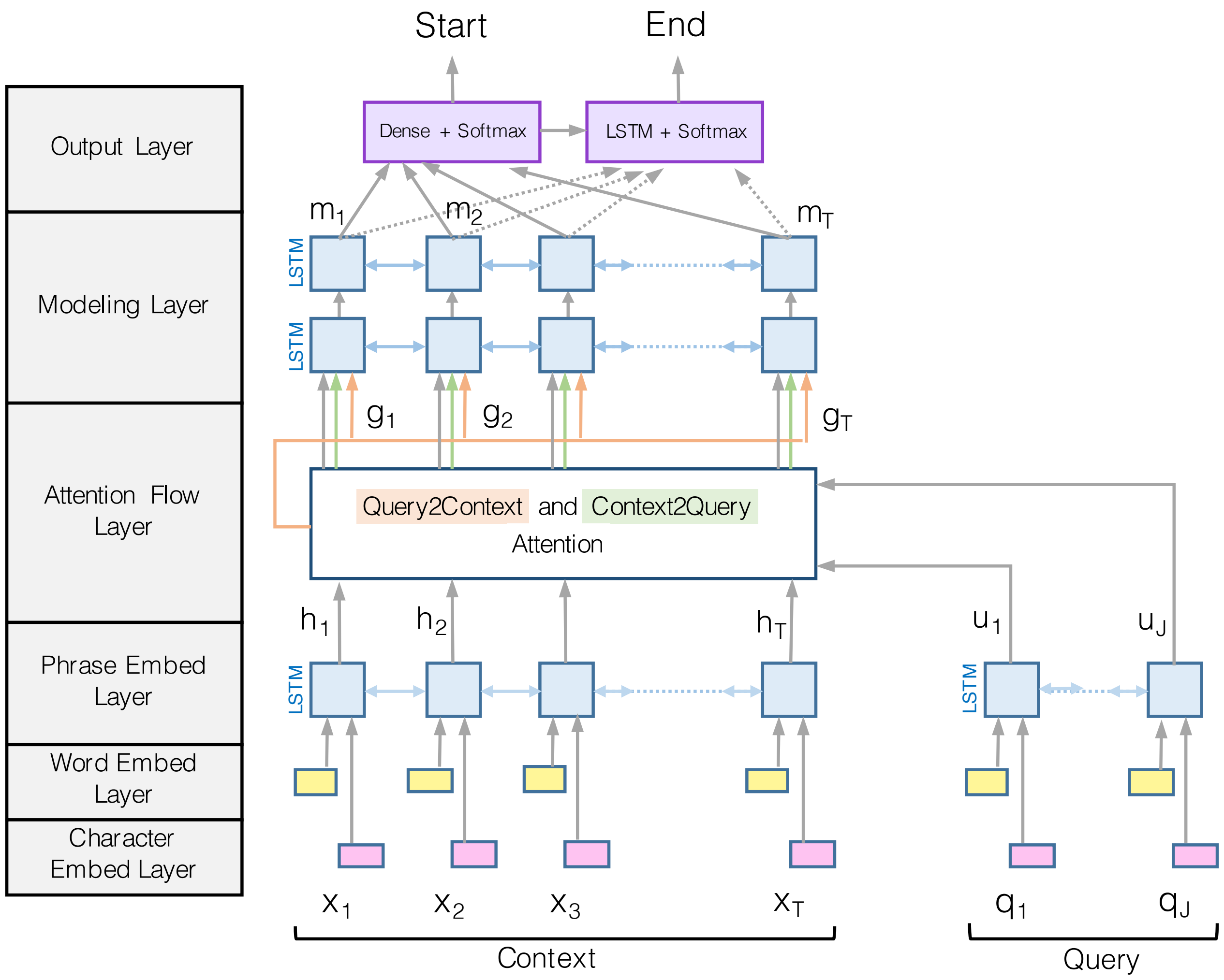
- Using neural attention as a controller (Xiong et al., 2016)
- Using neural attention within RNN (Wang & Jiang, 2016)
- Most of these attentions are uni-directional

- BiDAF (our model)
  - uses neural attention *as a layer*,
  - Is separated from modeling part (RNN),
  - Is bidirectional

Image Classifier and BiDAF



VGG-16



BiDAF (ours)

Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016)

The immune system is a system of many biological structures and processes within an organism that protects against disease. To function properly, an immune system must detect a wide variety of agents, known as pathogens, from viruses to parasitic worms, and distinguish them from the organism's own healthy tissue. In many species, the immune system can be classified into subsystems, such as the innate immune system versus the adaptive immune system, or humoral immunity versus cell-mediated immunity. In humans, the blood–brain barrier, blood–cerebrospinal fluid barrier, and similar fluid–brain barriers separate the peripheral immune system from the neuroimmune system which protects the brain.

What is the immune system?
Answer 1: a system of many biological structures and processes within an organism that protects against disease
Answer 2: system of many biological structures and processes
Answer 3: a system of many biological structures and processes within an organism
Answer 4: a system of many biological structures and processes within an organism

- Most popular articles from Wikipedia
- Questions and answers from Turkers
- 90k train, 10k dev, ? test (hidden)
- Answer must lie in the context
- Two metrics: Exact Match (**EM**) and **F1**

SQuAD Results (http://stanford-qa.com) as of 12pm Today

Test Set Leaderboard

Since the release of our dataset (and paper), the community has made rapid progress! Here are the ExactMatch (EM) and F1 scores of the best models evaluated on the test and development sets of v1.1.

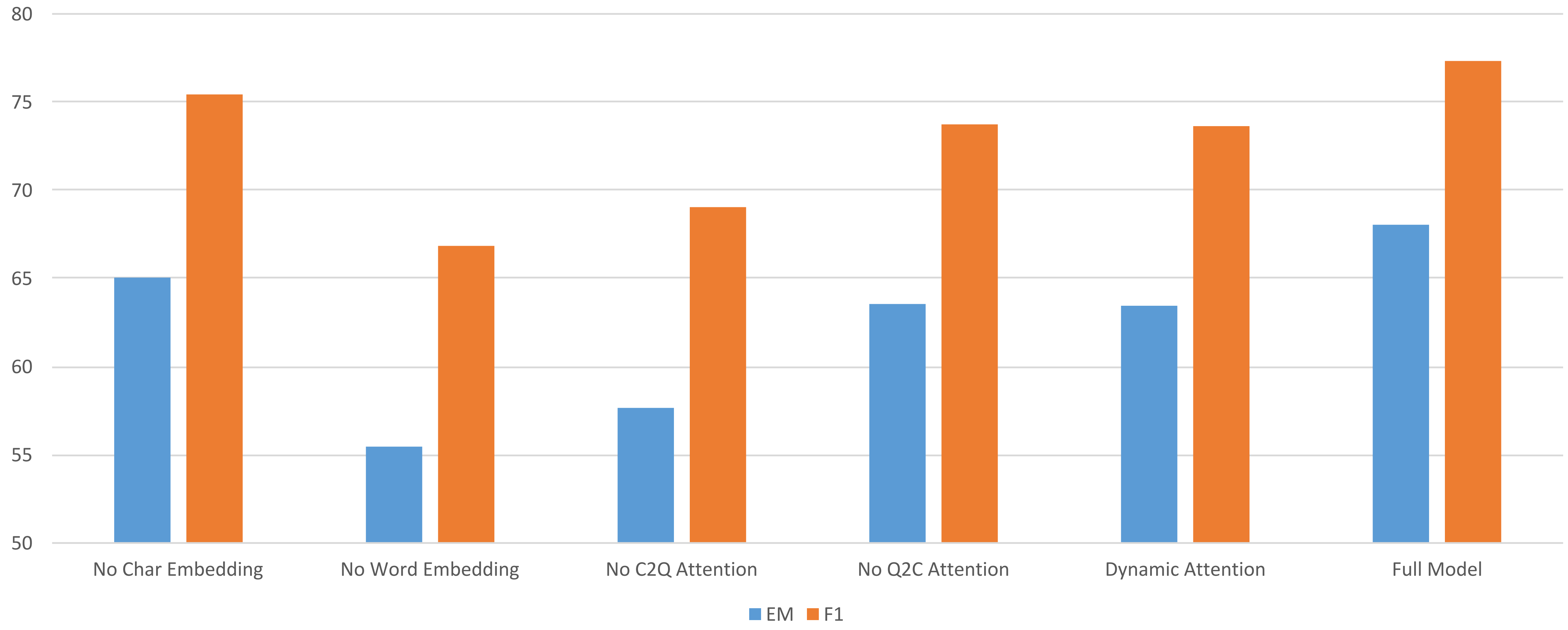
Rank	Model	Test EM	Test F1
1	BiDAF (ensemble) Allen Institute for AI & University of Washington (Seo et al. '16)	73.3	81.1
2	Dynamic Coattention Networks (ensemble) Salesforce Research (Xiong & Zhong et al. '16)	71.6	80.4
2	r-net (ensemble) Microsoft Research Asia	72.1	79.7
4	r-net (single model) Microsoft Research Asia	68.4	77.5
5	BiDAF (single model) Allen Institute for AI & University of Washington (Seo et al. '16)	68.0	77.3
5	Multi-Perspective Matching (ensemble) IBM Research	68.2	77.2

SQuAD Results

	EM	F1
Stanford[1] (baseline)	40.4	51.0
IBM[2]	62.5	71.0
CMU[3]	62.5	73.3
Singapore Management[4] (ensemble)	67.9	77.0
IBM Research (ensemble)	68.2	77.2
Salesforce Research[6] (ensemble)	71.6	80.4
Microsoft Research Asia (ensemble)	72.1	79.7
Ours (ensemble)	**73.3**	**81.1**

1: Rajpurkar et al. (2016)
2: Yu et al. (2016)
3: Yang et al. (2016)
4: Wang & Jiang (2016)
6: Xiong et al. (2016)

Ablations on dev data



Interactive Demo

http://allenai.github.io/bi-att-flow/demo

Attention Visualizations

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 .

at, the, at, Stadium, Levi, in, Santa, Ana

[]

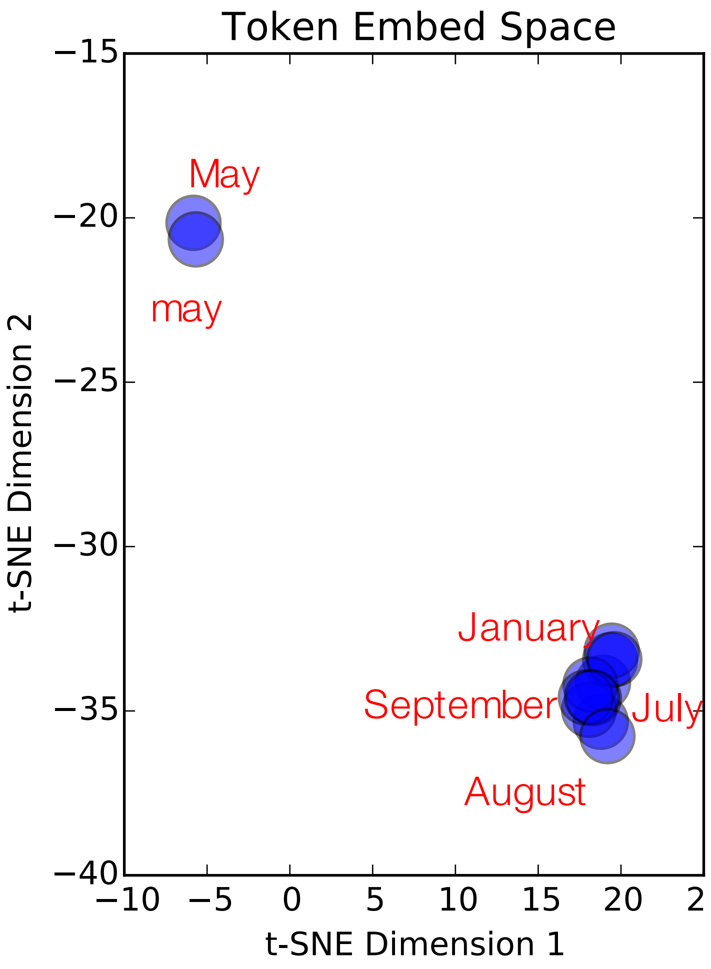
Super, Super, Super, Super, Super

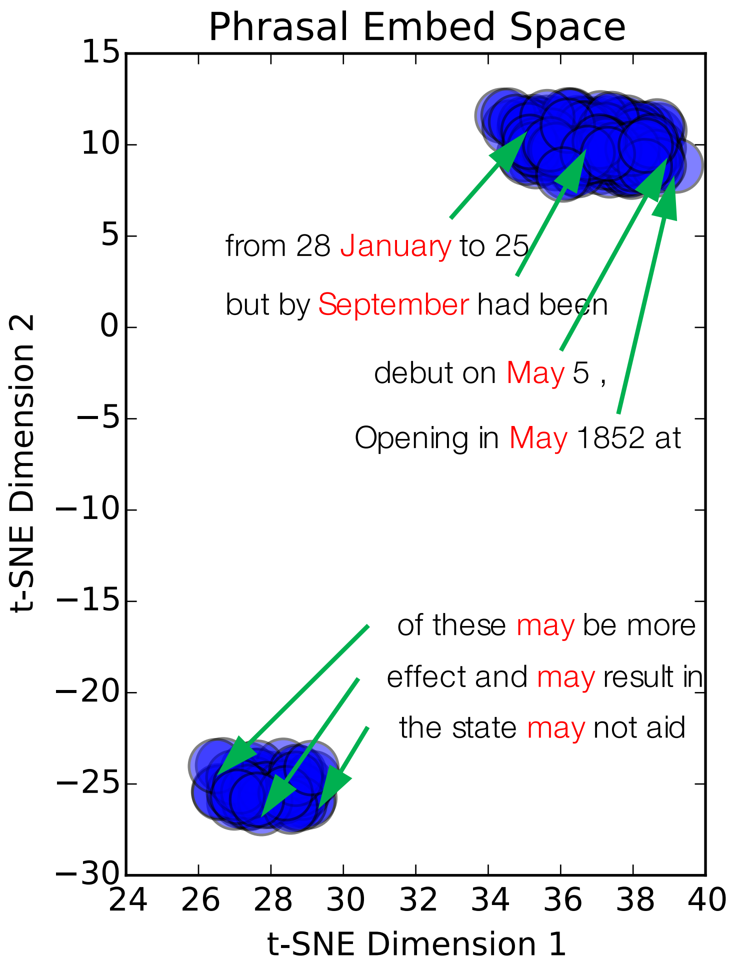
Bowl, Bowl, Bowl, Bowl, Bowl

50

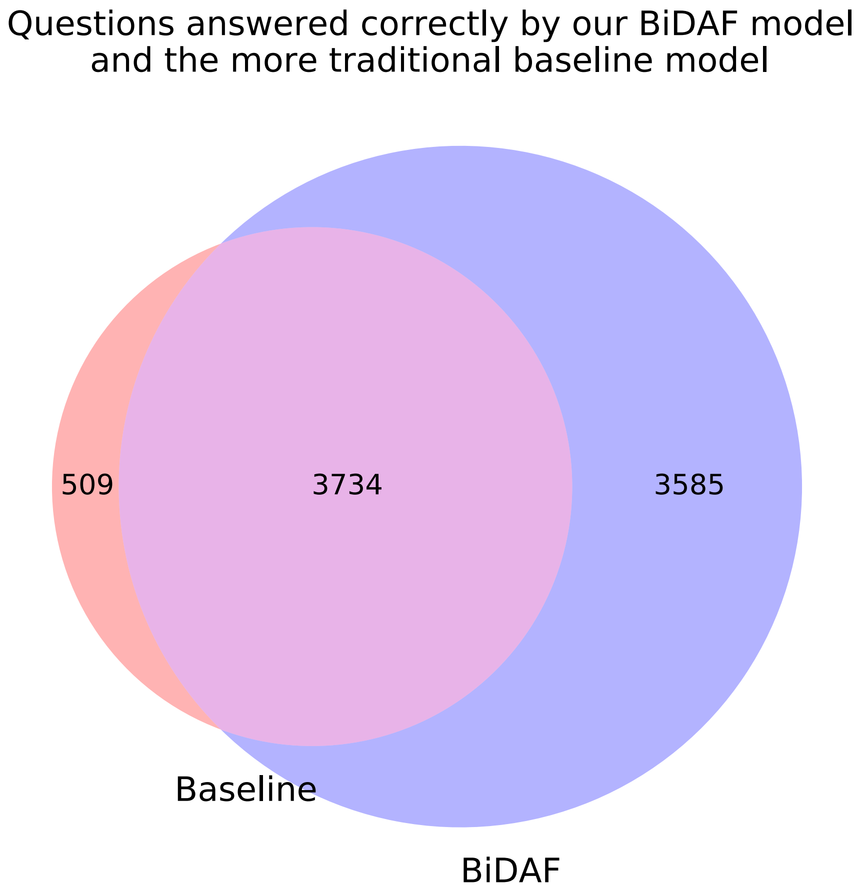
initiatives

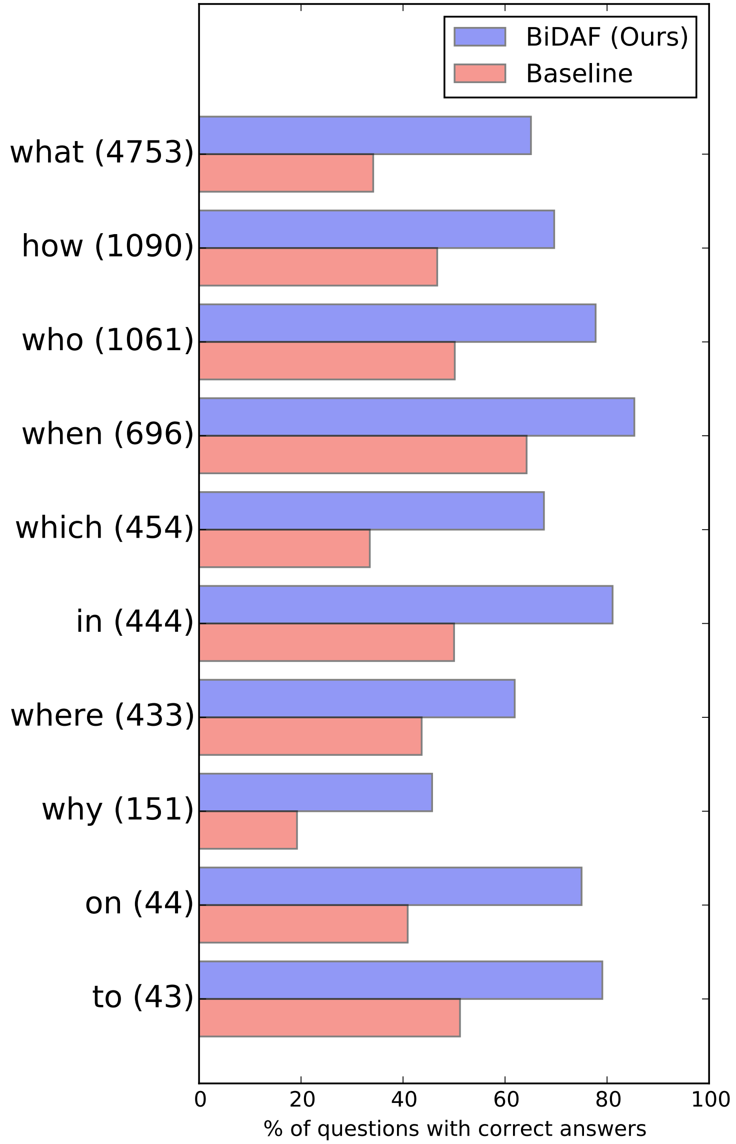
Embedding Visualization at Word vs Phrase Layers





How does it compare with feature-based models?





CNN/DailyMail Cloze Test (Hermann et al., 2015)

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

- Cloze Test (Predicting Missing words)
- Articles from CNN/DailyMail
- Human-written summaries
- Missing words are always entities
- CNN – 300k article-query pairs
- DailyMail – 1M article-query pairs

CNN/DailyMail Cloze Test Results

	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
Stanford AR (Chen 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	-	-
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

Some limitations of SQuAD

Reasoning	Description	Example	Percentage
Lexical variation (synonymy)	Major correspondences between the question and the answer sentence are synonyms.	Q: What is the Rankine cycle sometimes **called**? Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle.	33.3%
Lexical variation (world knowledge)	Major correspondences between the question and the answer sentence require world knowledge to resolve.	Q: Which **governing bodies** have veto power? Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process.	9.1%
Syntactic variation	After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications.	Q: What Shakespeare scholar **is currently on the faculty**? Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.	64.1%
Multiple sentence reasoning	There is anaphora, or higher-level fusion of multiple sentences is required.	Q: What collection does **the V&A Theatre & Performance galleries** hold? Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance.	13.6%
Ambiguous	We don't agree with the crowd-workers' answer, or the question does not have a unique answer.	Q: What is the main goal of criminal punishment? Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment.	6.1%

Two Question Answering Systems with Neural Attention

- Bidirectional Attention Flow (BiDAF)
  - On Stanford Question Answering Dataset and CNN/DailyMail Cloze Test

- Query-Reduction Networks (QRN)
  - On bAbI QA and dialog datasets

Reasoning Question Answering

Task 1: Single Supporting Fact
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 3: Three Supporting Facts
John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

Task 4: Two Argument Relations
The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom

Dialog System

U: Can you book a table in Rome in Italian Cuisine

S: How many people in your party?

U: For four people please.

S: What price range are you looking for?

Dialog task vs QA

- Dialog system can be considered as QA system:
  - Last user’s utterance is the query
  - All previous conversations are context to the query
  - The system’s next response is the answer to the query
- Poses a few unique challenges
  - Dialog system requires tracking states
  - Dialog system needs to look at multiple sentences in the conversation
  - Building end-to-end dialog system is more challenging

Our approach: Query-Reduction

Reduced query:

<START> — *Where is the apple?*

Sandra got the apple there. — *Where is Sandra?*

Sandra dropped the apple. — *Where is Sandra?*

Daniel took the apple there. — *Where is Daniel?*

Sandra went to the hallway. — *Where is Daniel?*

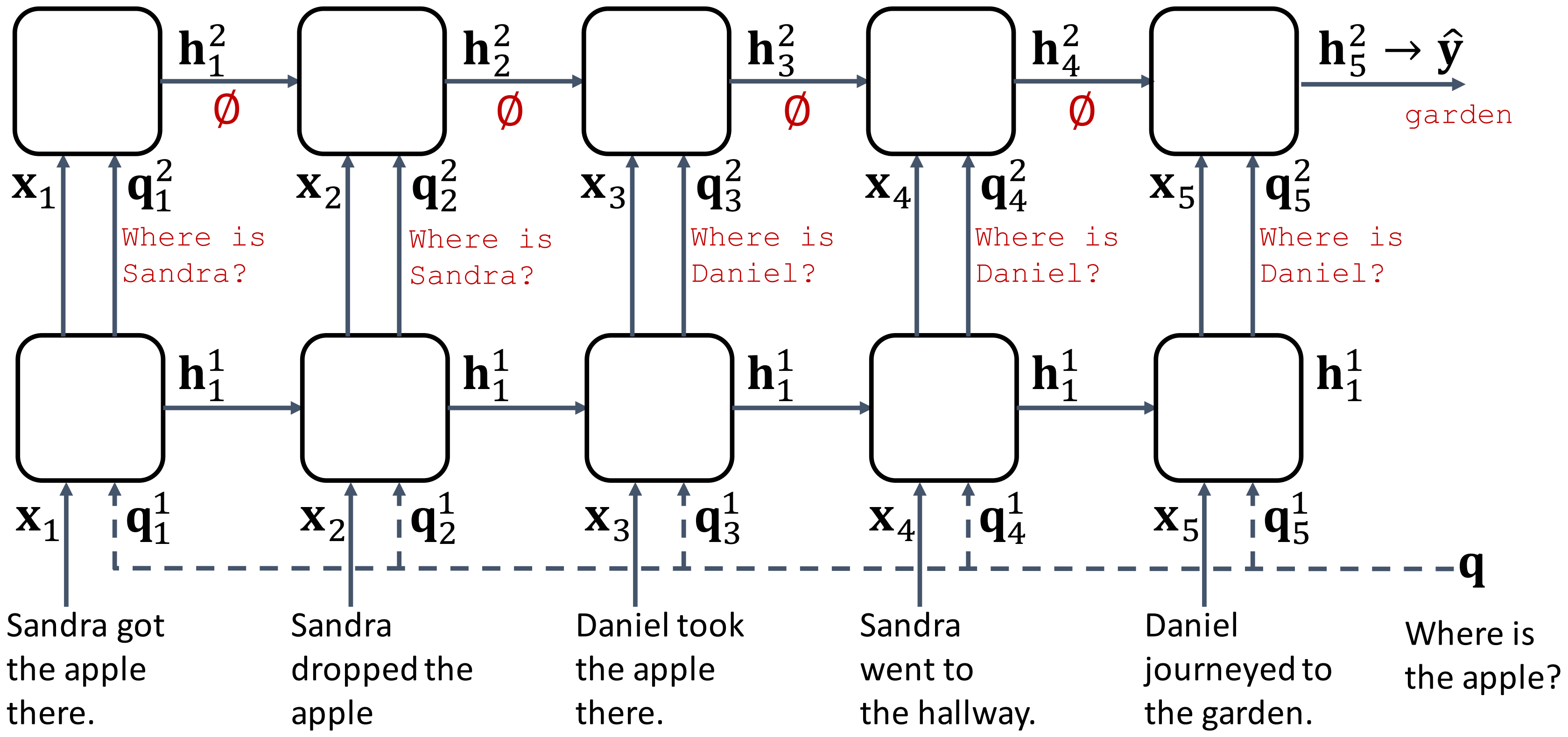
Daniel journeyed to the garden. — *Where is Daniel? → garden*

Q: Where is the apple?

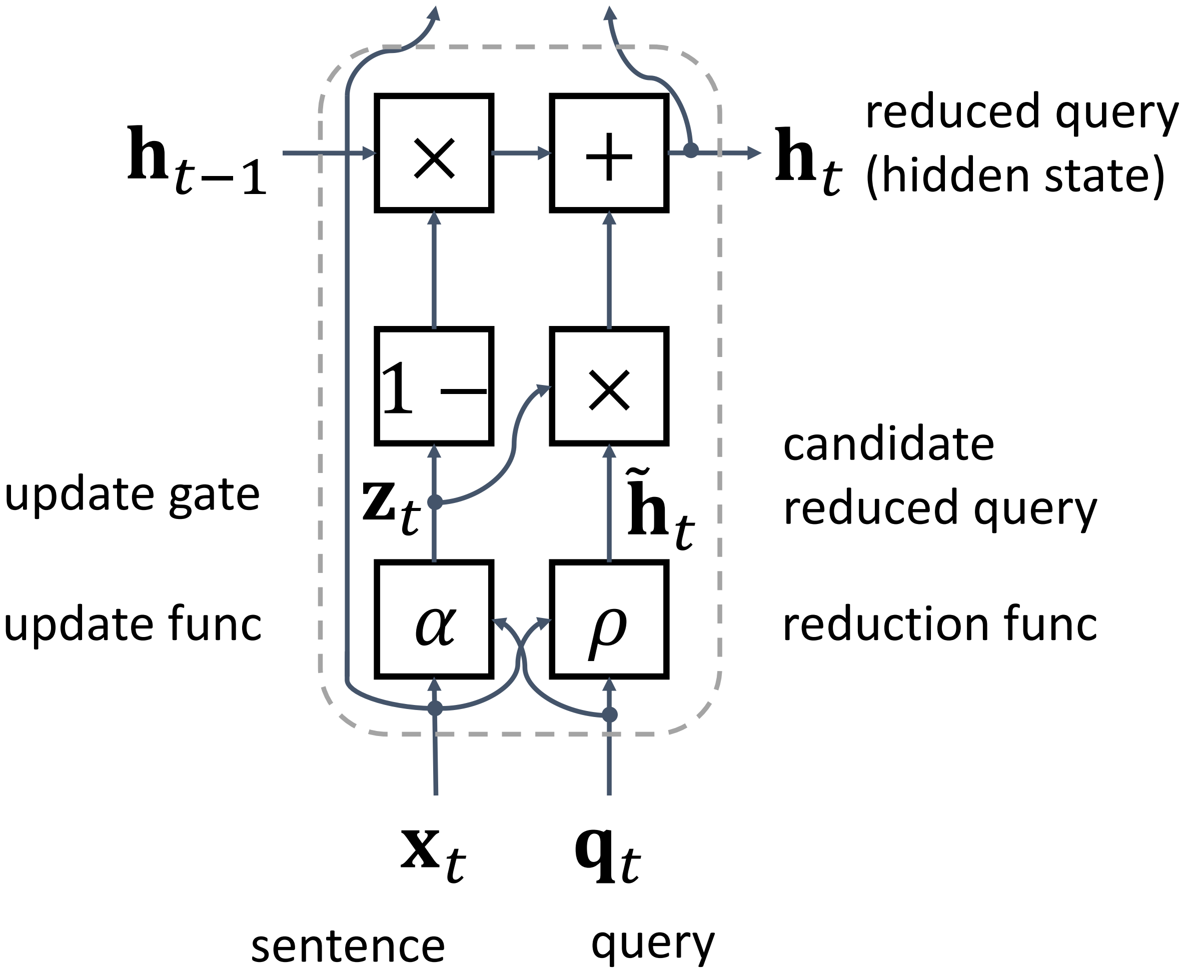
A: garden

Query-Reduction Networks

- Reduce the query into an easier-to-answer query over the sequence of state-changing triggers (sentences), *in vector space*



QRN Cell



$$z_t = \alpha(\mathbf{x}_t, \mathbf{q}_t)$$

$$\tilde{\mathbf{h}}_t = \boldsymbol{\rho}(\mathbf{x}_t, \mathbf{q}_t)$$

$$\mathbf{h}_t = z_t \tilde{\mathbf{h}}_t + (1 - z_t)\mathbf{h}_{t-1}$$

Characteristics of QRN

- Update gate can be considered as local attention
  - QRN chooses to consider / ignore each candidate reduced query
  - The decision is made locally (as opposed to global softmax attention)
- Subclass of Recurrent Neural Network (RNN)
  - Two inputs, hidden state, gating mechanism
  - Able to handle sequential dependency (attention cannot)
- Simpler recurrent update enables *parallelization* over time
  - Candidate hidden state (reduced query) is computed from inputs only
  - Hidden state can be explicitly computed as a function of inputs

Parallelization

$$z_t = \alpha(\mathbf{x}_t, \mathbf{q}_t)$$

$$\tilde{\mathbf{h}}_t = \boldsymbol{\rho}(\mathbf{x}_t, \mathbf{q}_t)$$

computed from inputs only, so can be trivially parallelized

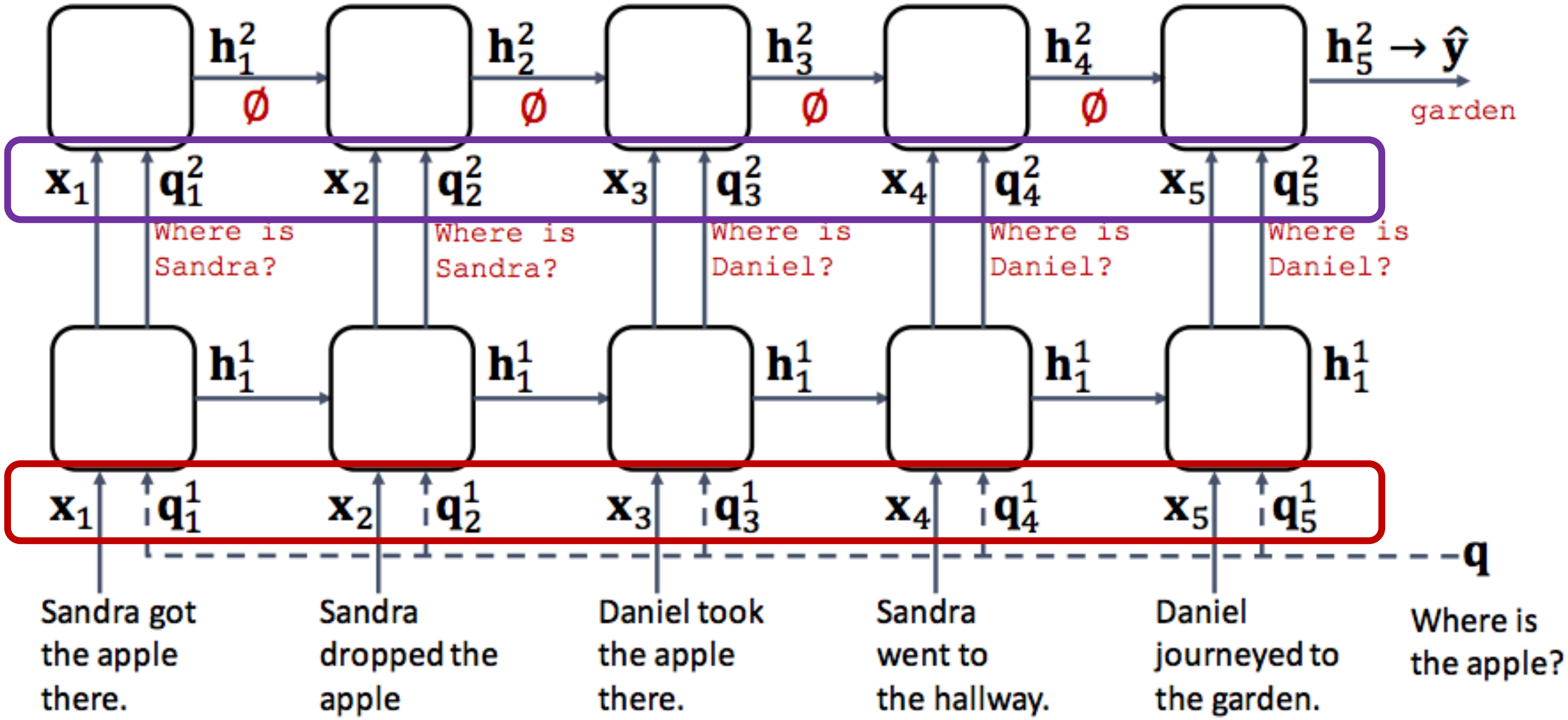
$$\mathbf{h}_t = z_t \tilde{\mathbf{h}}_t + (1 - z_t)\mathbf{h}_{t-1}$$

$$\Downarrow$$

$$\mathbf{h}_t = \sum_{i=1}^{t} \left[\prod_{j=i+1}^{t} 1 - z_j \right] z_i \tilde{\mathbf{h}}_i$$

Can be explicitly expressed as the geometric sum of previous candidate hidden states

Parallelization



Characteristics of QRN

- Update gate can be considered as local attention
- Subclass of Recurrent Neural Network (RNN)
- Simpler recurrent update enables *parallelization* over time

QRN sits between neural attention mechanism and recurrent neural networks, taking the advantage of both paradigms.

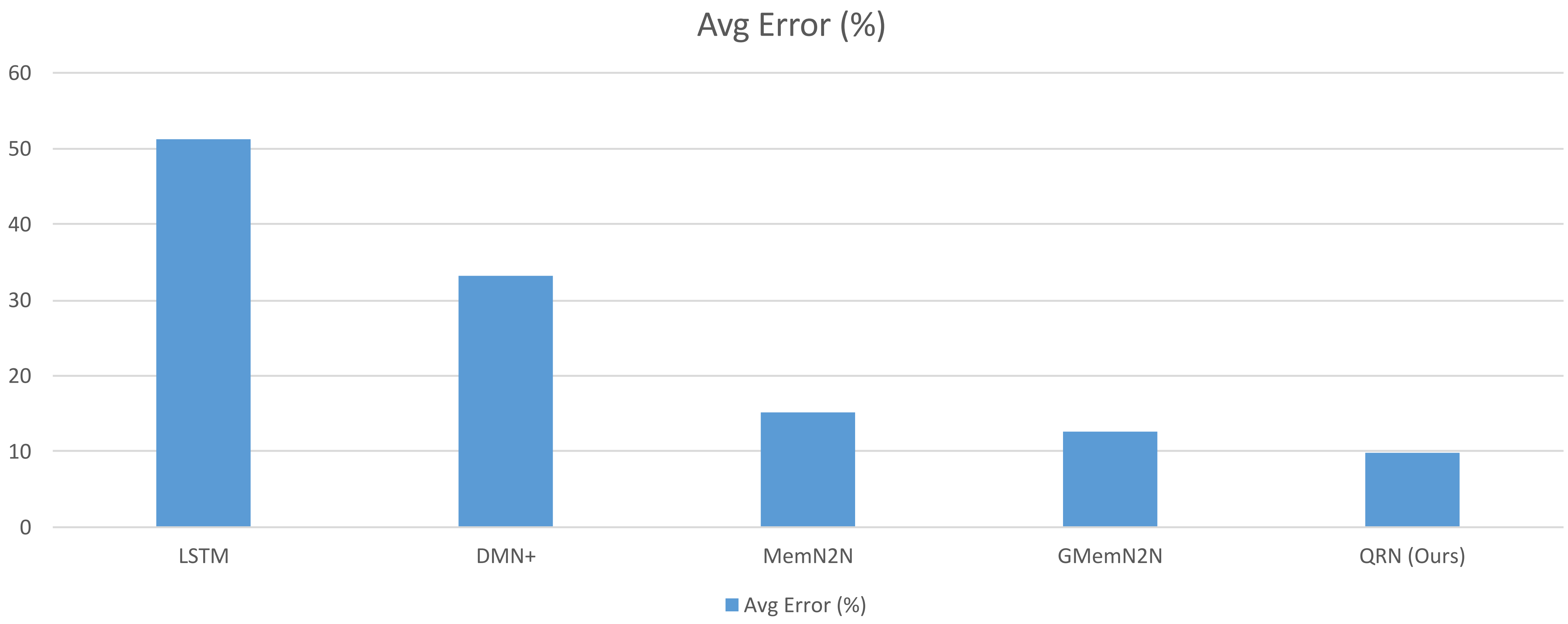
bAbI QA Dataset

- 20 different tasks
- 1k story-question pairs for each task (10k also available)
- Synthetically generated
- Many questions require looking at multiple sentences
- For end-to-end system supervised by answers only

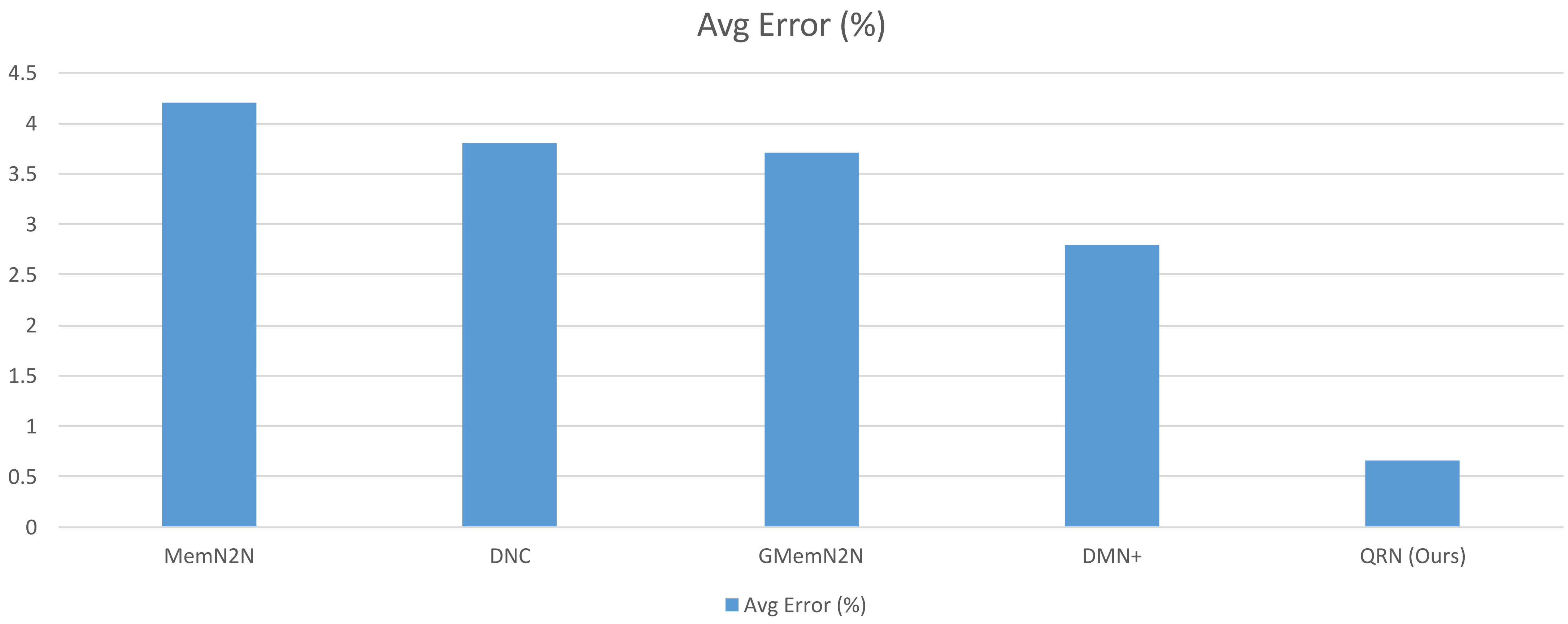
What's different from SQuAD?

- Synthetic
- More than lexical / syntactic understanding
- Different kinds of inferences
  - induction, deduction, counting, path finding, etc.
- Reasoning over multiple sentences
- Interesting testbed towards developing complex QA system (and dialog system)

bAbI QA Results (1k)



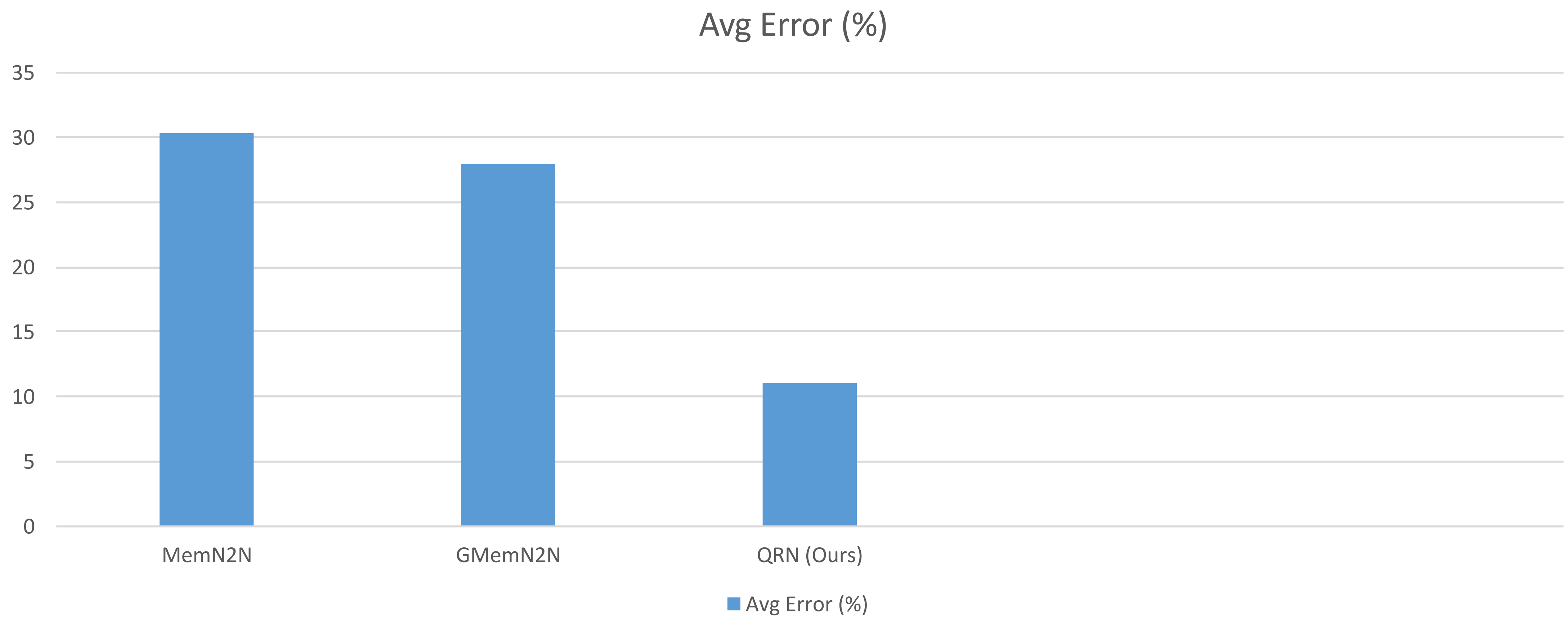
bAbI QA Results (10k)



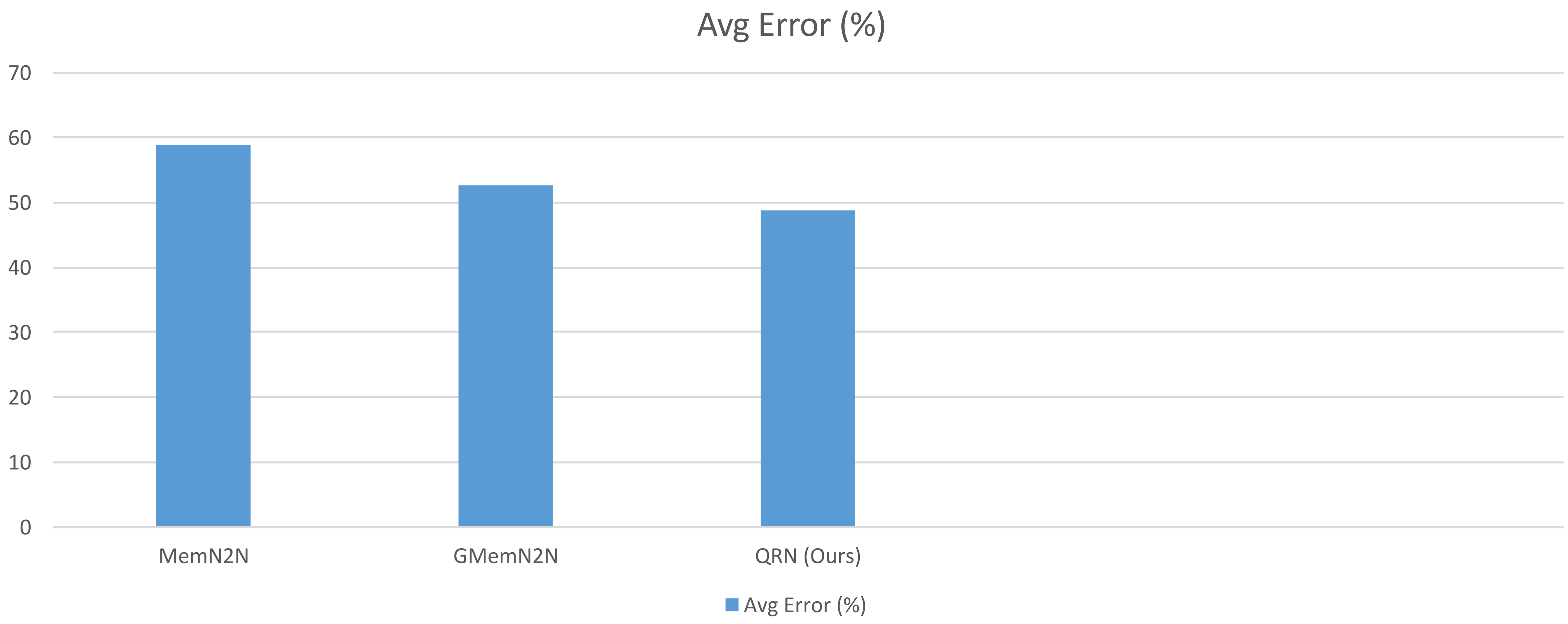
Dialog Datasets

- bAbI Dialog Dataset
  - Synthetic
  - 5 different tasks
  - 1k dialogs for each task
- DSTC2* Dataset
  - Real dataset
  - Evaluation metric is different from original DSTC2: response generation instead of “state-tracking”
  - Each dialog is 800+ utterances
  - 2407 possible responses

bAbI Dialog Results (OOV)



DSTC2* Dialog Results



bAbI QA Visualization

	Layer 1			Layer 2
Task 2: Two Supporting Facts	z^1	$\overrightarrow{r}^1$	$\overleftarrow{r}^1$	z^2
Sandra picked up the apple there.	0.95	0.89	0.98	0.00
Sandra dropped the apple.	0.83	0.05	0.92	0.01
Daniel grabbed the apple there.	0.88	0.93	0.98	0.00
Sandra travelled to the bathroom.	0.01	0.18	0.63	0.02
Daniel went to the hallway.	0.01	0.24	0.62	0.83
Where is the apple?	hallway			

z^l = Local attention (update gate) at layer *l*

DSTC2 (Dialog) Visualization

	Layer 1			Layer 2
Task 6 DSTC2 dialog	z^1	$\overrightarrow{r}^1$	$\overleftarrow{r}^1$	z^2
Spanish food.	0.84	0.07	0.00	0.82
You are lookng for a spanish restaurant right?	0.98	0.02	0.49	0.75
Yes.	0.01	1.00	0.33	0.13
What part of town do you have in mind?	0.20	0.73	0.41	0.11
I don't care.	0.00	1.00	0.02	0.00
What price range would you like?	0.72	0.46	0.52	0.72
I don't care.	API CALL spanish R-location R-price			

$z^l =$ Local attention (update gate) at layer *l*

Conclusion

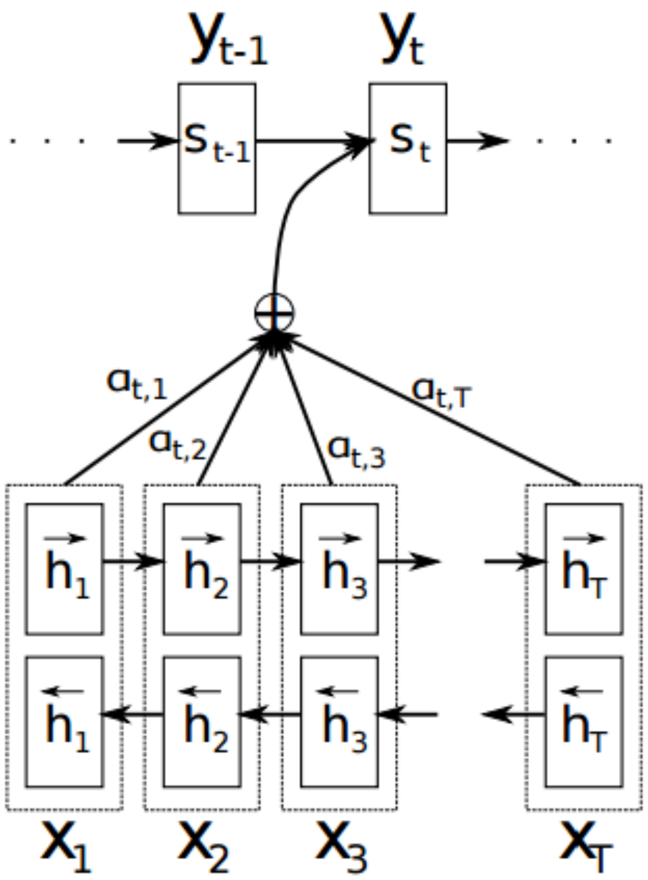
- Presented two novel approaches for QA tasks using neural attention
- **Bidirectional Attention Flow**: using attention as a layer, on both directions (context to query, query to context)
- **Query-reduction Networks**: a sequential model that takes advantage of both attention and RNN for reasoning over multiple sentences

Thanks!

Why do we need attention?

- RNN has long-term dependency problem
  - Vanishing gradients (Pascanu et al., 2013)
  - Inherently unstable over a long period of time (Weston et al., 2016)
- Attention provides shortcut access to relevant information
  - Directly retrieves the context vector from a distant location
- Critical to most modern sequence models
  - Machine translation
  - Question answering, machine comprehension

Neural Attention in Sequence Modeling



(Bahdanau et al., 2015)

- Apply RNN on context vectors
- Apply RNN on query vectors
  - At each time step, use neural attention to soft-select a single context vector
  - Use the selected context vector, along with current query vector and current hidden state, to obtain the next hidden state