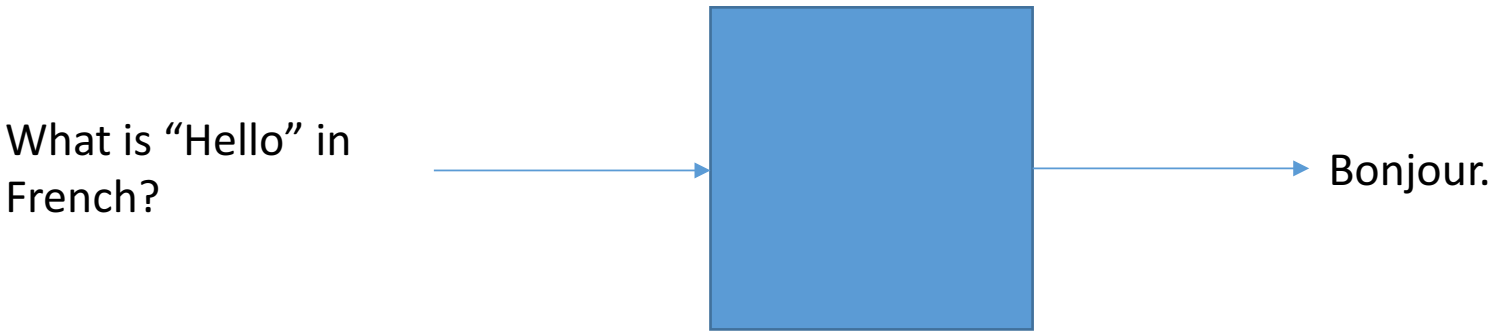
Learning to reason by reading text and answering questions

Minjoon Seo
Natural Language Processing Group
University of Washington
May 26, 2017
@ Kakao Brain

What is reasoning?

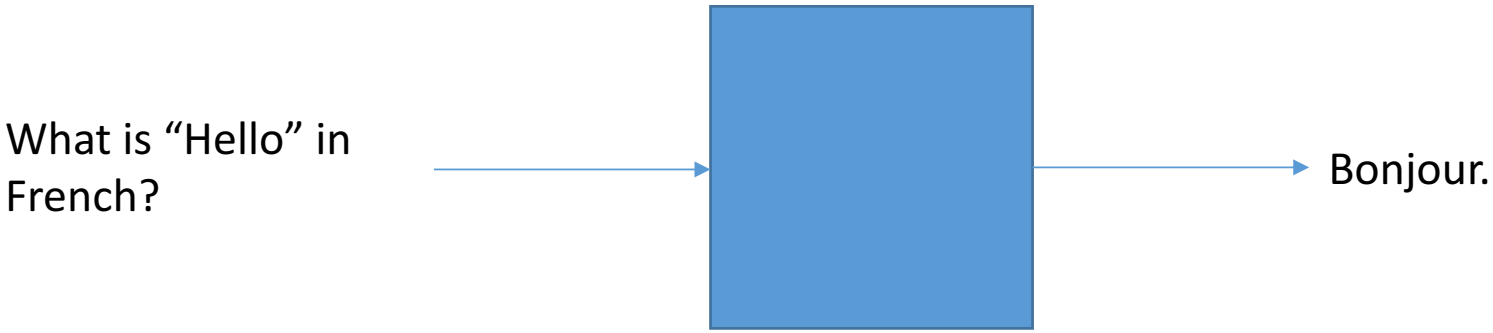
Simple Question Answering Model



Examples

- Most neural machine translation systems (Cho et al., 2014; Bahdanau et al. , 2014)
  - Need very high hidden state size (~1000)
  - No need to query the database (context) → very fast
- Most dependency, constituency parser (Chen et al., 2014; Klein et al., 2003)
- Sentiment classification (Socher et al., 2013)
  - Classifying whether a sentence is positive or negative
- Most neural image classification systems
  - The question is always “What is in the image?”
- Most classification systems

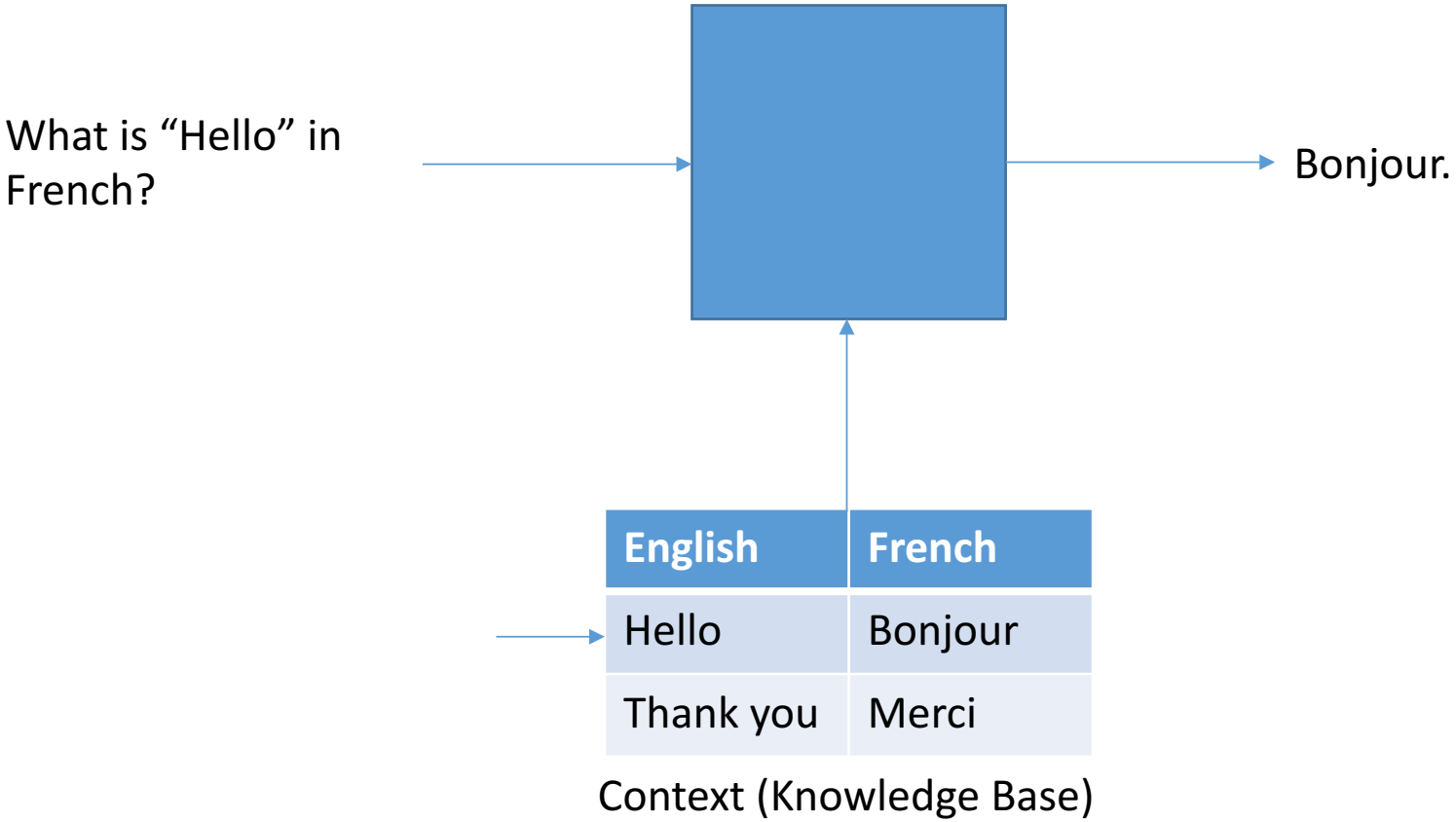
Simple Question Answering Model



Problem: parametric model has finite capacity.

“You can’t even fit a sentence into a single vector” -Dan Roth

QA Model with Context

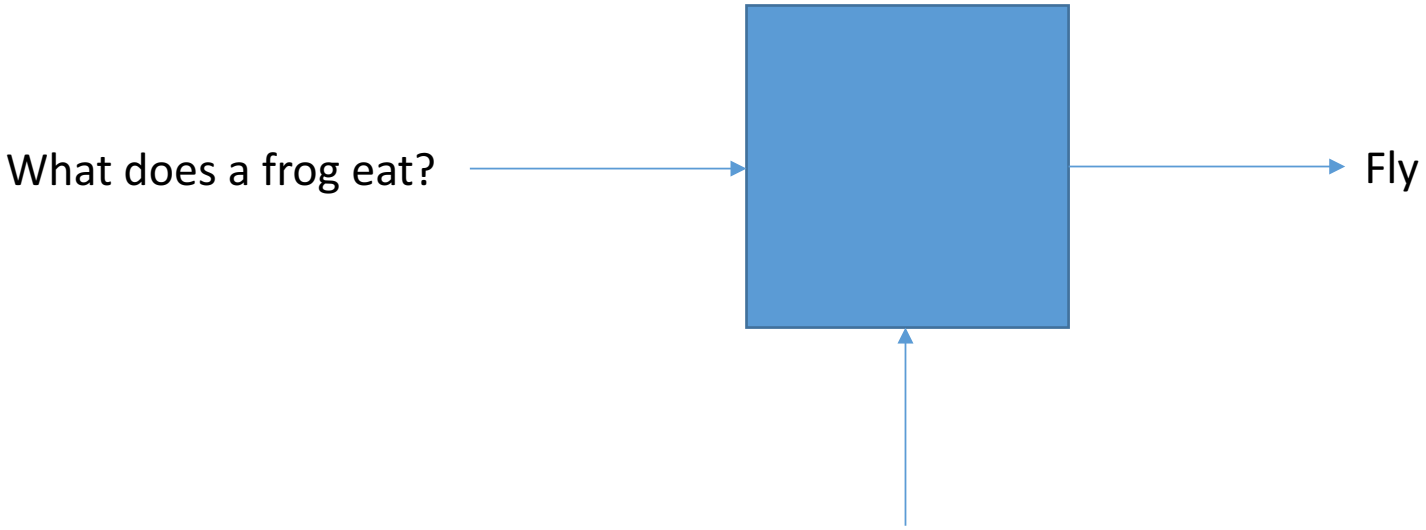


Examples

- Wiki QA (Yang et al., 2015)
- QA Sent (Wang et al., 2007)
- WebQuestions (Berant et al., 2013)
- WikiAnswer (Wikia)
- Free917 (Cai and Yates, 2013)

- Many deep learning models with external memory (e.g. Memory Networks)

QA Model with Context

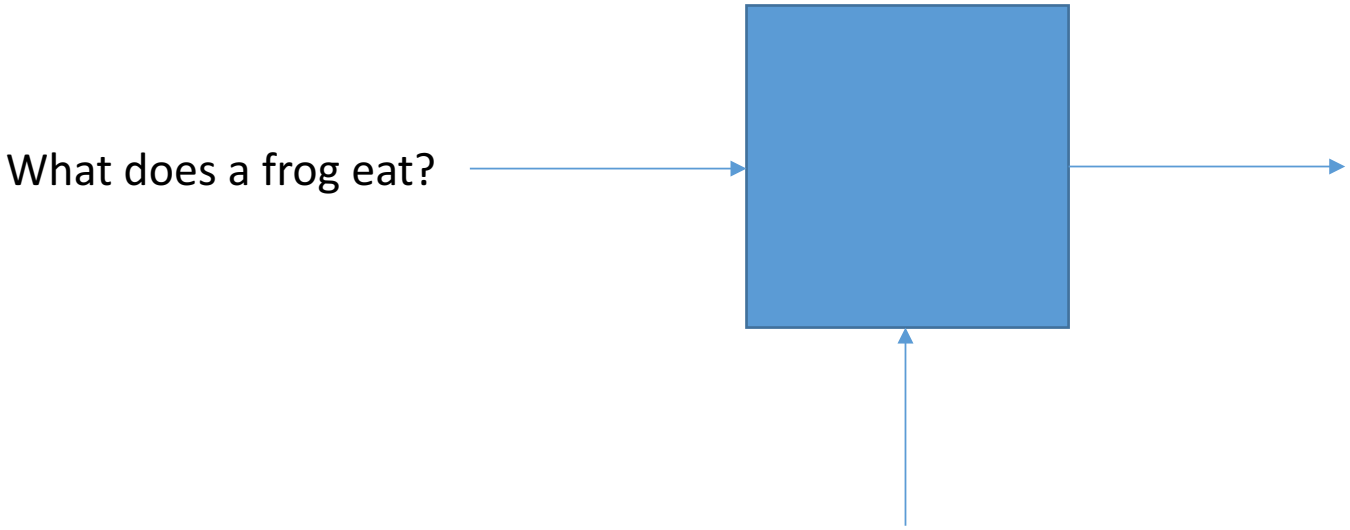


Eats	IsA
(Amphibian, insect)	(Frog, amphibian)
(insect, flower)	(Fly, insect)

Context (Knowledge Base)

Something is missing …

QA Model with Reasoning Capability



Eats	IsA	First Order Logic
(Amphibian, insect)	(Frog, amphibian)	IsA(A, B) ^ IsA(C, D) ^ Eats(B, D) → Eats(A, C)
(insect, flower)	(Fly, insect)	

Context (Knowledge Base)

Examples

- Semantic parsing
  - GeoQuery (Krishnamurthy et al., 2013; Artzi et al., 2015)
- Science questions
  - Aristo Challenge (Clark et al., 2015)
  - ProcessBank (Berant et al., 2014)
- Machine comprehension
  - MCTest (Richardson et al., 2013)

“Vague” line between non-reasoning QA and reasoning QA

- Non-reasoning:
  - The required information is explicit in the context
  - The model often needs to handle lexical / syntactic variations
- Reasoning:
  - The required information may *not* be explicit in the context
  - Need to combine multiple facts to derive the answer
- There is no clear line between the two!

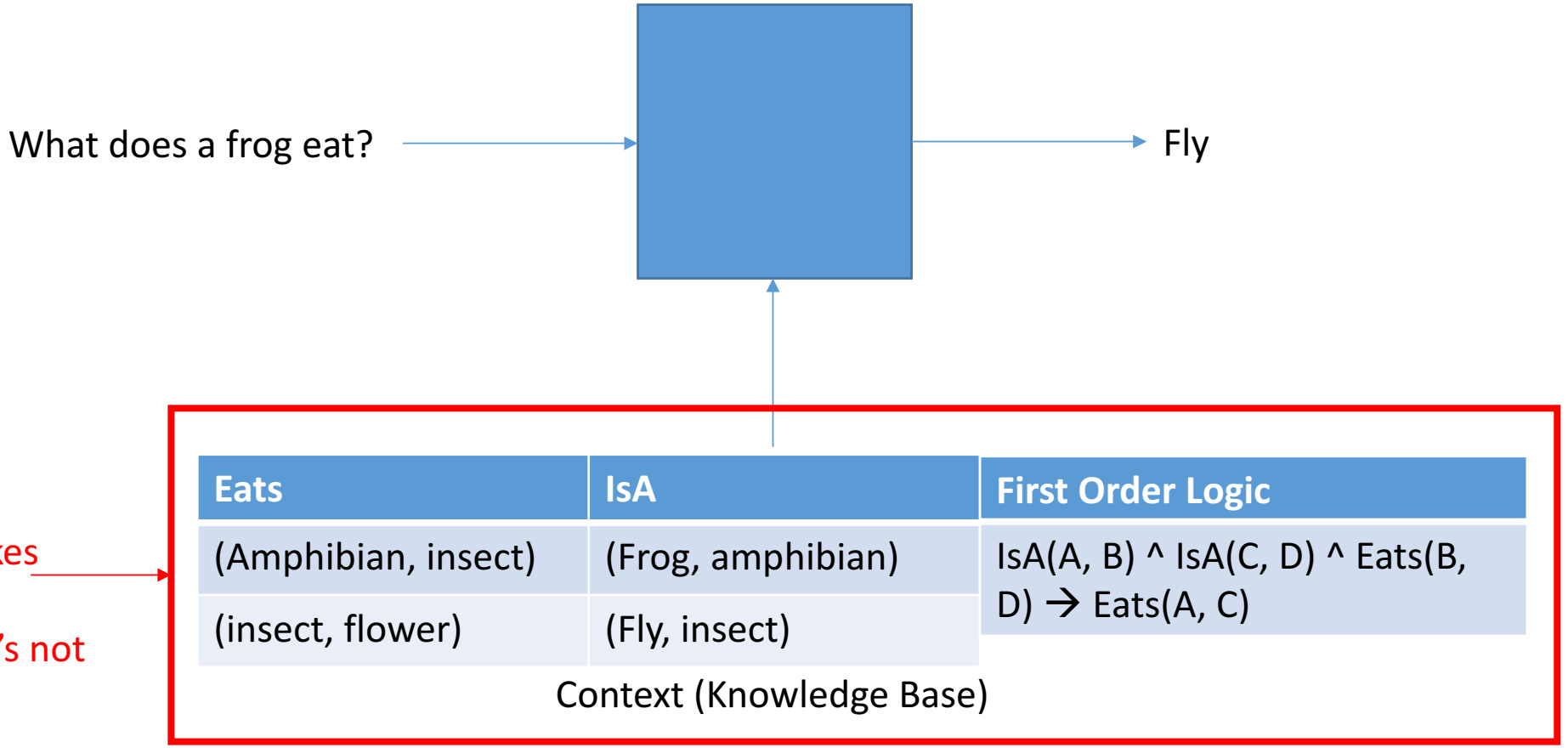
If our objective is to “answer” difficult questions …

- We can try to make the machine more capable of reasoning (better model)

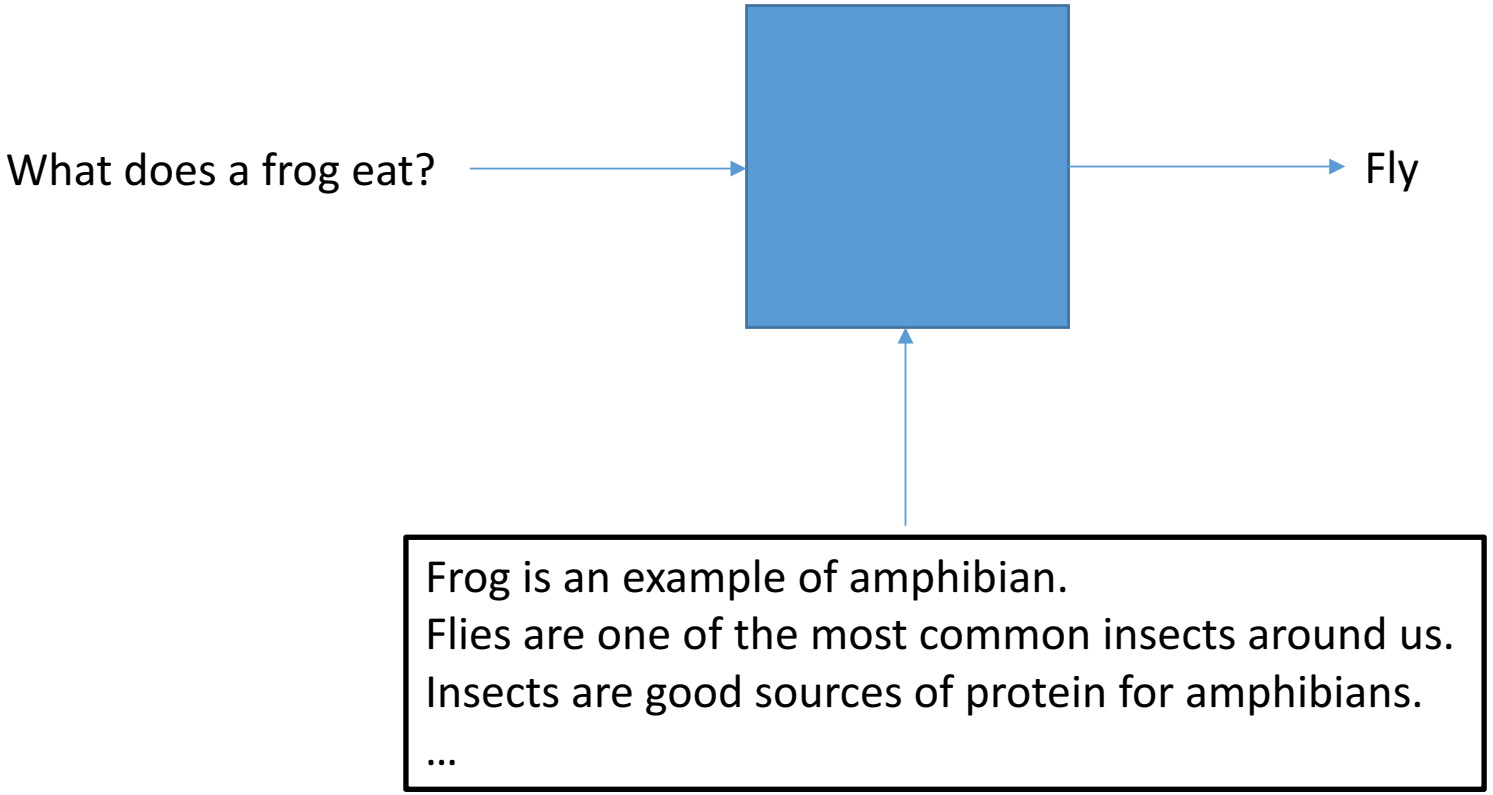
OR

- We can try to make more information explicit in the context (more data)

QA Model with Reasoning Capability

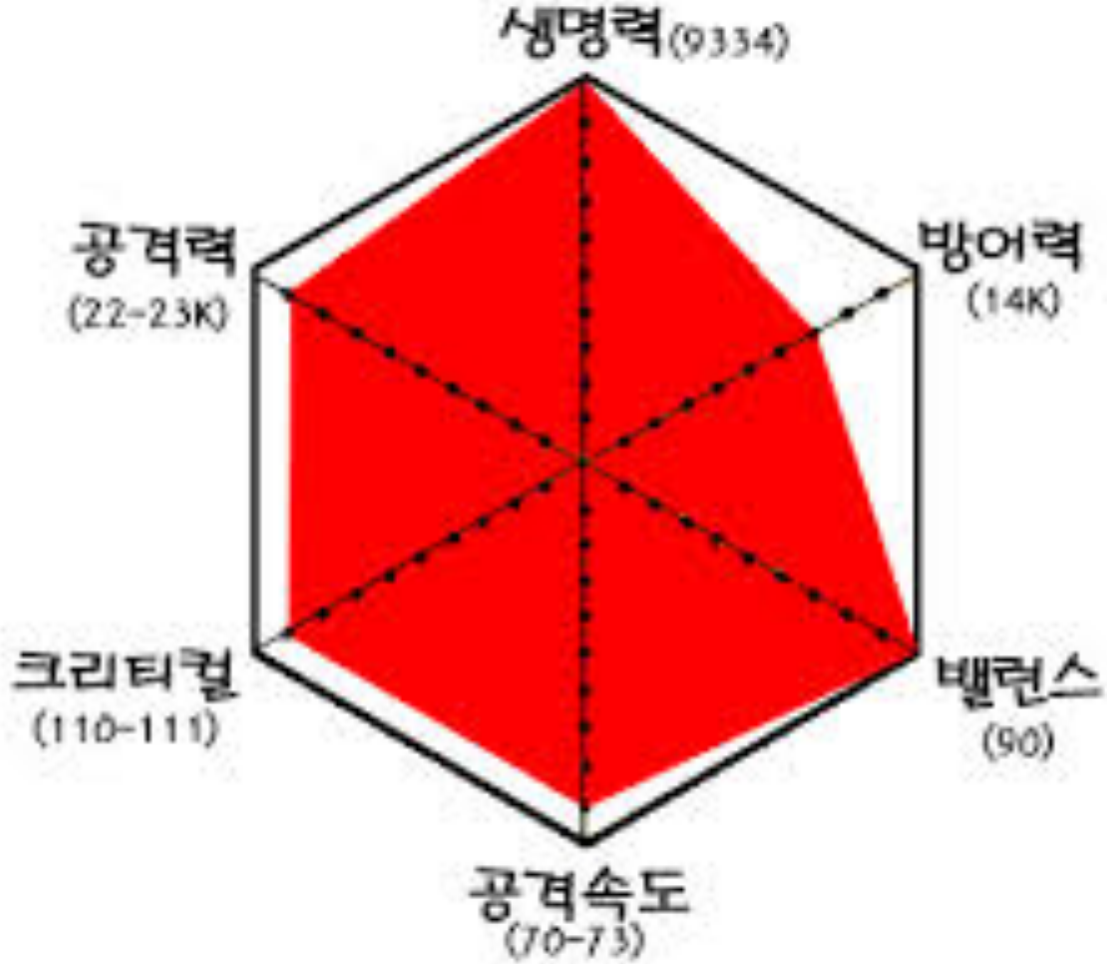


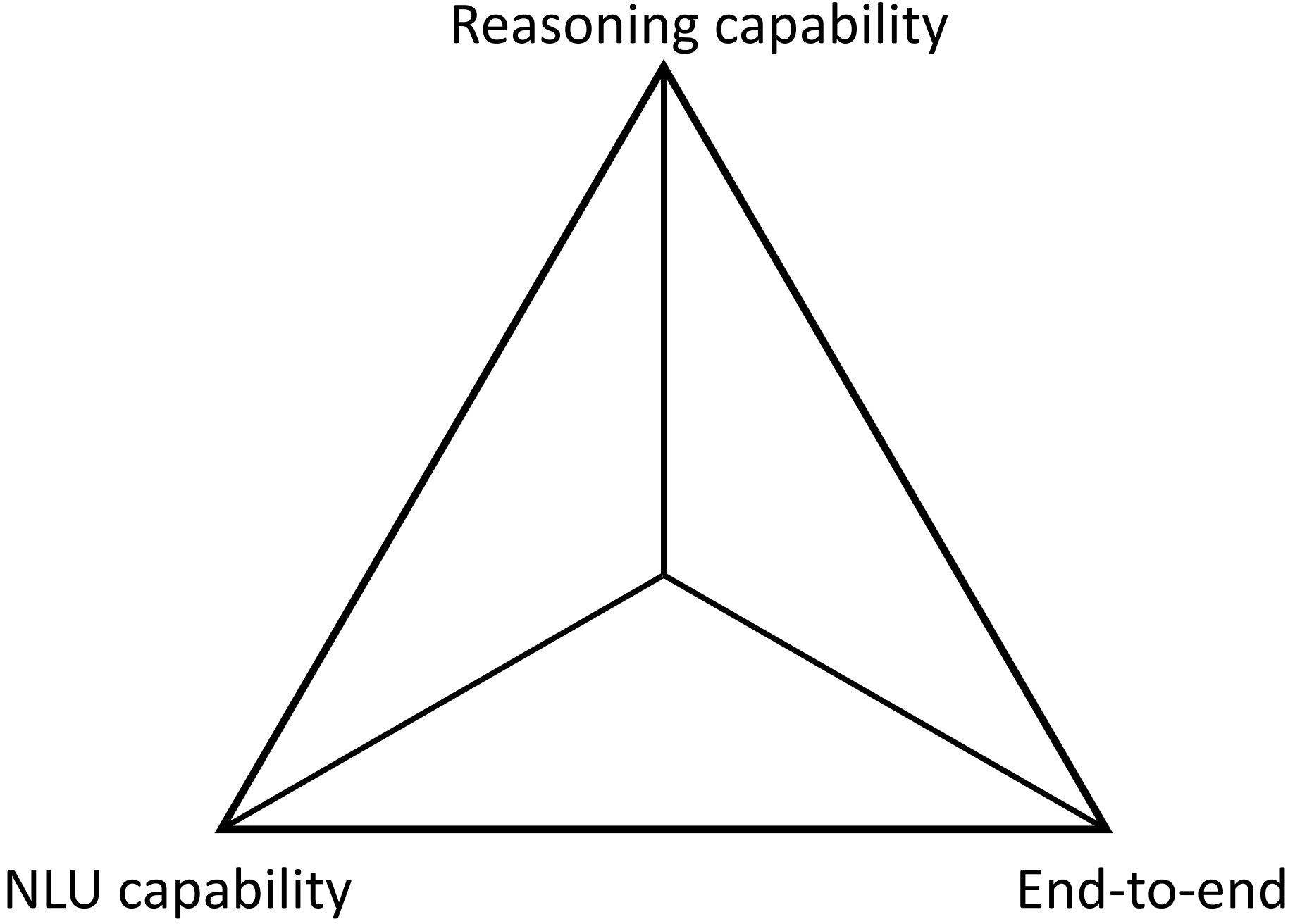
Reasoning QA Model with Unstructured Data



I am interested in…

- **Natural language understanding**
  - Natural language has diverse surface forms (lexically, syntactically)
- **Learning to read text and reason by question answering (dialog)**
  - Text is unstructured data
  - Deriving new knowledge from existing knowledge
- **End-to-end training**
  - Minimizing human efforts


생명력(9334)
방어력
(14K)
밸런스
(90)
공격속도
(70-73)
크리티컬
(110-111)
공격력
(22-23K)


Reasoning capability
NLU capability
End-to-end

Geometry QA

AAAI 2014
EMNLP 2015

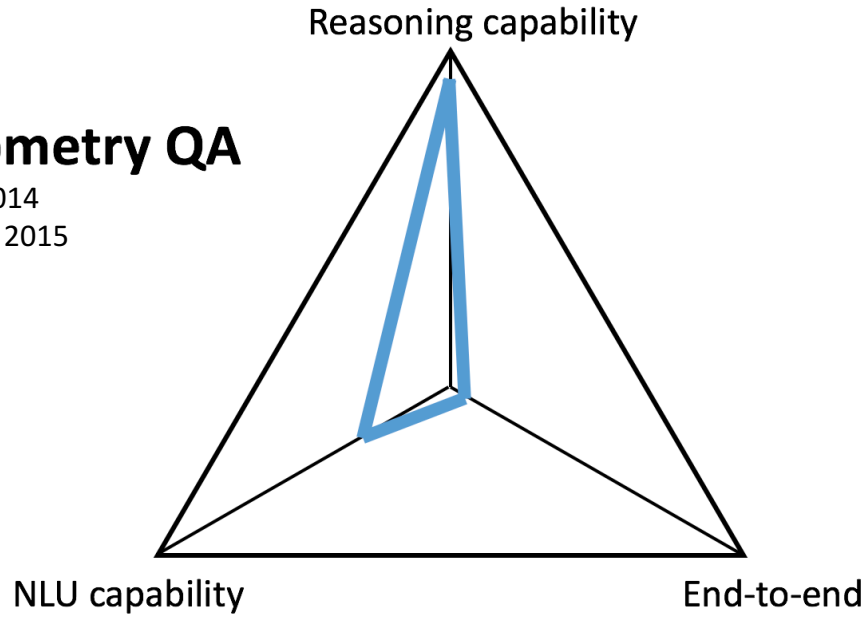
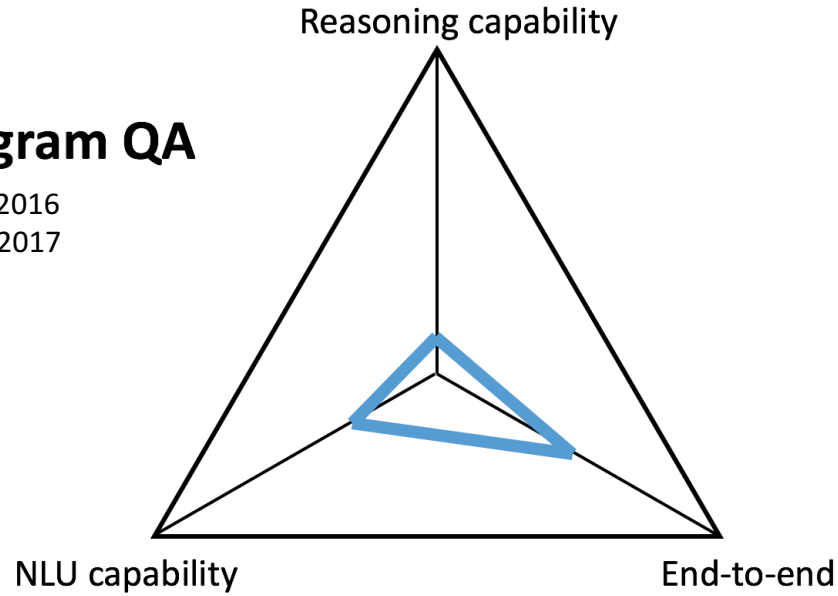


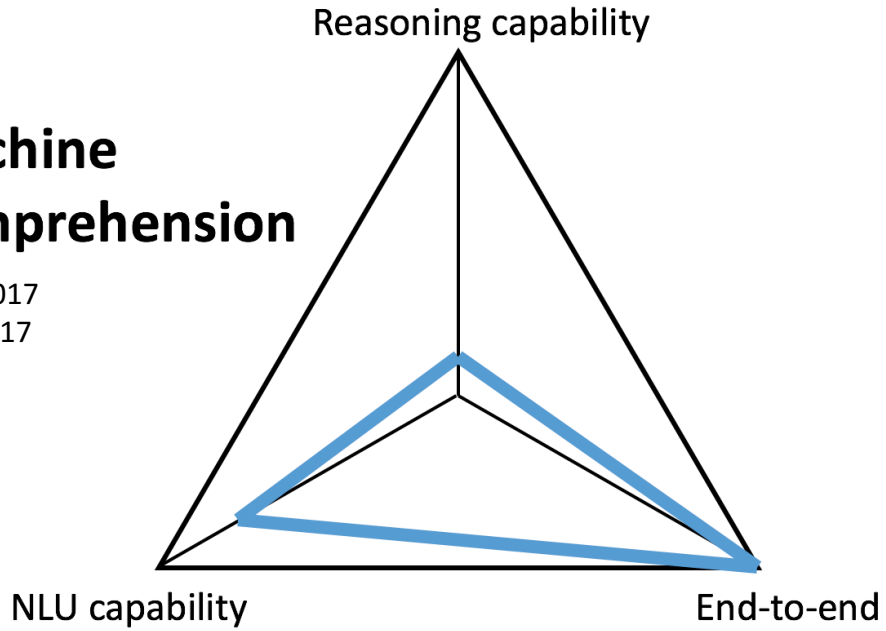
Diagram QA

ECCV 2016
CVPR 2017



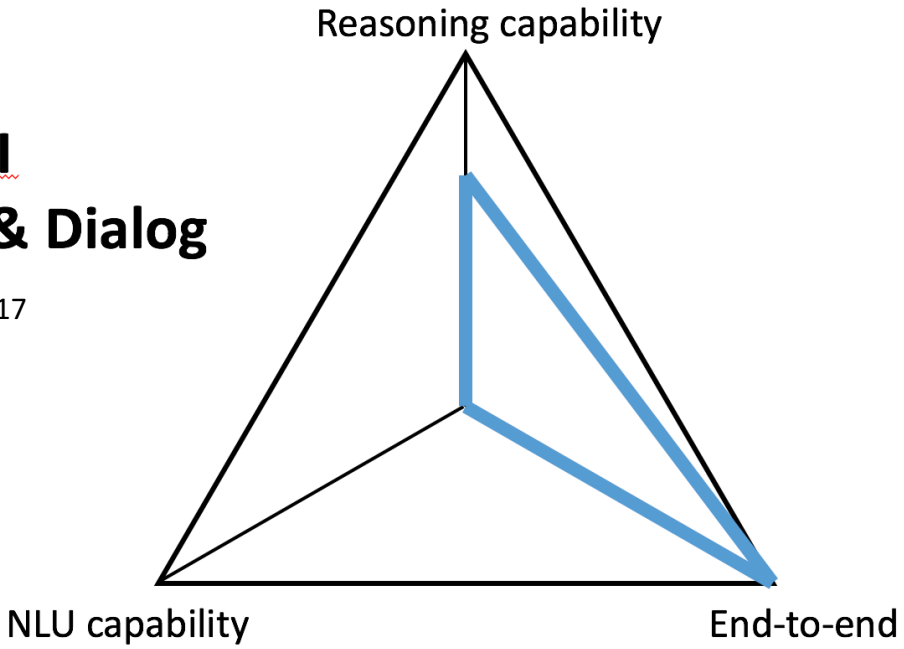
Machine Comprehension

ICLR 2017
ACL 2017

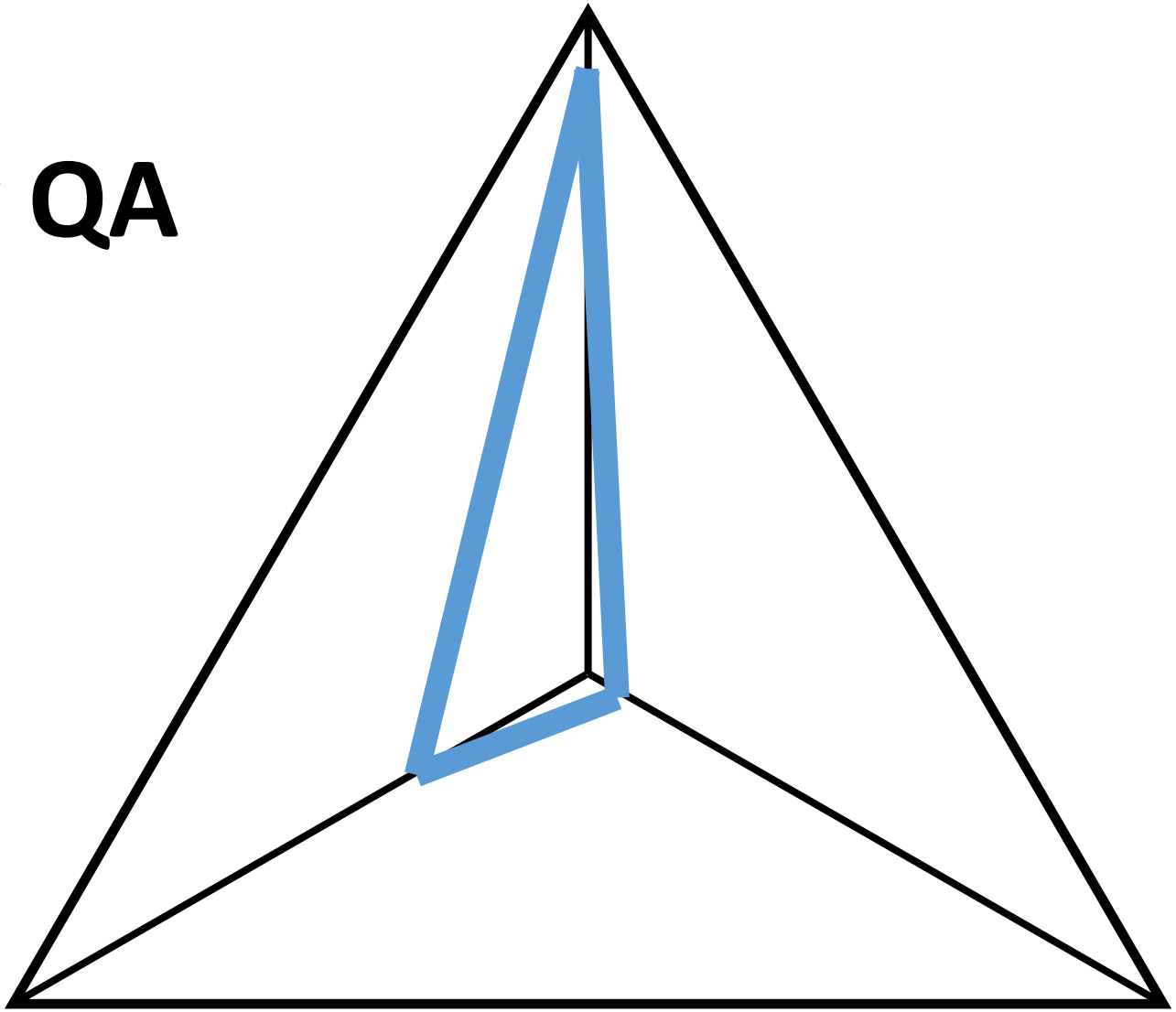


bAbI QA & Dialog

ICLR 2017



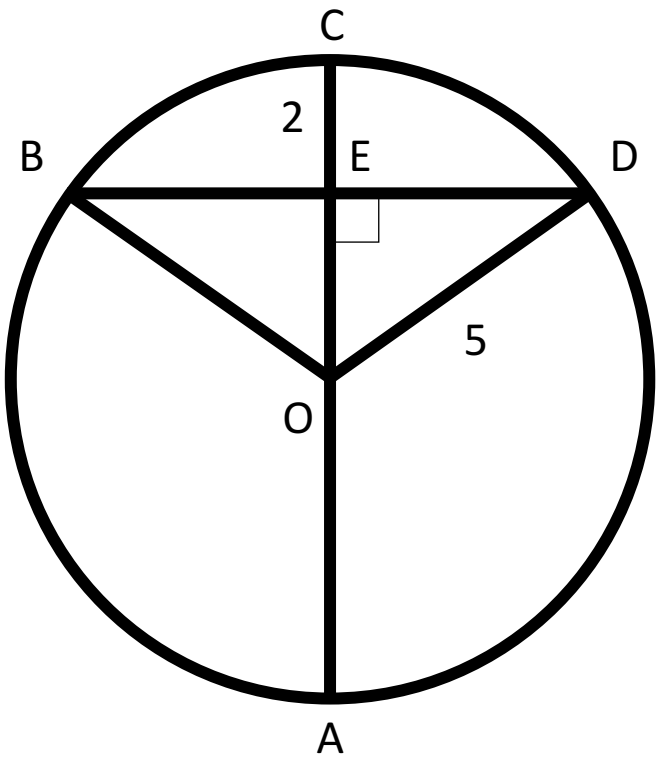
Geometry QA



Geometry QA

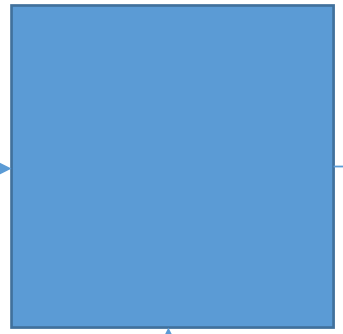
In the diagram at the right, circle O has a radius of 5, and CE = 2. Diameter AC is perpendicular to chord BD. What is the length of BD?

a) 2 b) 4 c) 6
d) 8 e) 10



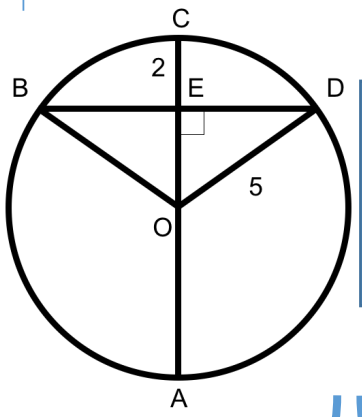
Geometry QA Model

What is the length of BD?



8

In the diagram at the right, circle O has a radius of 5, and CE = 2. Diameter AC is perpendicular to chord BD.



First Order Logic

Local context

Global context

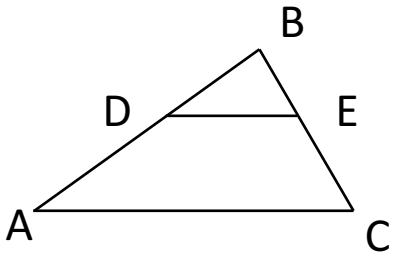
Method

- Learn to map question to logical form
- Learn to map local context to logical form
  - Text → logical form
  - Diagram → logical form
- Global context is already formal!
  - **Manually** defined
  - “If AB = BC, then <CAB = <ACB”
- Solver on all logical forms
  - We created a *reasonable* numerical solver

Mapping question / text to logical form

Text Input

In triangle ABC, line DE is parallel with line AC, DB equals 4, AD is 8, and DE is 5. Find AC.
(a) 9 (b) 10 (c) 12.5 (d) 15 (e) 17



Logical form

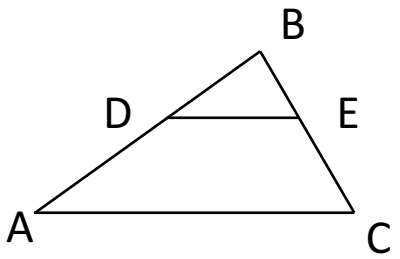
IsTriangle(ABC) ∧ Parallel(AC, DE) ∧ Equals(LengthOf(DB), 4) ∧ Equals(LengthOf(AD), 8) ∧ Equals(LengthOf(DE), 5) ∧ Find(LengthOf(AC))

Difficult to directly map text to a long logical form!

Mapping question / text to logical form

Text Input

In triangle ABC, line DE is parallel with line AC, DB equals 4, AD is 8, and DE is 5. Find AC.
(a) 9 (b) 10 (c) 12.5 (d) 15 (e) 17



Our method

Over-generated literals	Text scores	Diagram scores
IsTriangle(ABC)	*0.96*	*1.00*
Parallel(AC, DE)	*0.91*	*0.99*
Parallel(AC, DB)	*0.74*	*0.02*
Equals(LengthOf(DB), 4)	*0.97*	*n/a*
Equals(LengthOf(AD), 8)	*0.94*	*n/a*
Equals(LengthOf(DE), 5)	*0.94*	*n/a*
Equals(4, LengthOf(AD))	*0.31*	*n/a*
…	…	…

Selected subset

Logical form

IsTriangle(ABC) ∧ Parallel(AC, DE) ∧ Equals(LengthOf(DB), 4) ∧ Equals(LengthOf(AD), 8) ∧ Equals(LengthOf(DE), 5) ∧ Find(LengthOf(AC))

Numerical solver

- Translate literals to numeric equations

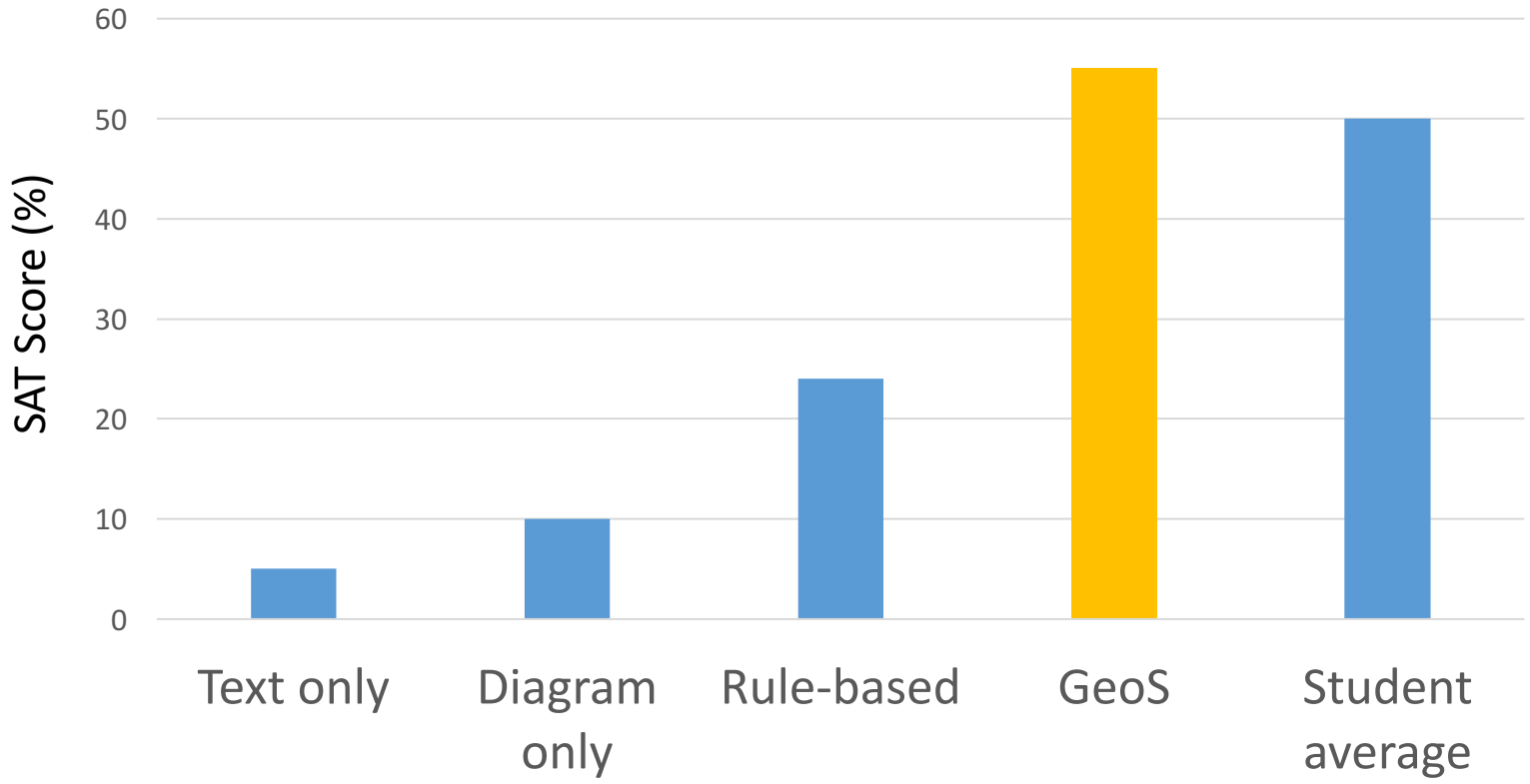
Literal	Equation
Equals(LengthOf(AB),d)	$(A_x-B_x)^2+(A_y-B_y)^2-d^2 = 0$
Parallel(AB, CD)	$(A_x-B_x)(C_y-D_y)-(A_y-B_y)(C_x-D_x) = 0$
PointLiesOnLine(B, AC)	$(A_x-B_x)(B_y-C_y)-(A_y-B_y)(B_x-C_x) = 0$
Perpendicular(AB,CD)	$(A_x-B_x)(C_x-D_x)+(A_y-B_y)(C_y-D_y) = 0$

- Find the solution to the equation system
- Use off-the-shelf numerical minimizers (Wales and Doye, 1997; Kraft, 1988)
- Numerical solver can choose not to answer question

Dataset

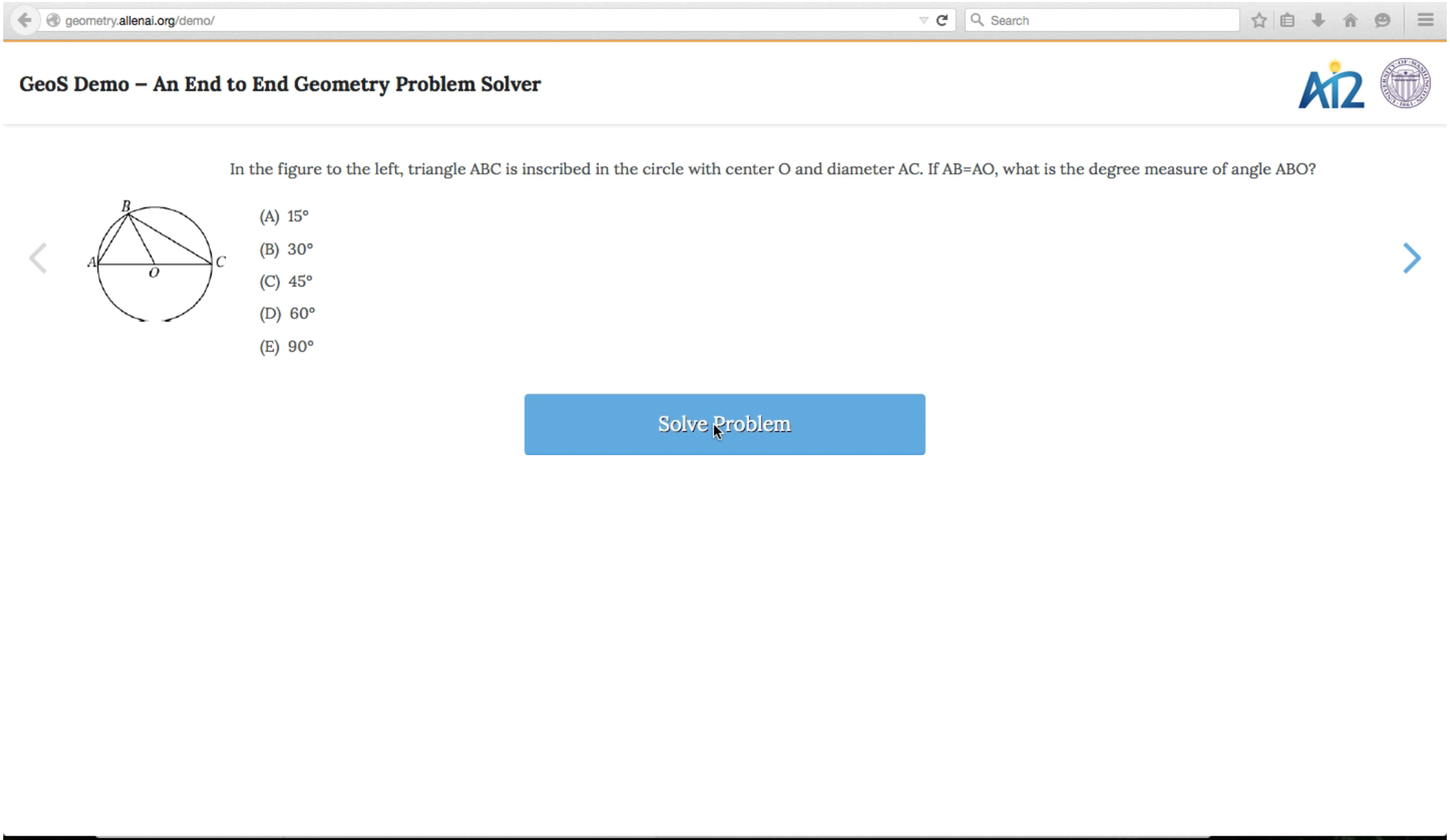
- **Training questions** (67 questions, 121 sentences)
  - Seo et al., 2014
  - High school geometry questions
- **Test questions** (119 questions, 215 sentences)
  - We collected them
  - SAT (US college entrance exam) geometry questions
- We manually annotated the text parse of all questions

Results (EMNLP 2015)



*** 0.25 penalty for incorrect answer

Demo (geometry.allenai.org/demo)



Limitations

- Dataset is small
- Required level of reasoning is very high
  - A lot of manual efforts (annotations, rule definitions, etc.)
  - End-to-end system is simply hopeless

- Collect more data?
- Change task?
- Curriculum learning? (Do more *hopeful* tasks first?)

Diagram QA

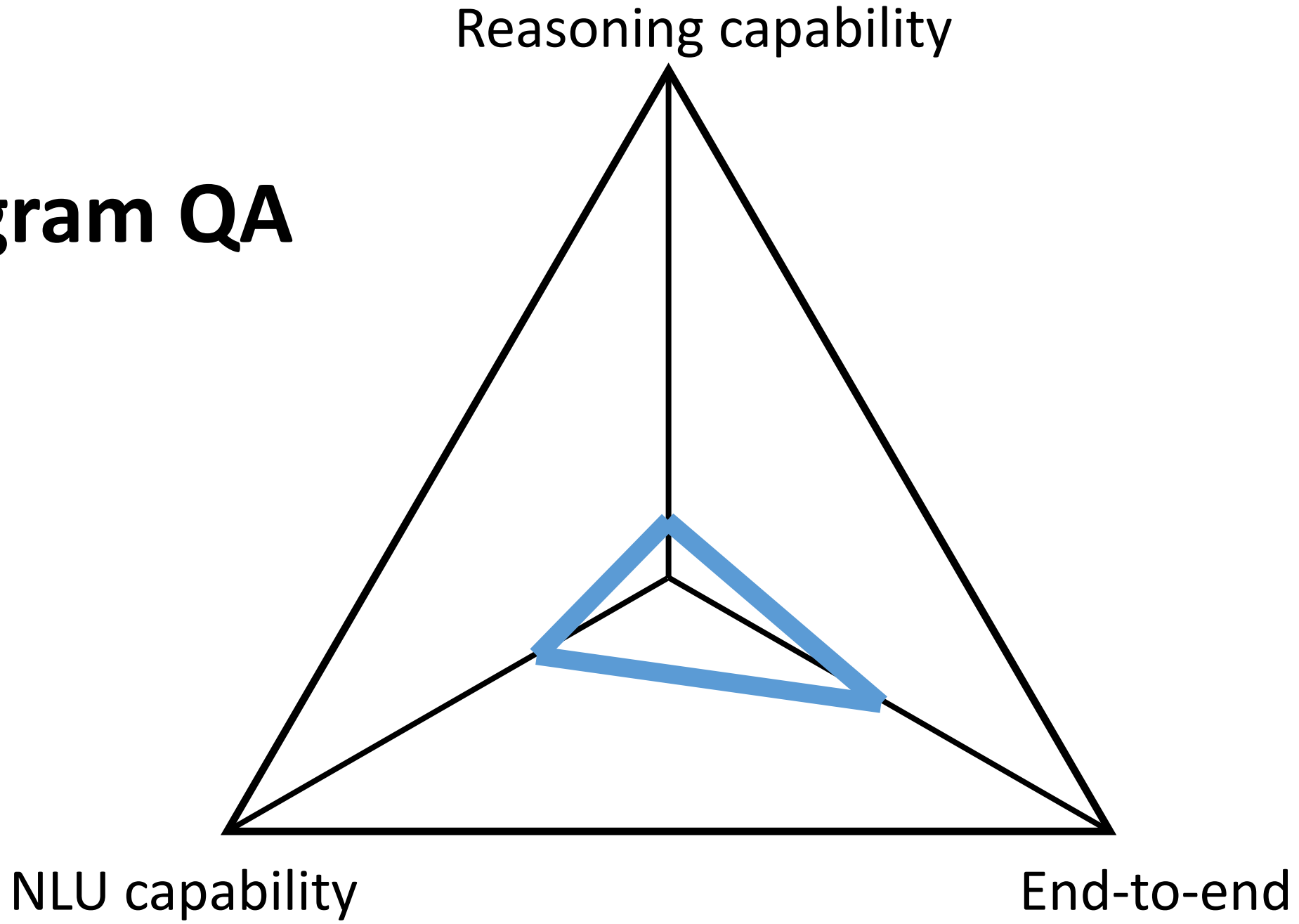
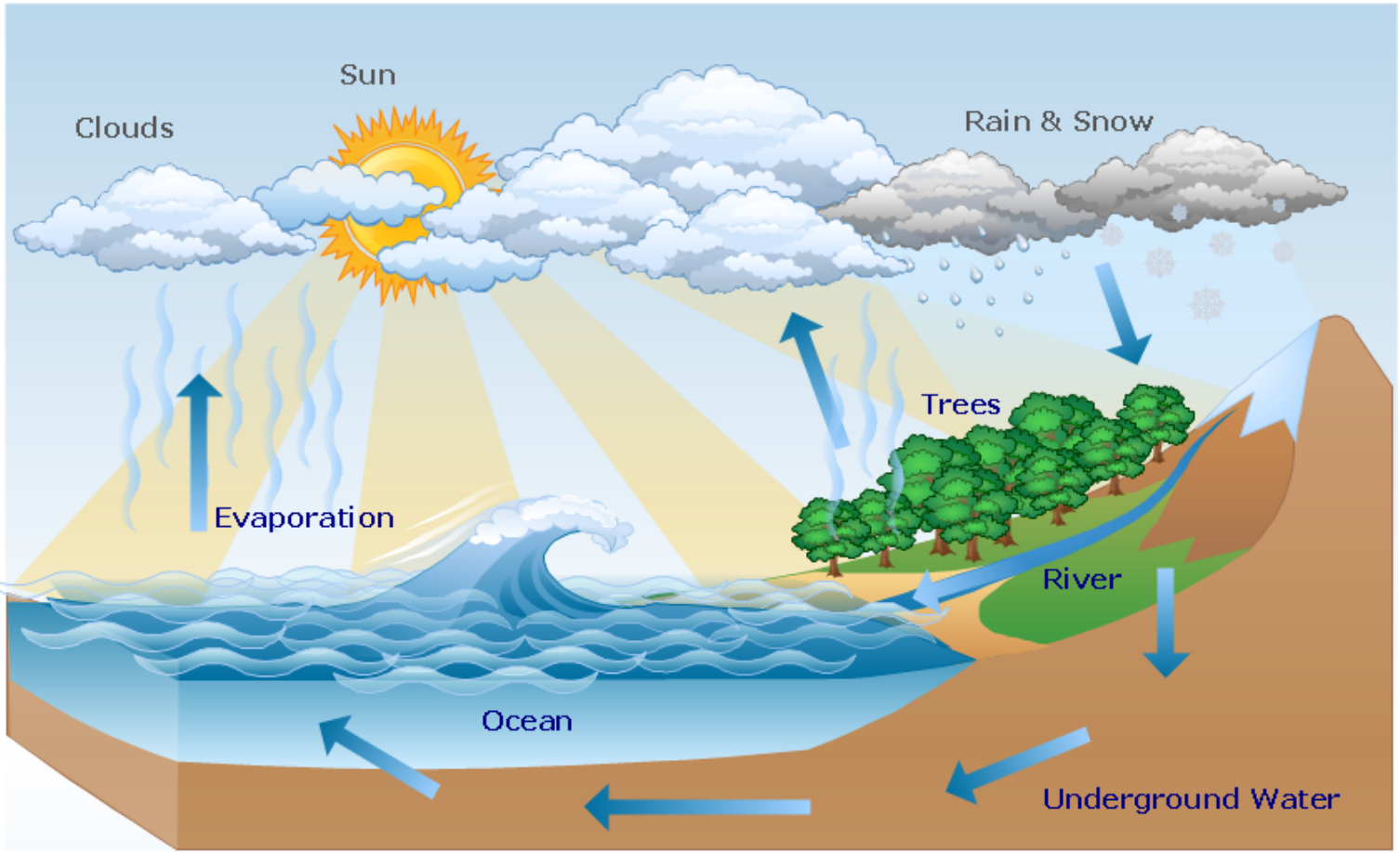


Diagram QA



Q: The process of water being heated by sun and becoming gas is called

A: Evaporation

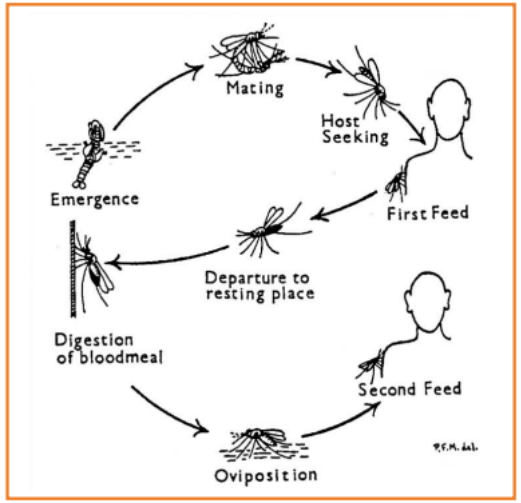
Is DQA subset of VQA?

- Diagrams and real images are very different
- Diagram components are simpler than real images
- Diagram contains a lot of information in a single image
- Diagrams are few (whereas real images are almost infinitely many)

Problem

What comes before second feed?

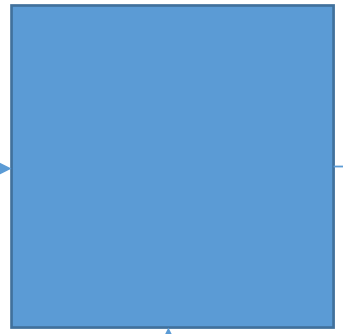
8



Difficult to latently learn relationships

Strategy

What does a frog eat?



Fly

Diagram Graph

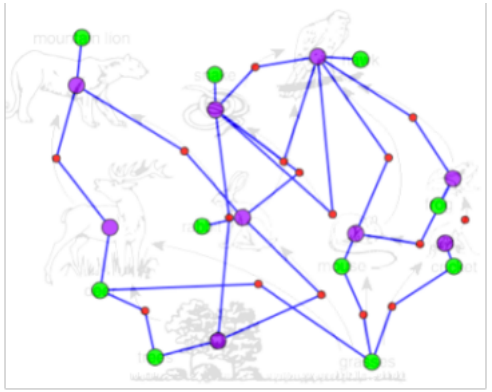
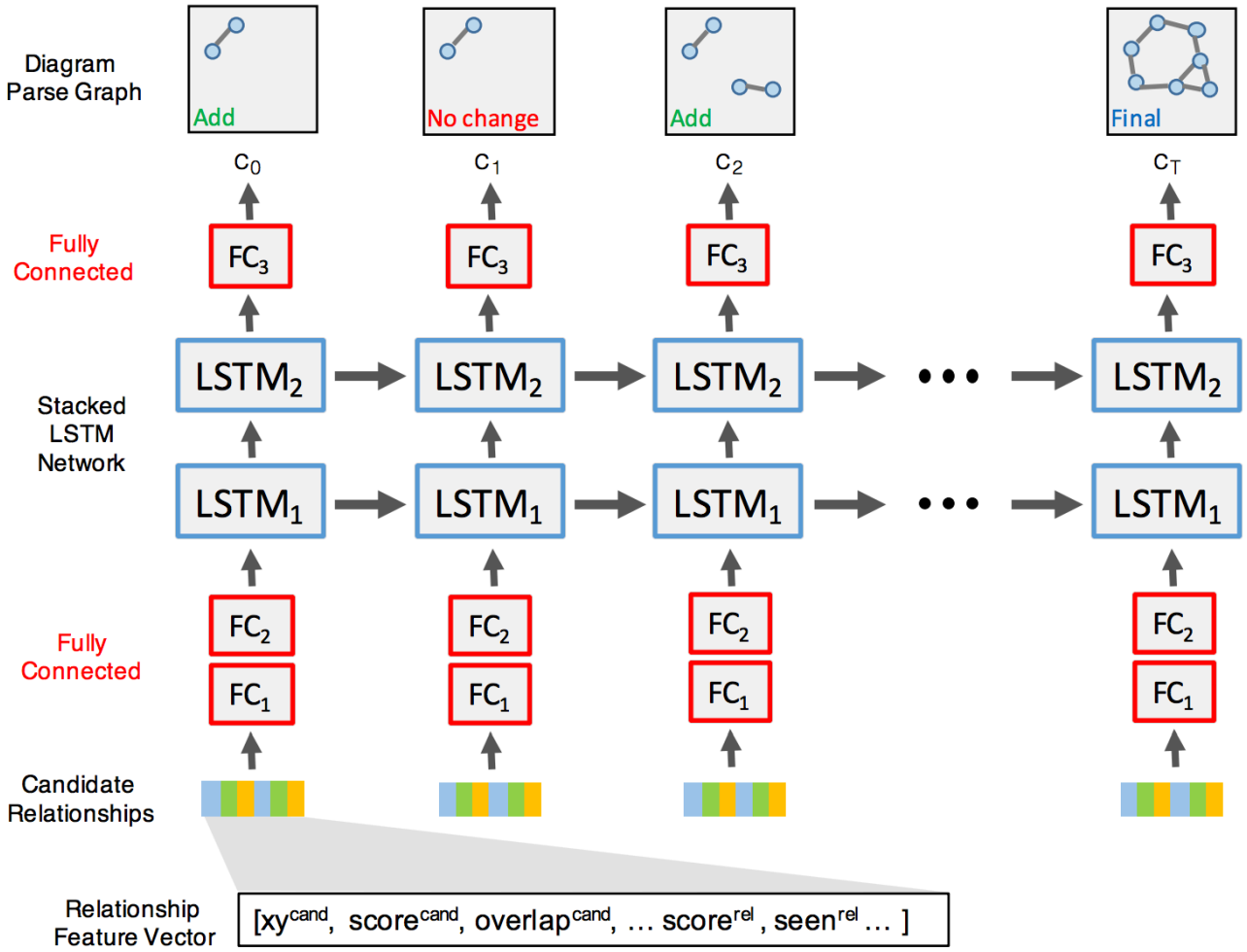
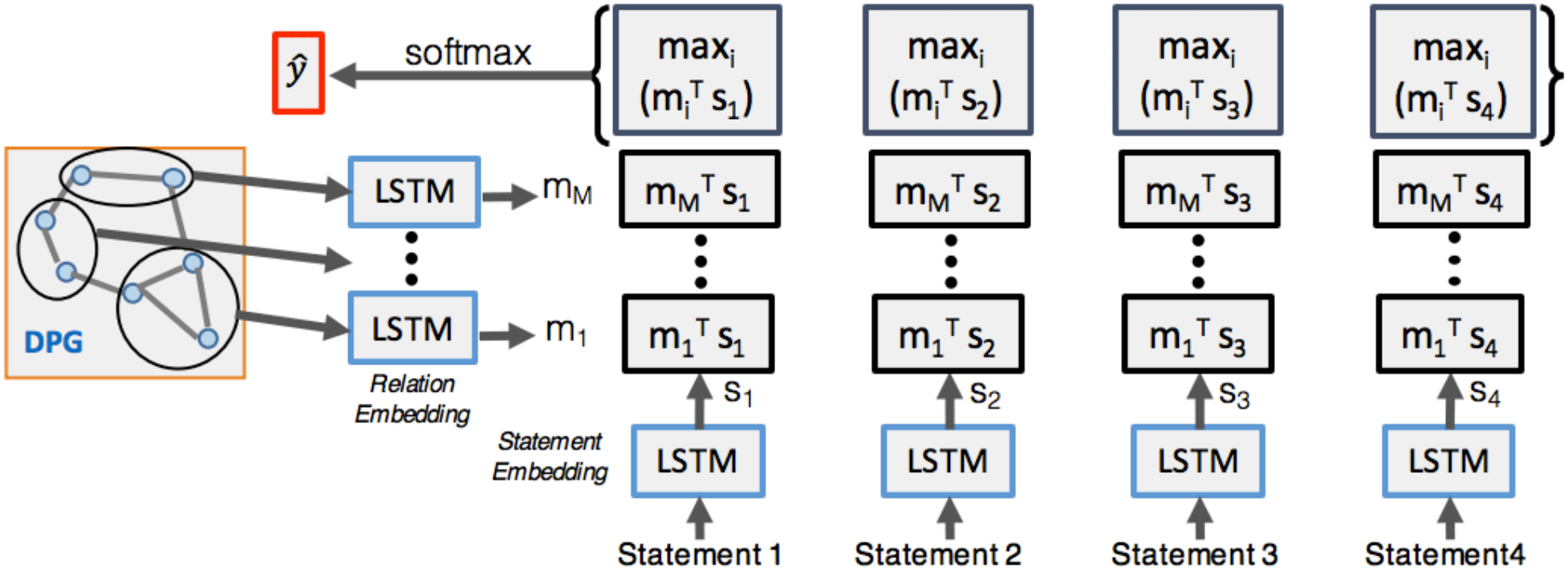


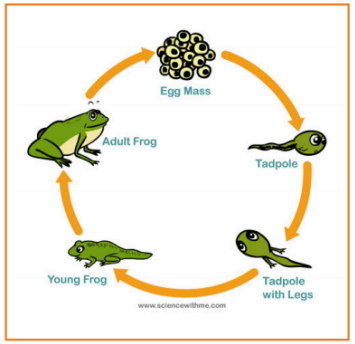
Diagram Parsing



Question Answering

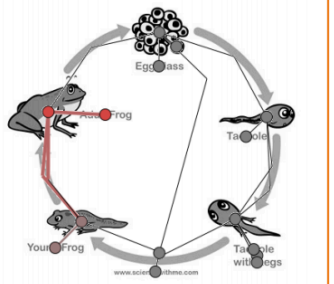


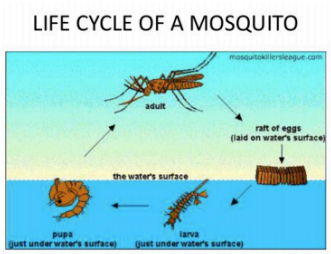
Attention visualization



The diagram depicts
The life cycle of

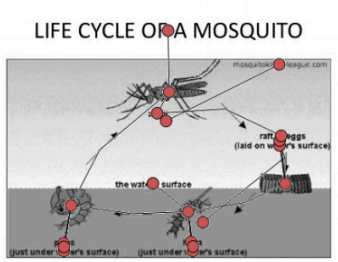
a) frog	**0.924**
b) bird	0.02
c) insecticide	0.054
d) insect	0.002

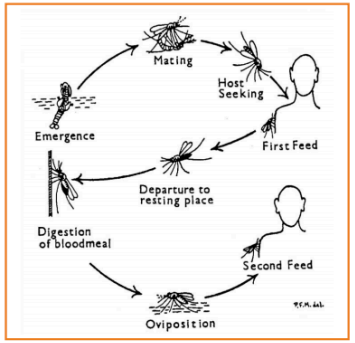




How many stages of
Growth does the diagram
Feature?

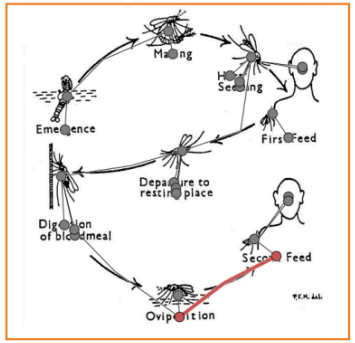
a) 4	**0.924**
b) 2	0.02
c) 3	0.054
d) 1	0.002





What comes before
Second feed?

a) digestion	0.0
b) First feed	0.15
c) indigestion	0.0
d) oviposition	0.85



Results (ECCV 2016)

Method	Training data	Accuracy
Random (expected)	-	25.00
LSTM + CNN	VQA	29.06
LSTM + CNN	AI2D	32.90
Ours	AI2D	**38.47**

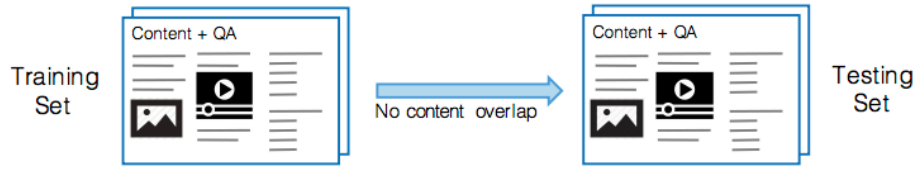
Limitations

- You can’t really call this *reasoning*…
  - Rather matchting algorithm
  - No complex inference involved

- You need a lot of prior knowledge to answer some questions!
  - E.g. “Fly is an insect”, “Frog is an amphibian”

Textbook QA textbookqa.org (CVPR 2017)

Multi-modal Machine Comprehension (M^3C)



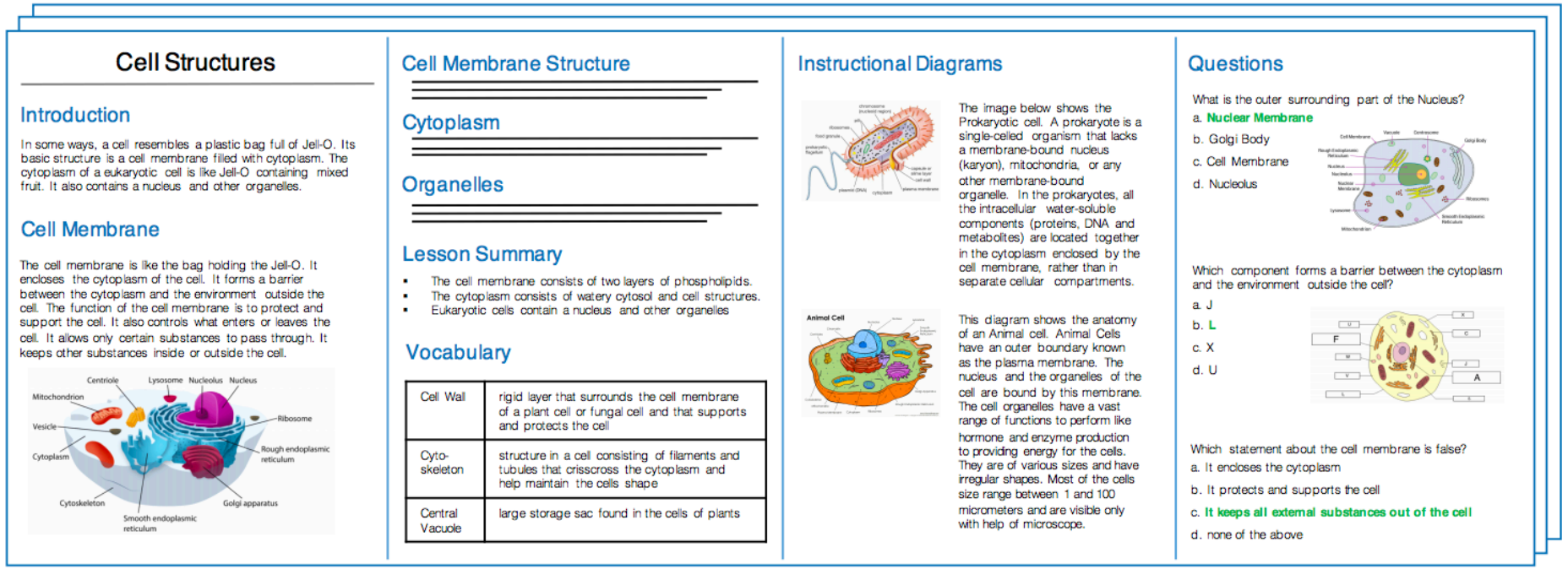
Textbook Question Answering (TQA)

1076 lessons from middle school curricula

Life Science | Earth Science | Physical Science

78,338 sentences
3,455 images
26,260 questions

Lessons in TQA



Cell Structures

Introduction

In some ways, a cell resembles a plastic bag full of Jell-O. Its basic structure is a cell membrane filled with cytoplasm. The cytoplasm of a eukaryotic cell is like Jell-O containing mixed fruit. It also contains a nucleus and other organelles.

Cell Membrane

The cell membrane is like the bag holding the Jell-O. It encloses the cytoplasm of the cell. It forms a barrier between the cytoplasm and the environment outside the cell. The function of the cell membrane is to protect and support the cell. It also controls what enters or leaves the cell. It allows only certain substances to pass through. It keeps other substances inside or outside the cell.

Cell Membrane Structure

Cytoplasm

Organelles

Lesson Summary

- The cell membrane consists of two layers of phospholipids.
- The cytoplasm consists of watery cytosol and cell structures.
- Eukaryotic cells contain a nucleus and other organelles

Vocabulary

Cell Wall	rigid layer that surrounds the cell membrane of a plant cell or fungal cell and that supports and protects the cell
Cyto-skeleton	structure in a cell consisting of filaments and tubules that crisscross the cytoplasm and help maintain the cells shape
Central Vacuole	large storage sac found in the cells of plants

Instructional Diagrams

The image below shows the Prokaryotic cell. A prokaryote is a single-celled organism that lacks a membrane-bound nucleus (karyon), mitochondria, or any other membrane-bound organelle. In the prokaryotes, all the intracellular water-soluble components (proteins, DNA and metabolites) are located together in the cytoplasm enclosed by the cell membrane, rather than in separate cellular compartments.

This diagram shows the anatomy of an Animal cell. Animal Cells have an outer boundary known as the plasma membrane. The nucleus and the organelles of the cell are bound by this membrane. The cell organelles have a vast range of functions to perform like hormone and enzyme production to providing energy for the cells. They are of various sizes and have irregular shapes. Most of the cells size range between 1 and 100 micrometers and are visible only with help of microscope.

Questions

What is the outer surrounding part of the Nucleus?

a. **Nuclear Membrane**
b. Golgi Body
c. Cell Membrane
d. Nucleolus

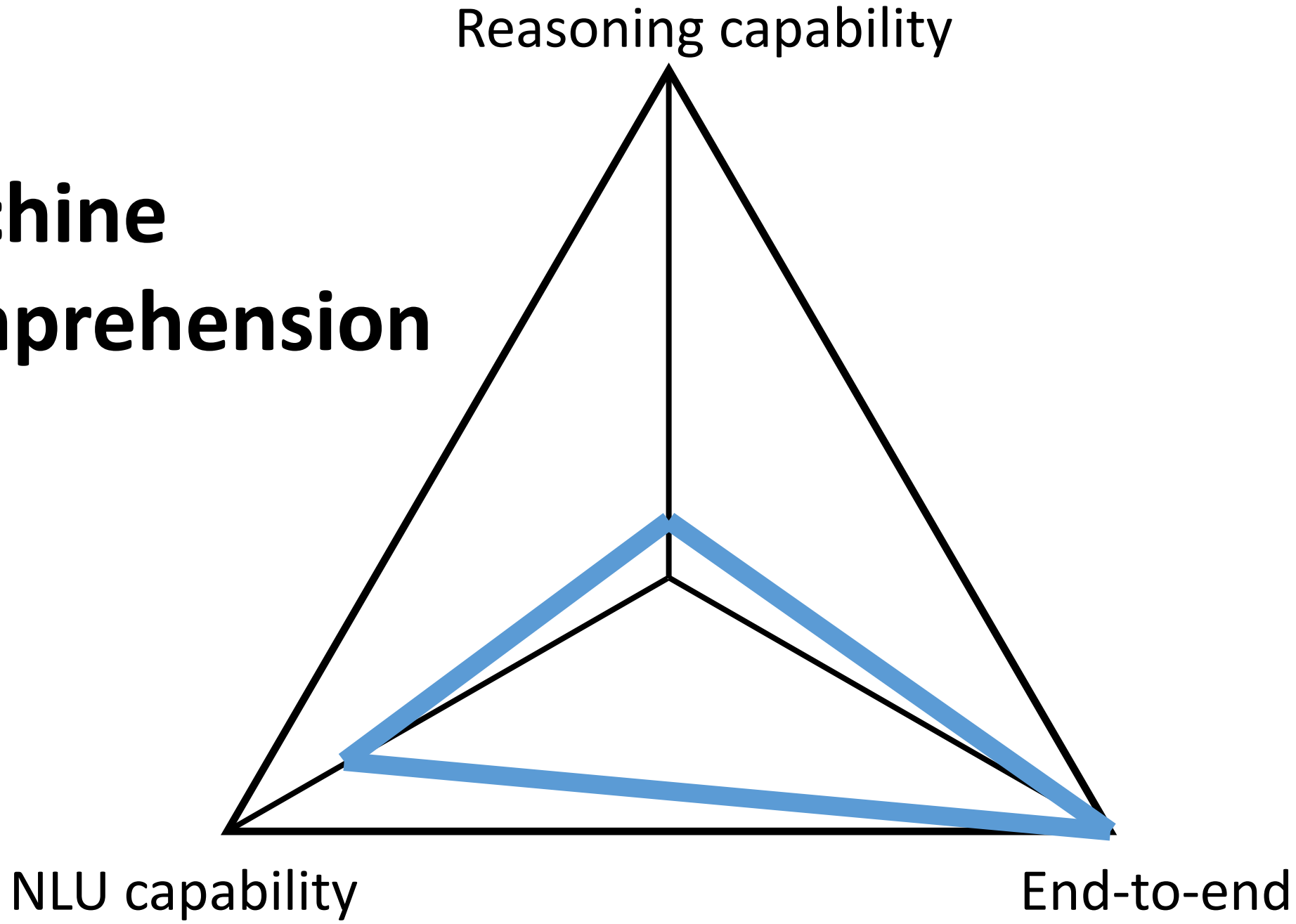
Which component forms a barrier between the cytoplasm and the environment outside the cell?

a. J
b. **L**
c. X
d. U

Which statement about the cell membrane is false?

a. It encloses the cytoplasm
b. It protects and supports the cell
c. **It keeps all external substances out of the cell**
d. none of the above

Machine Comprehension



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 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

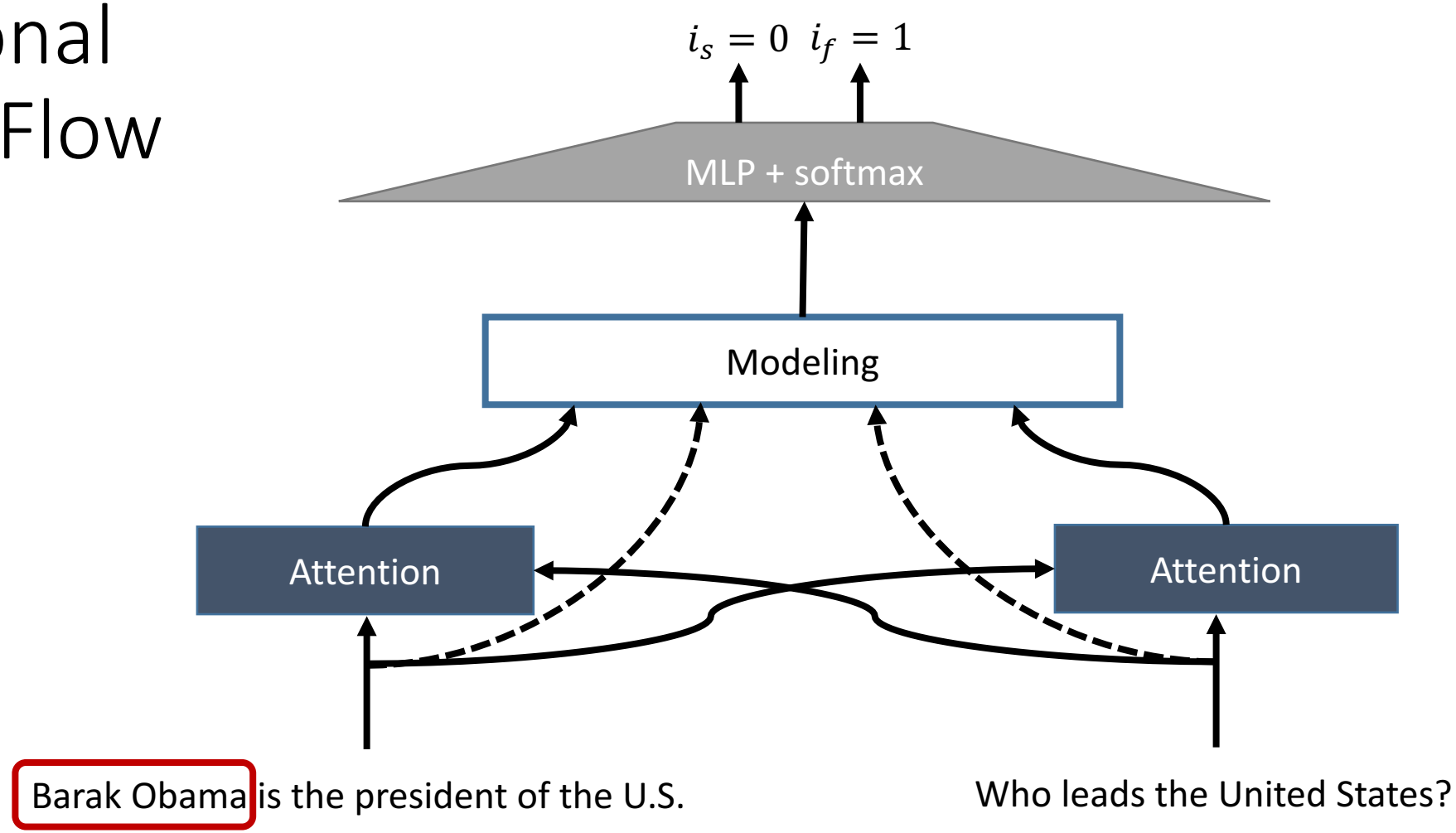
Why Neural Attention?

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.

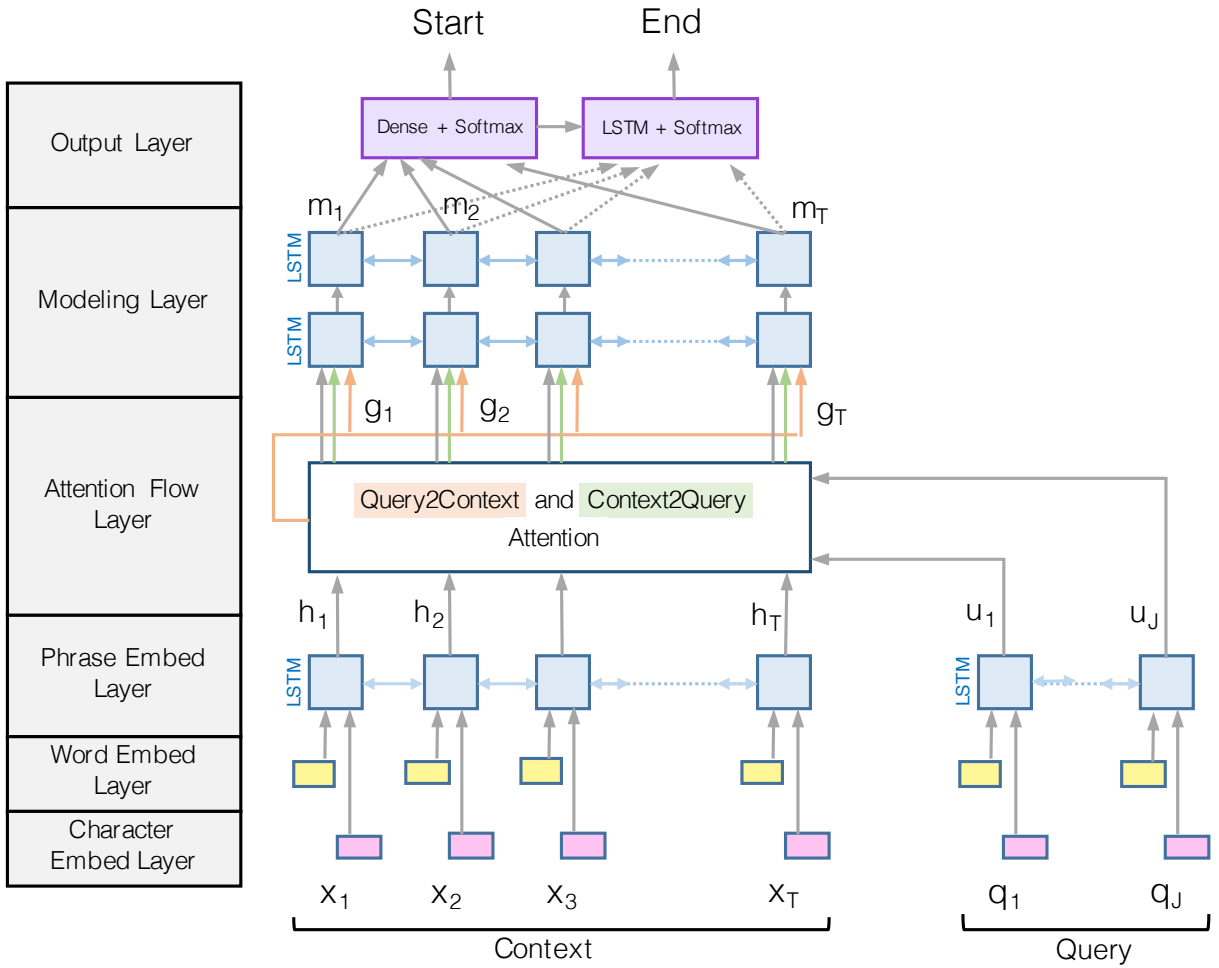
Q: Which NFL team represented the AFC at Super Bowl 50?

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

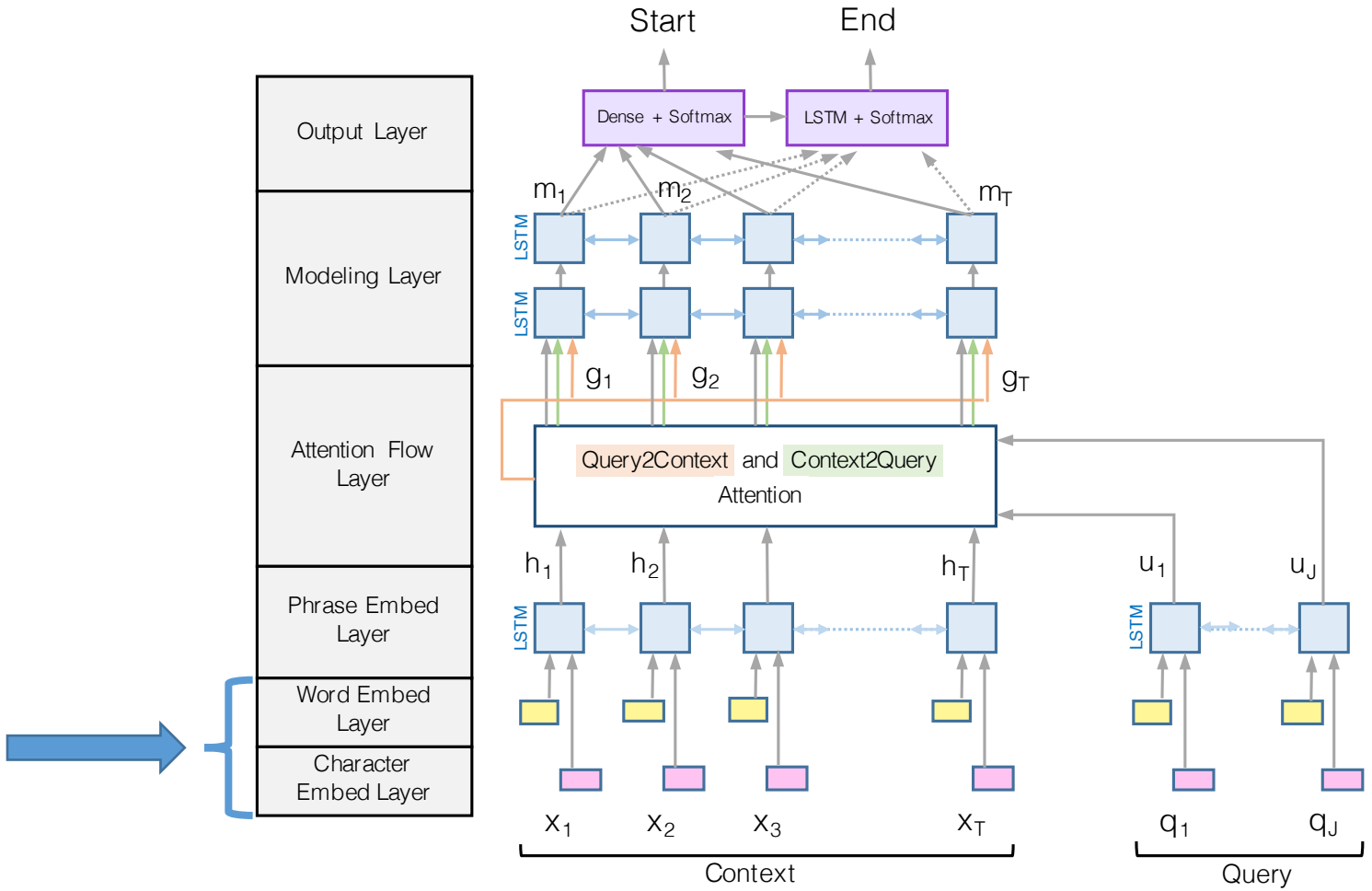
Our Model: Bi-directional Attention Flow (BiDAF)



(Bidirectional) Attention Flow



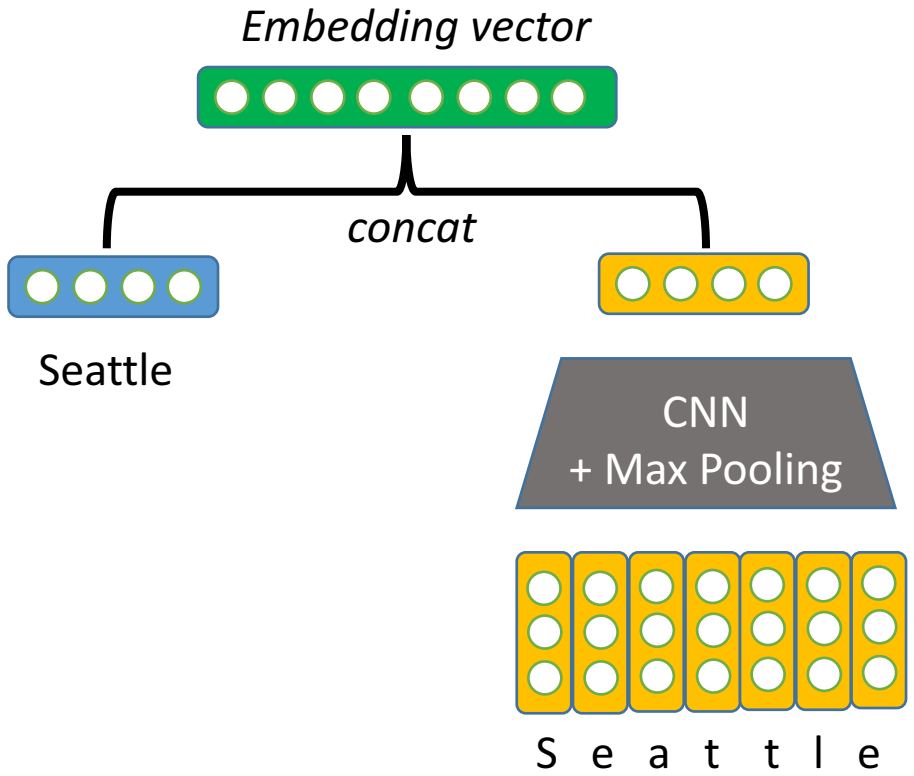
Char/Word Embedding Layers



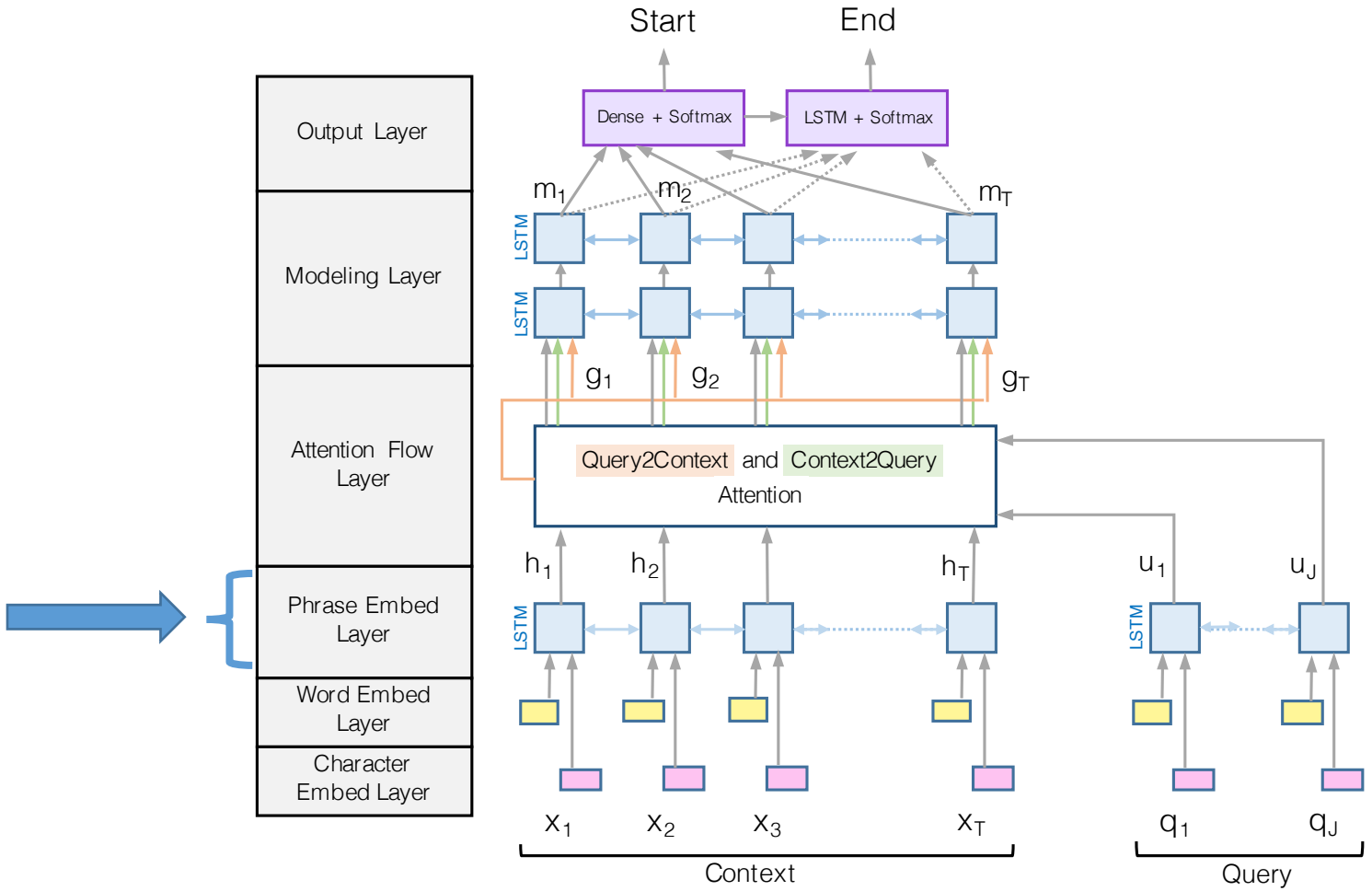
Character and Word Embedding

- 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)



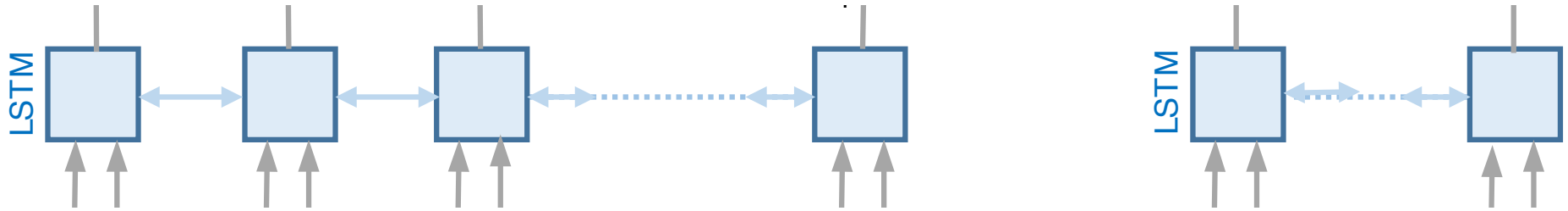
Phrase Embedding Layer



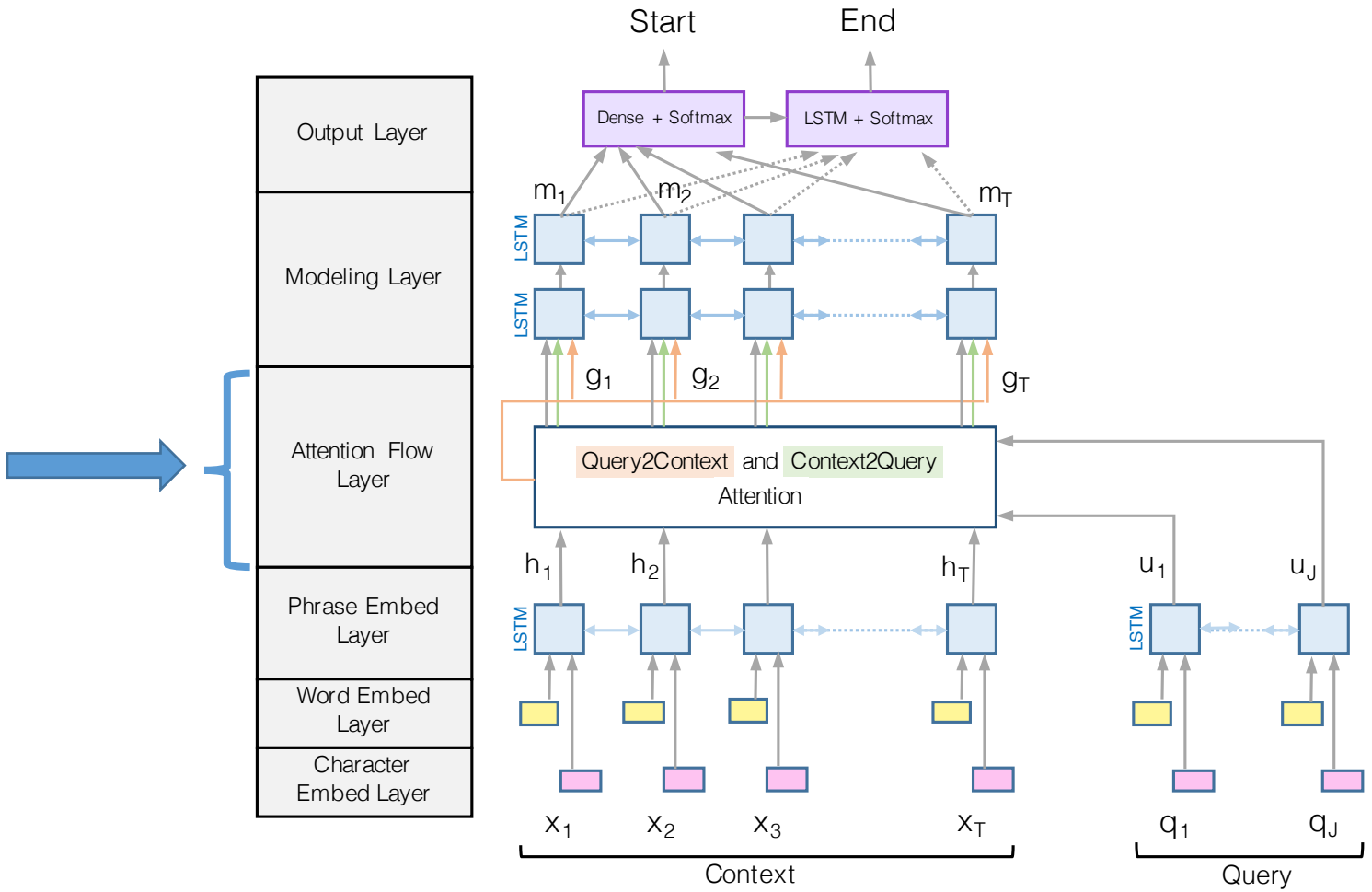
Phrase Embedding Layer

- **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



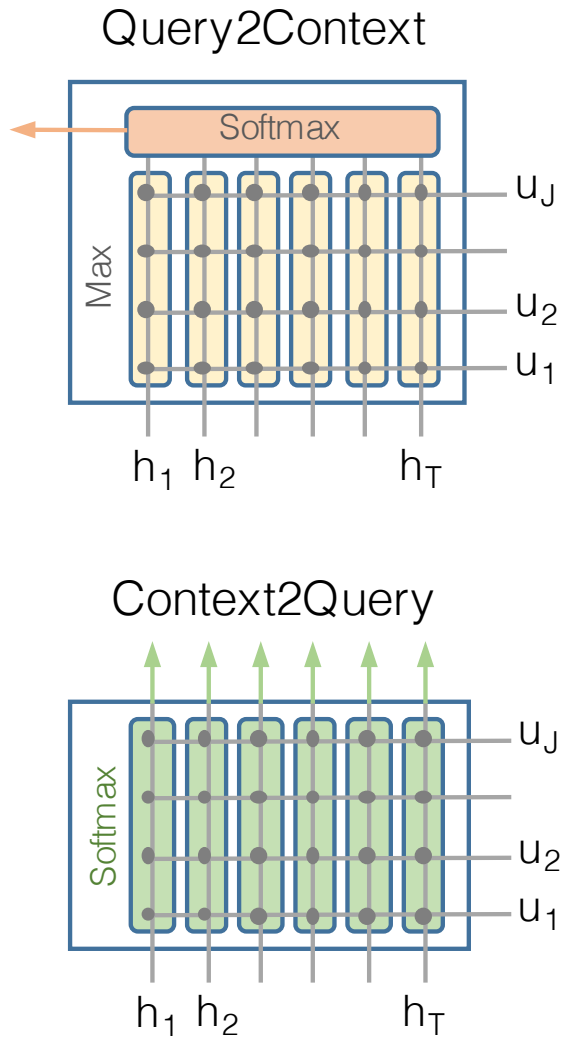
Attention Layer



Attention Layer

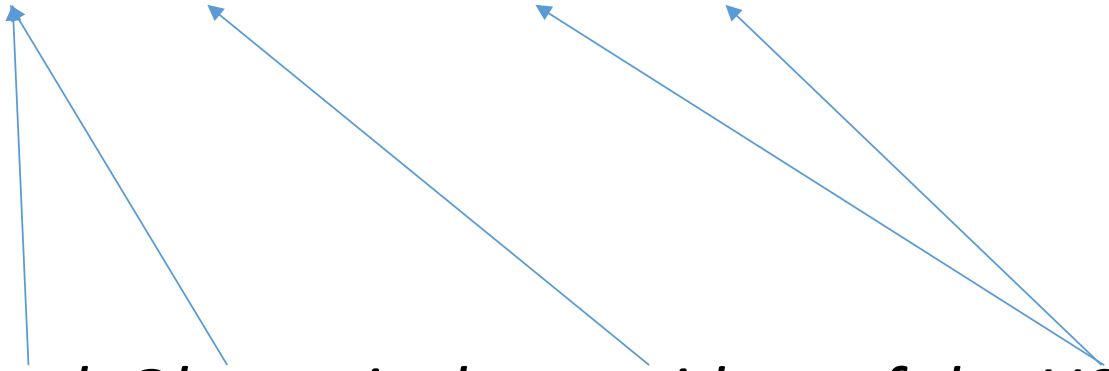
- **Inputs**: phrase-aware context and query words
- **Outputs**: query-aware representations of context words

- **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?*

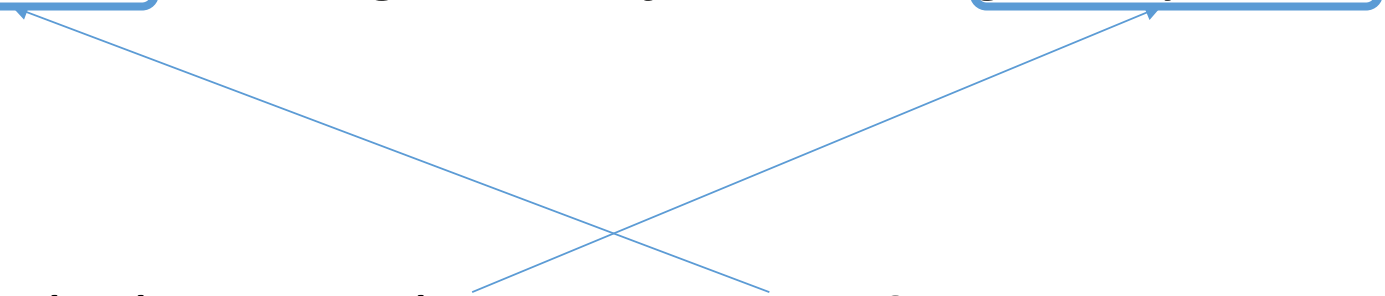


C: *Barak Obama is the president of the USA.*

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

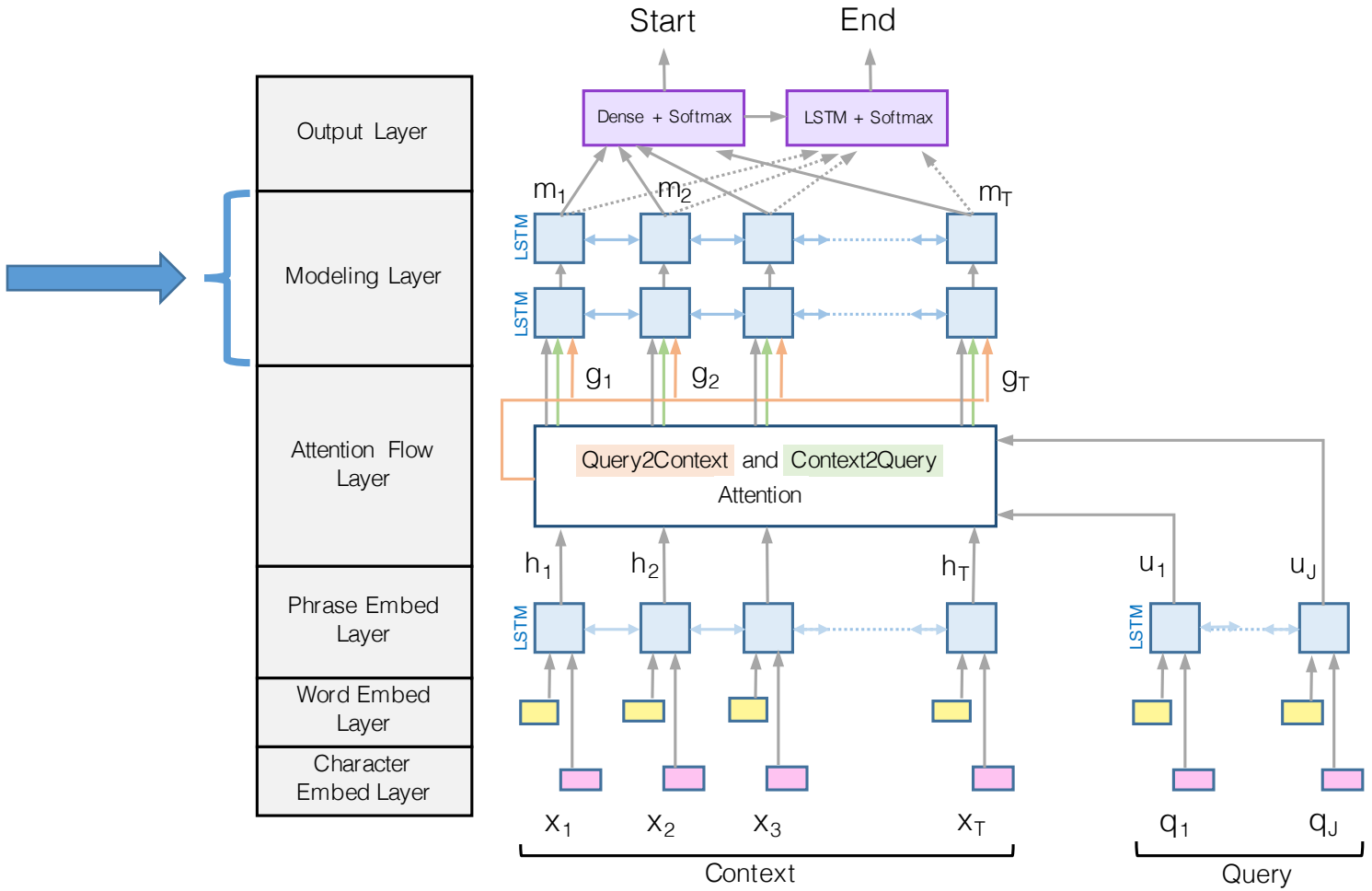
Query-to-Context Attention (Q2C)

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

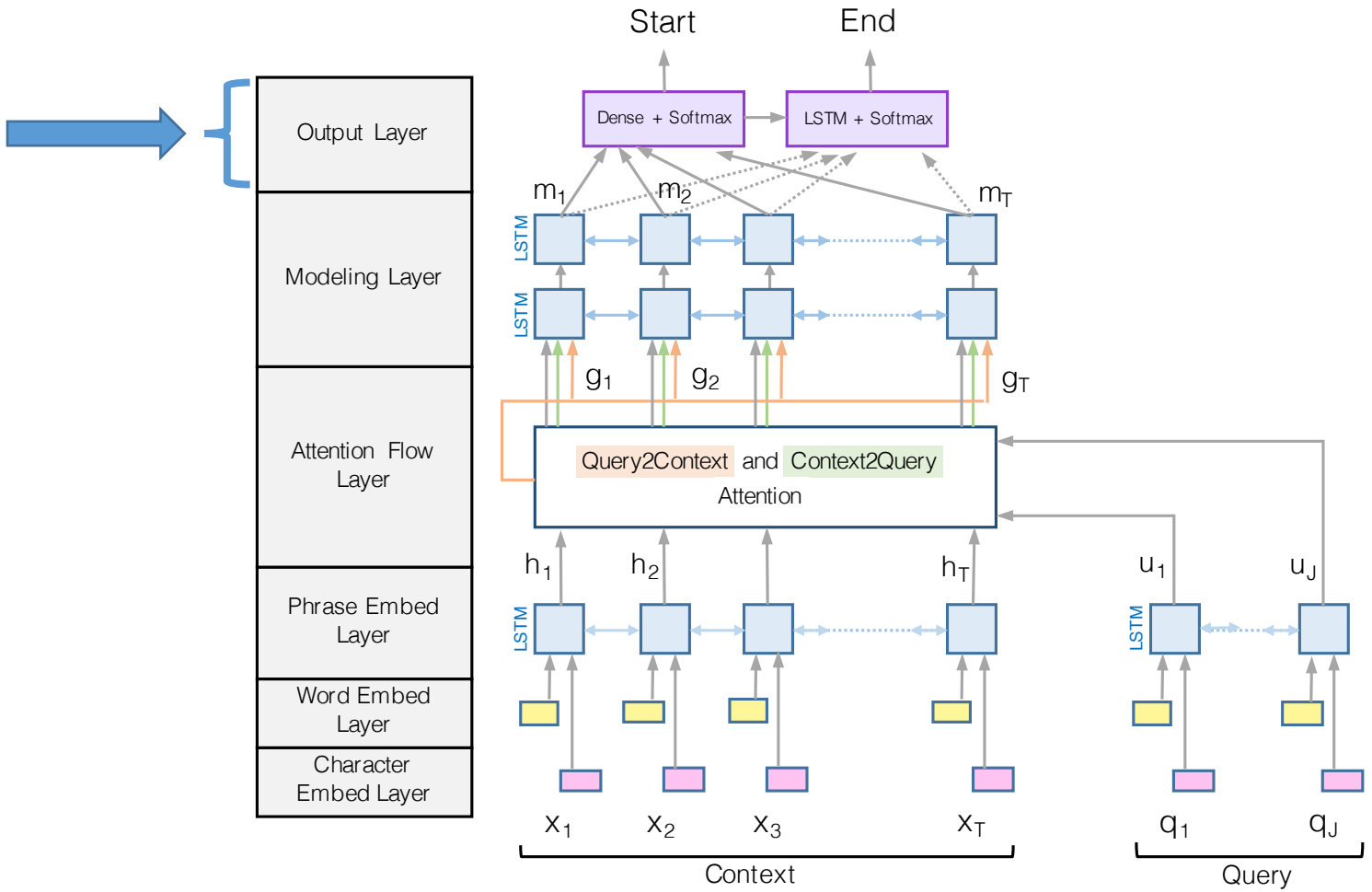


Modeling Layer

- **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

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

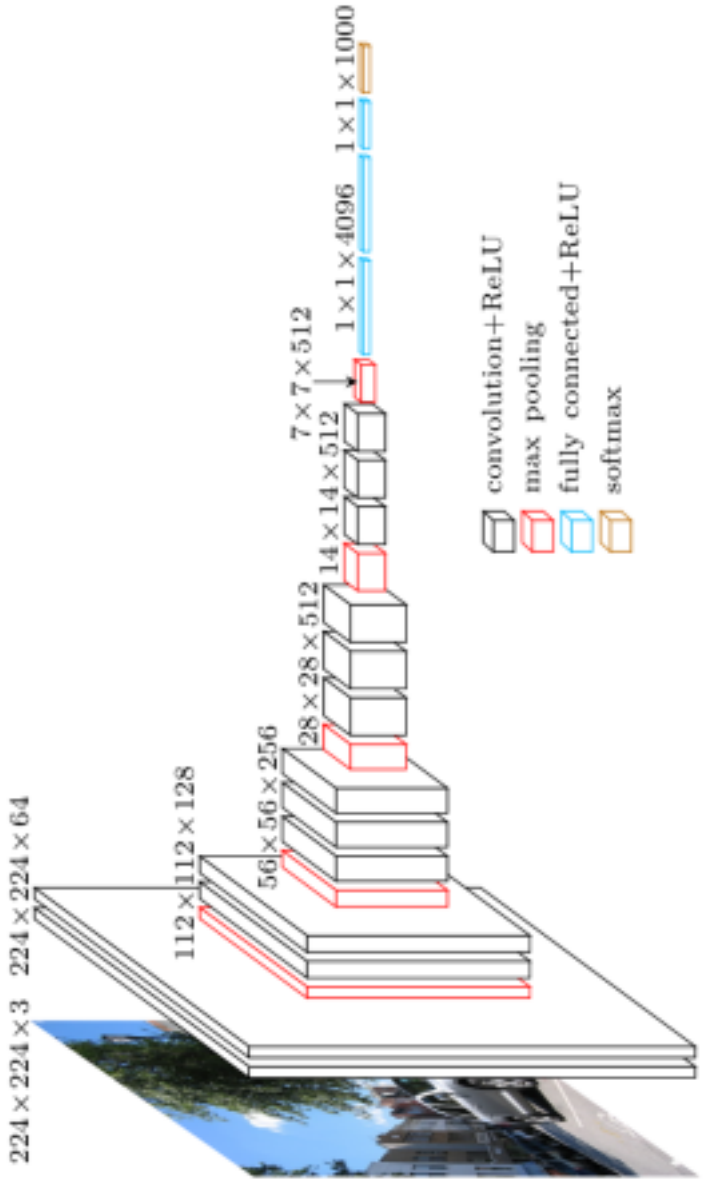
y_i^1	True start index of example i
y_i^2	True end index of example i
$\mathbf{p}^1$	Probability distribution of start index
$\mathbf{p}^2$	Probability distribution of stop index

Previous work

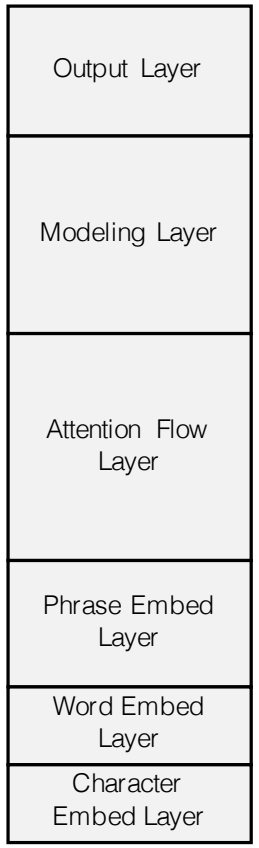
- Using neural attention as a controller (Xiong et al., 2016)
- Using neural attention within RNN (Wang & Jiang, 2016)
- Most of these attentions are uni-directional

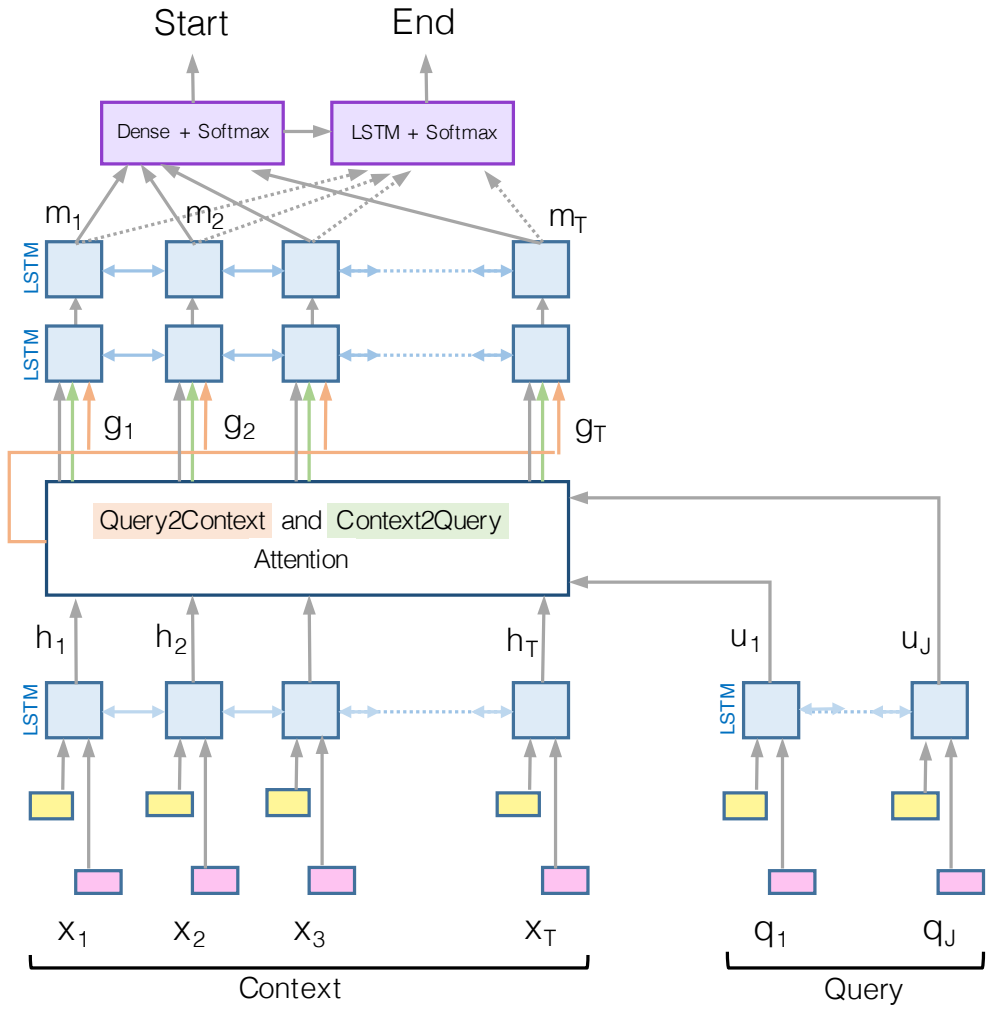
- BiDAF (our model)
  - uses neural attention *as a layer*,
  - Is separated from modeling part (RNN),
  - Is bidirectional

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 Dec 2

(ICLR 2017)

Test Set Leaderboard

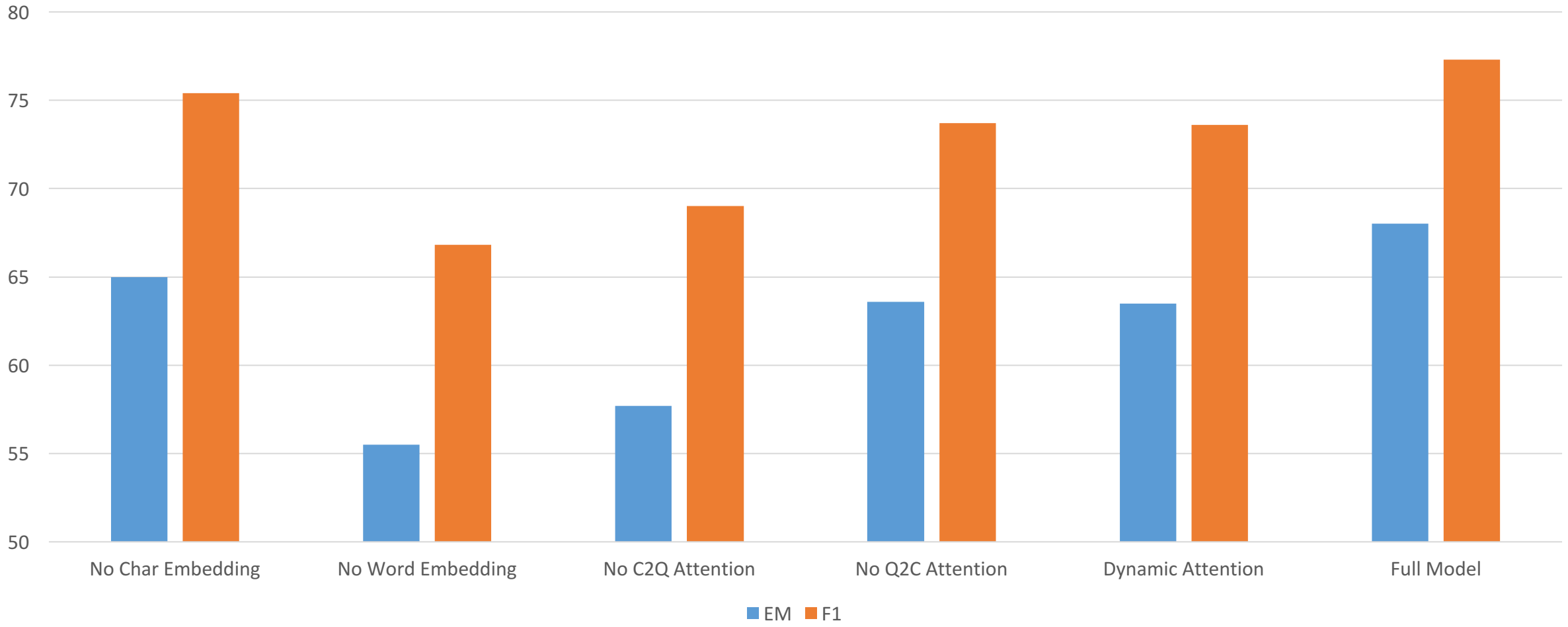
Since the release of our dataset (and paper), the community has made rapid progress! Here are the ExactMatch (EM) and F1 scores of the best models evaluated on the test and development sets of v1.1.

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

Now..

Rank	Model	EM	F1
1 Mar 2017	r-net (ensemble) *Microsoft Research Asia* http://aka.ms/rnet	76.922	84.006
2 May 2017	MEMEN (ensemble) *Eigen Technology & Zhejiang University*	75.37	82.658
3 Mar 2017	ReasoNet (ensemble) *MSR Redmond*	75.034	82.552
4 May 2017	r-net (single model) *Microsoft Research Asia* http://aka.ms/rnet	74.614	82.458
5 May 2017	Mnemonic Reader (ensemble) *NUDT & Fudan University* https://arxiv.org/abs/1705.02798	73.754	81.863
6 Apr 2017	SEDT+BiDAF (ensemble) *CMU* https://arxiv.org/abs/1703.00572	73.723	81.53
6 Feb 2017	BiDAF (ensemble) *Allen Institute for AI & University of Washington* https://arxiv.org/abs/1611.01603	73.744	81.525

Ablations on dev data

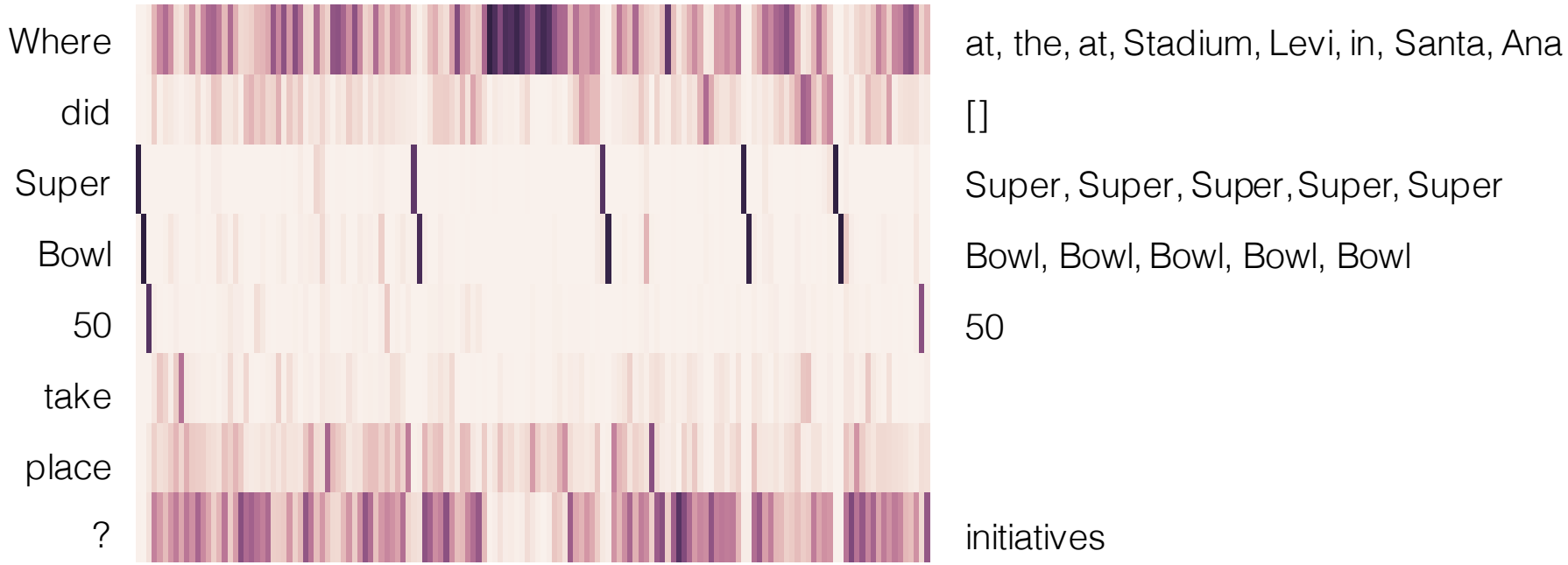


Interactive Demo

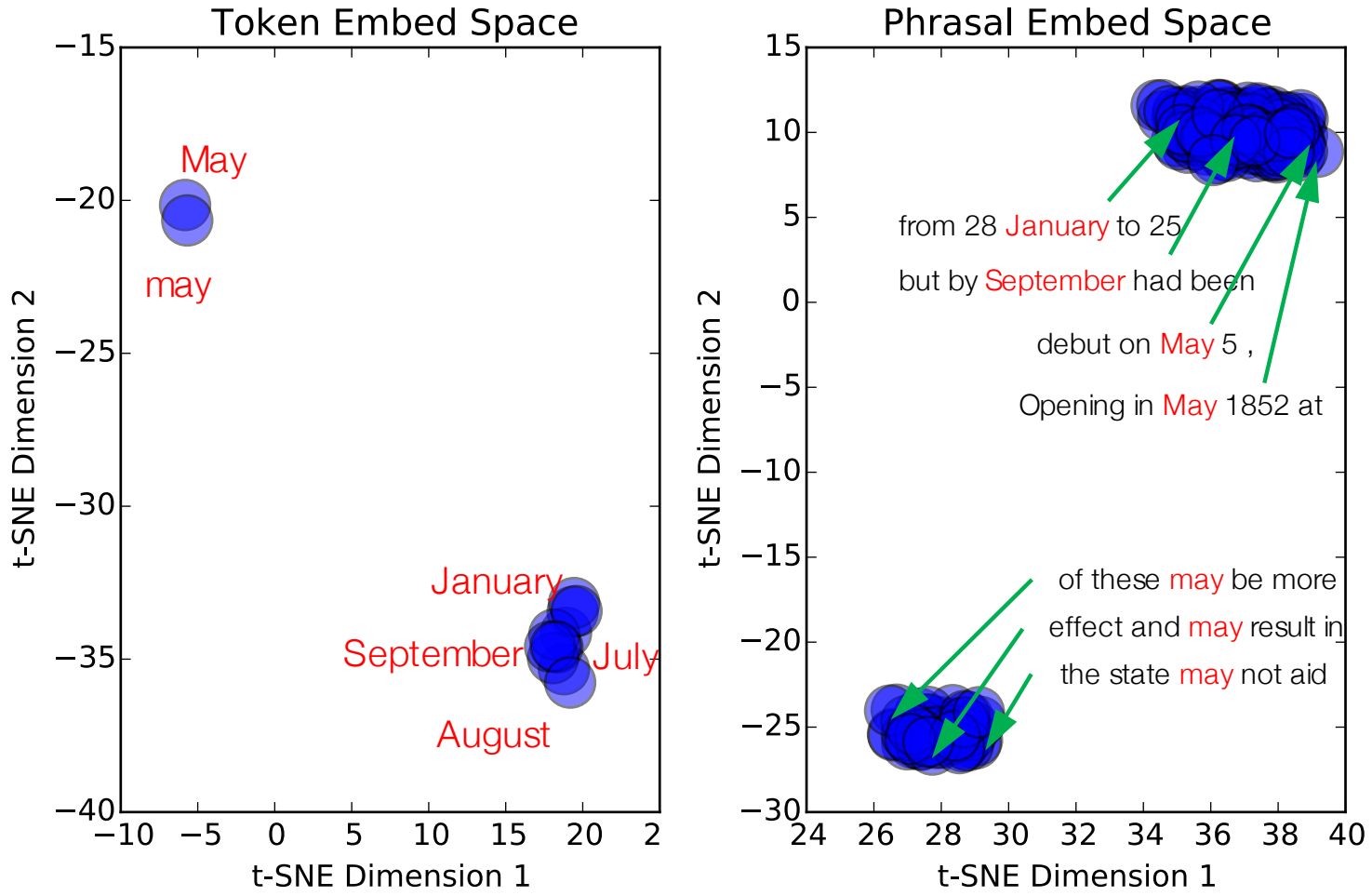
http://allenai.github.io/bi-att-flow/demo

Attention Visualizations

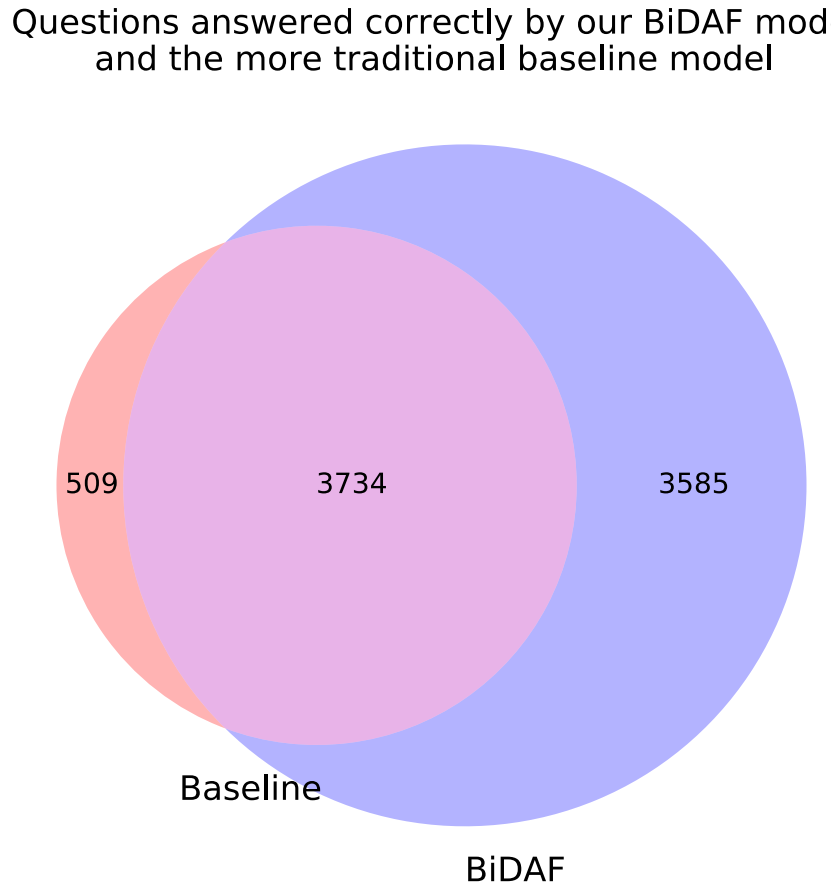
Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season . The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl 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 .

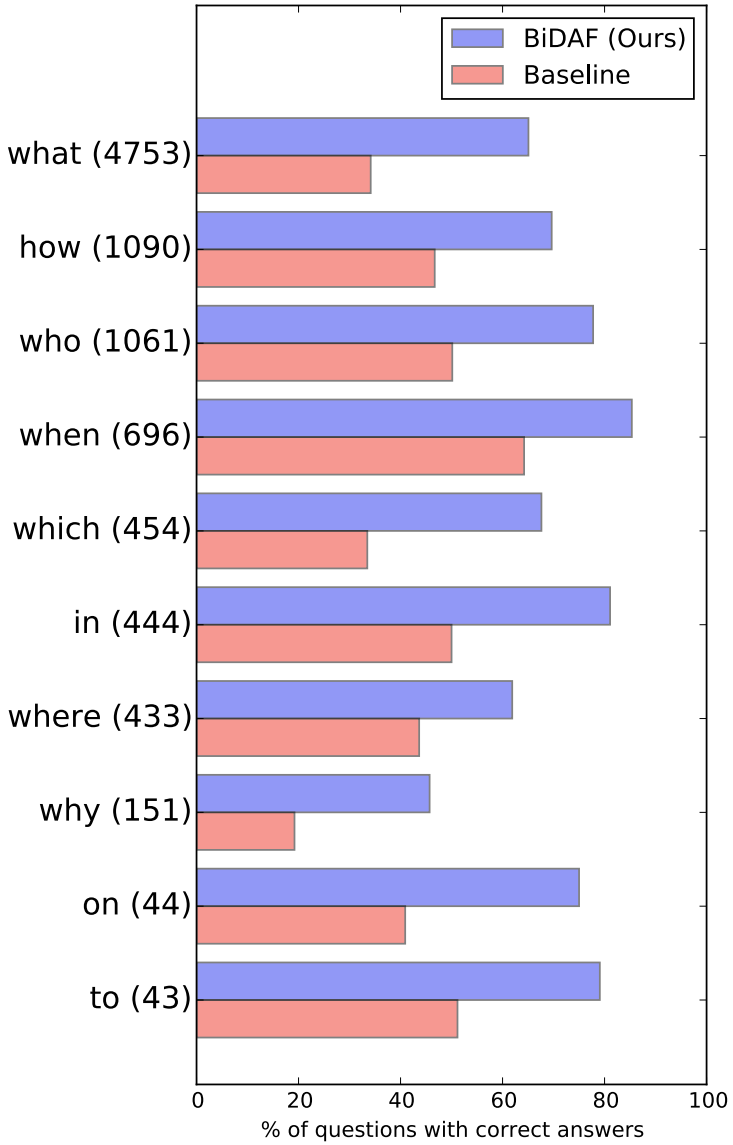


Embedding Visualization at Word vs Phrase Layers



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

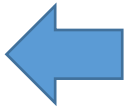
	CNN		DailyMail	
	val	test	val	test
Attentive Reader (Hermann et al., 2015)	61.6	63.0	70.5	69.0
MemNN (Hill et al., 2016)	63.4	6.8	-	-
AS Reader (Kadlec et al., 2016)	68.6	69.5	75.0	73.9
Stanford AR (Chen et al., 2016)	68.6	69.5	75.0	73.9
DER Network (Kobayashi et al., 2016)	71.3	72.9	-	-
Iterative Attention (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

Transfer Learning (ACL 2017)

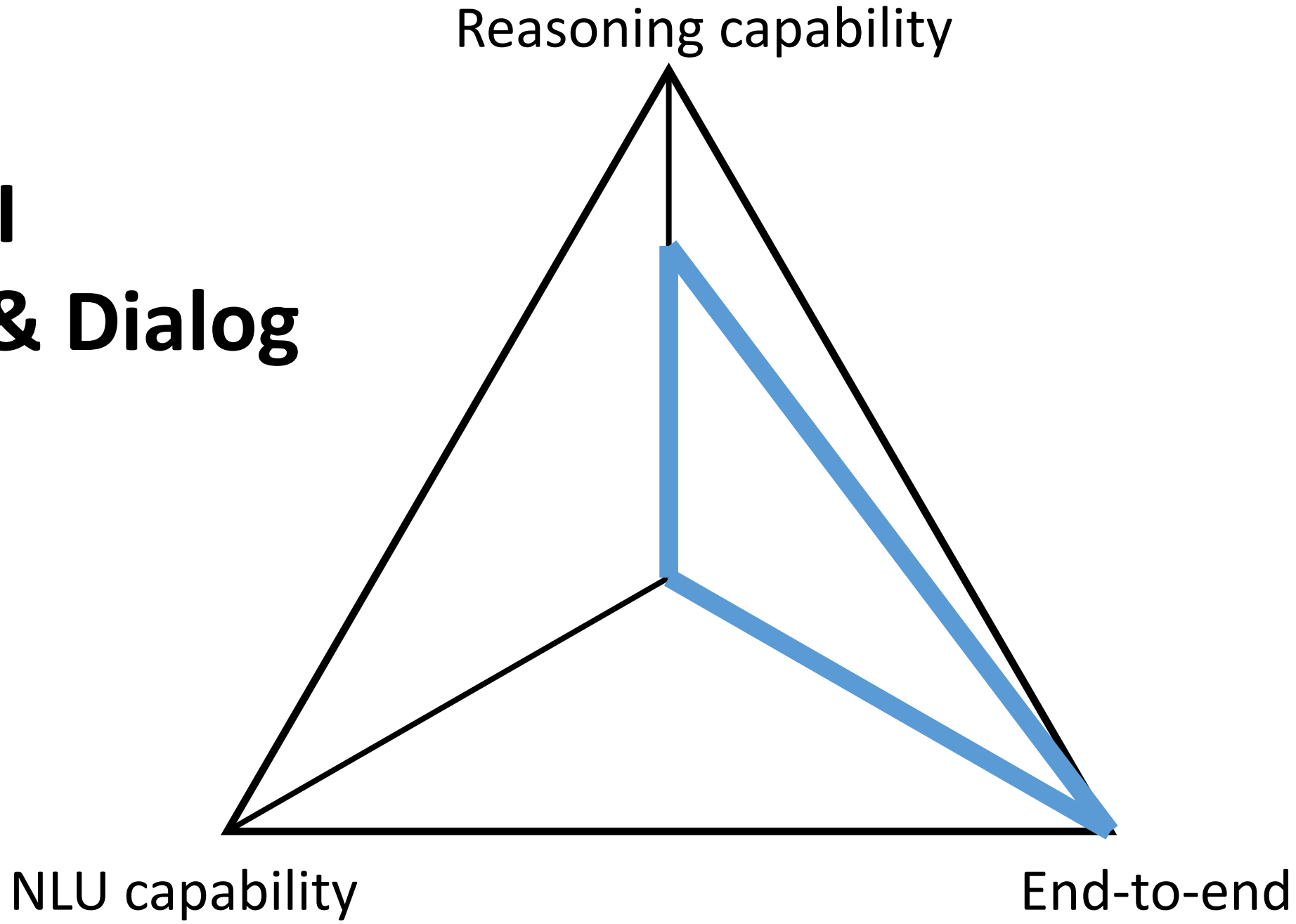
Pretrained dataset	Fine-tuned	WikiQA			SemEval-2016		
		MAP	MRR	P@1	MAP	MRR	AvgR
-	-	62.96	64.47	49.38	76.40	82.20	86.51
SQuAD-T	No	75.22	76.40	62.96	47.23	49.31	60.01
SQuAD	No	75.19	76.31	62.55	57.80	66.10	71.13
SQuAD-T	Yes	76.44	77.85	64.61	76.30	82.51	86.64
SQuAD	Yes	79.90	82.01	70.37	78.37	85.58	87.68
SQuAD*	Yes	**83.20**	**84.58**	**75.31**	**80.20**	**86.44**	**89.14**
Rank 1		74.33	75,45	-	79.19	86.42	88.82
Rank 2		74.17	75.88	64.61	77.66	84.93	88.05
Rank 3		70.69	72.65	-	77.58	85.21	88.14

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 crowdworkers' 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%



bAbI
QA & Dialog



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

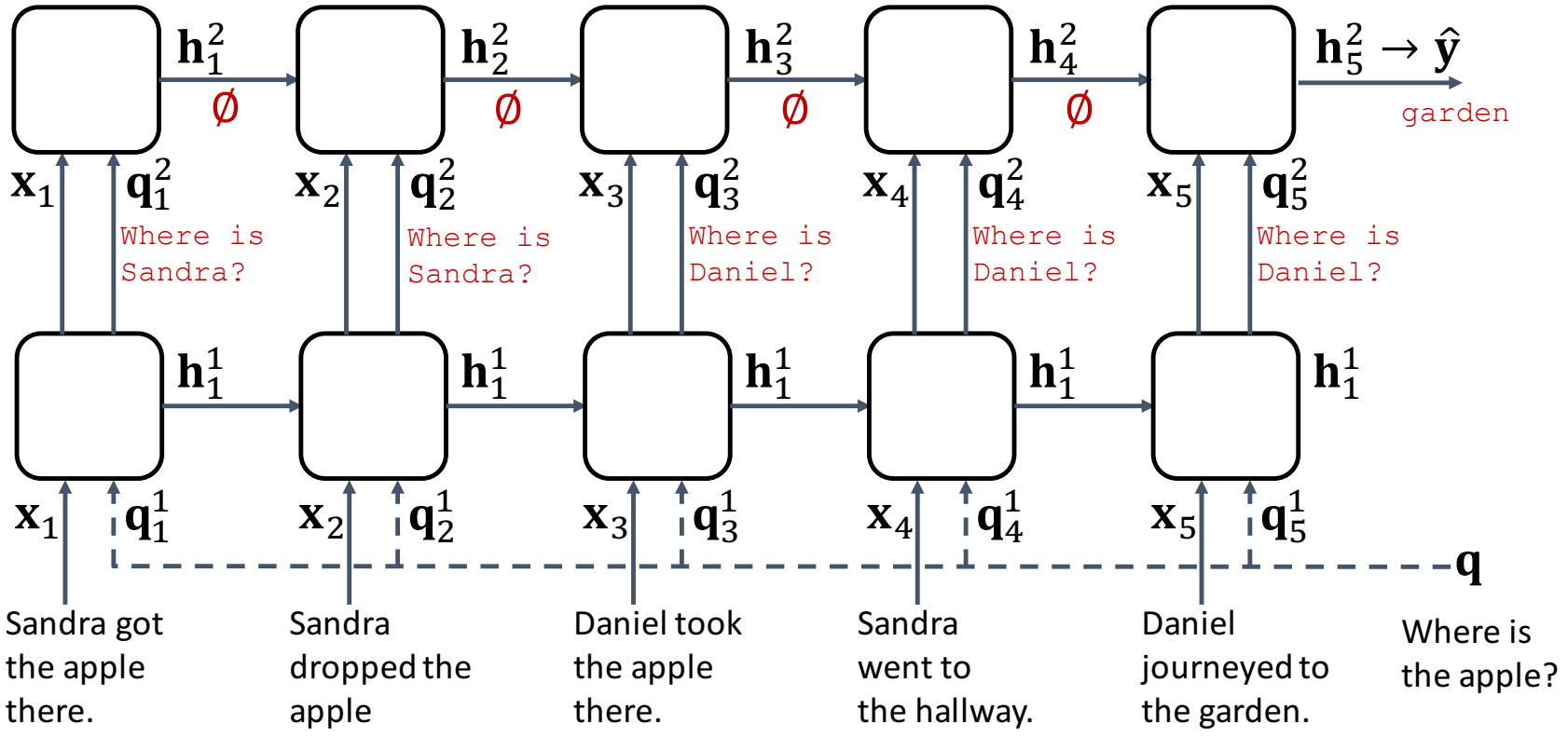
- Dialog system can be considered as QA system:
  - Last user's utterance is the query
  - All previous conversations are context to the query
  - The system's next response is the answer to the query
- Poses a few unique challenges
  - Dialog system requires tracking states
  - Dialog system needs to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

Our approach: Query-Reduction

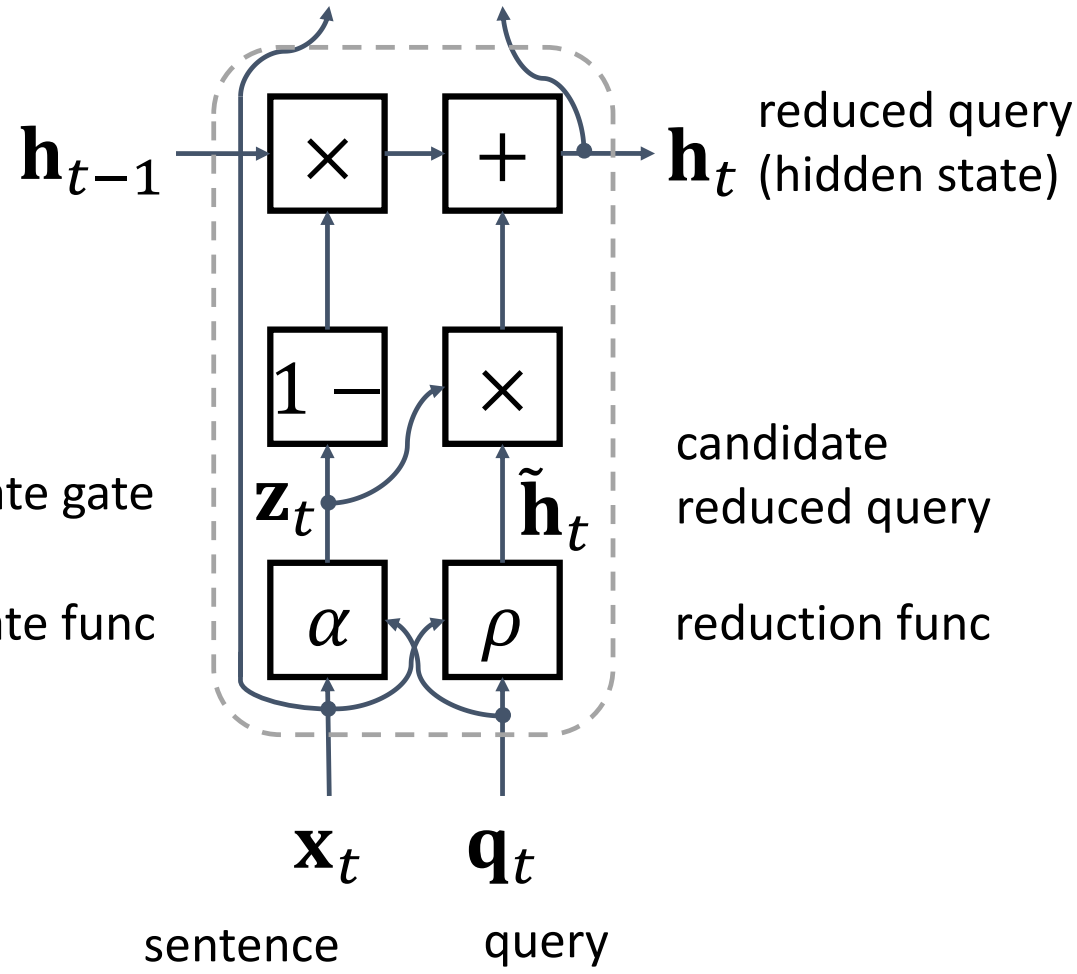
	Reduced query:
<START>	*Where is the apple?*
Sandra got the apple there.	*Where is Sandra?*
Sandra dropped the apple.	*Where is Sandra?*
Daniel took the apple there.	*Where is Daniel?*
Sandra went to the hallway.	*Where is Daniel?*
Daniel journeyed to the garden.	*Where is Daniel?* → *garden*
Q: Where 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

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

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

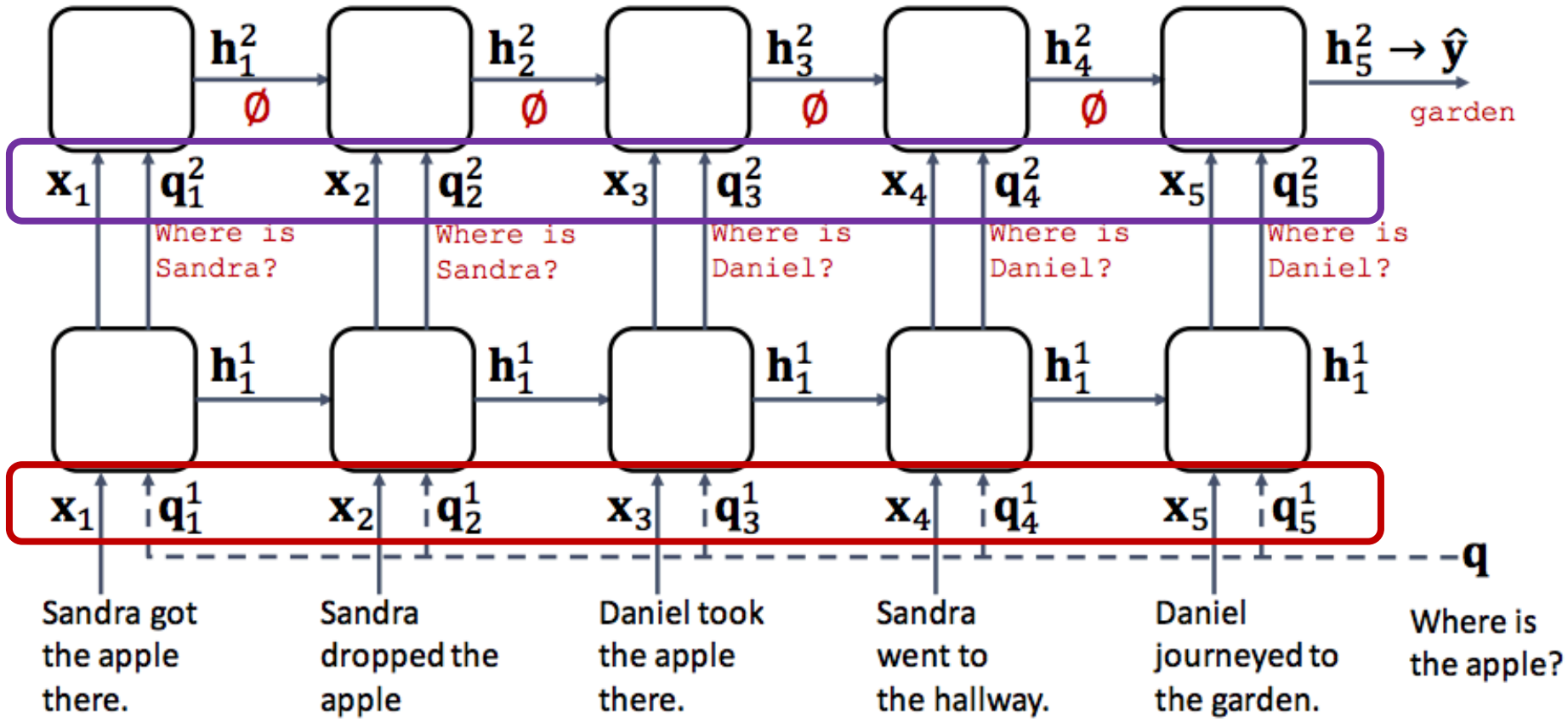
computed from inputs only, so can be trivially parallelized

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

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

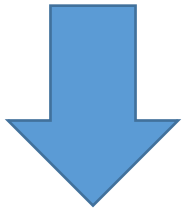
Can be explicitly expressed as the geometric sum of previous candidate hidden states

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.

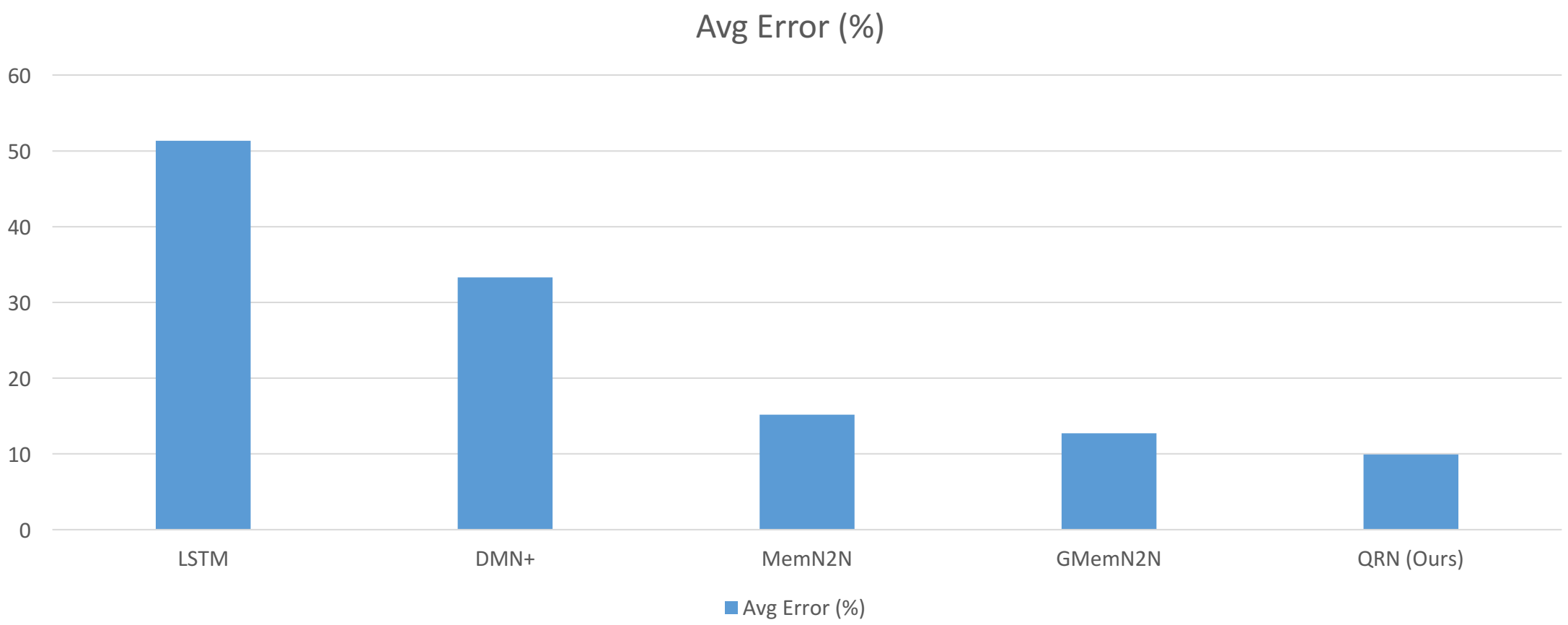
bAbI QA Dataset

- 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

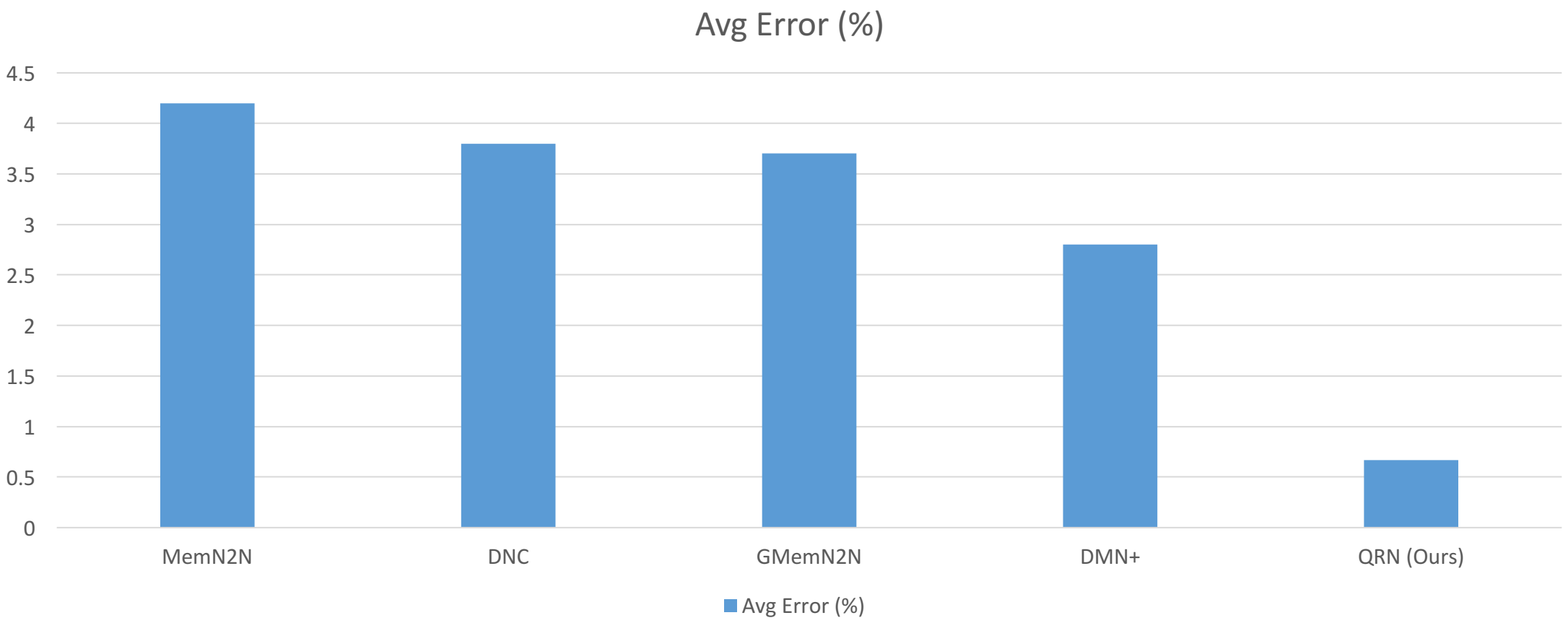
What's different from SQuAD?

- 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) (ICLR 2017)



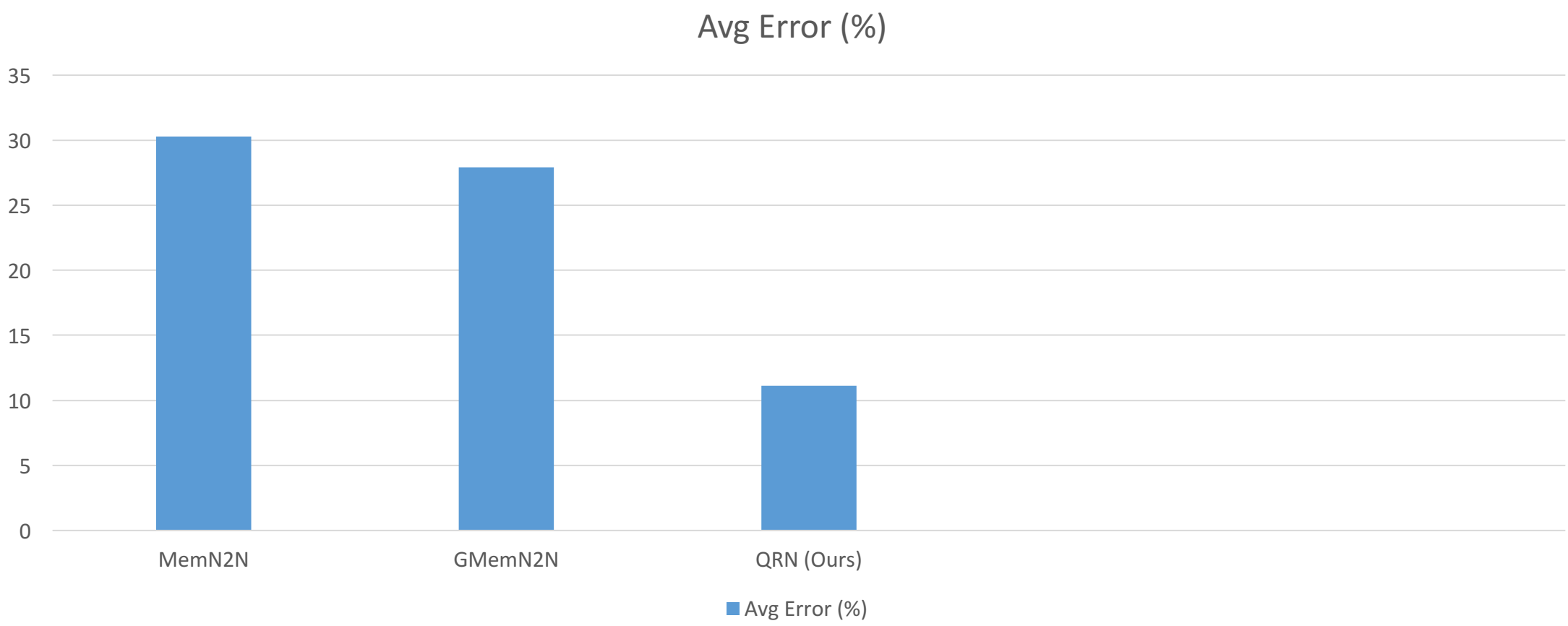
bAbI QA Results (10k)



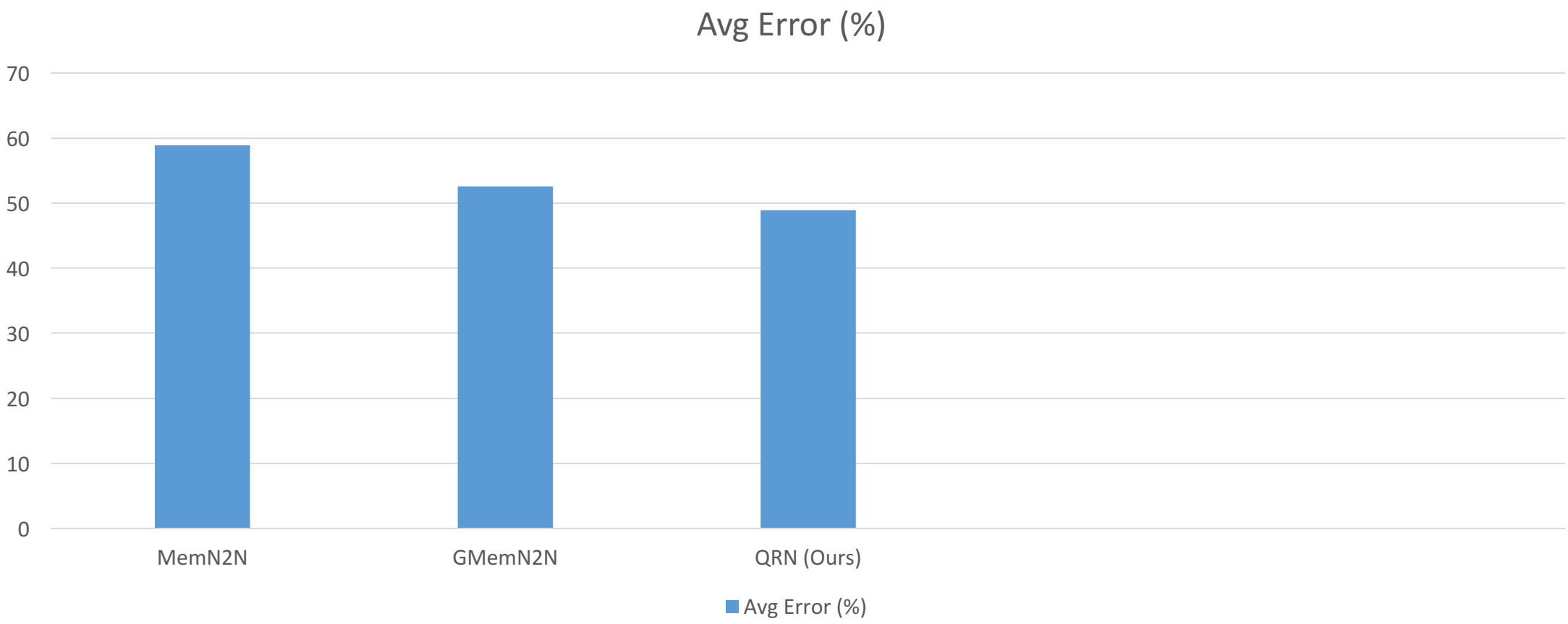
Dialog Datasets

- 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*

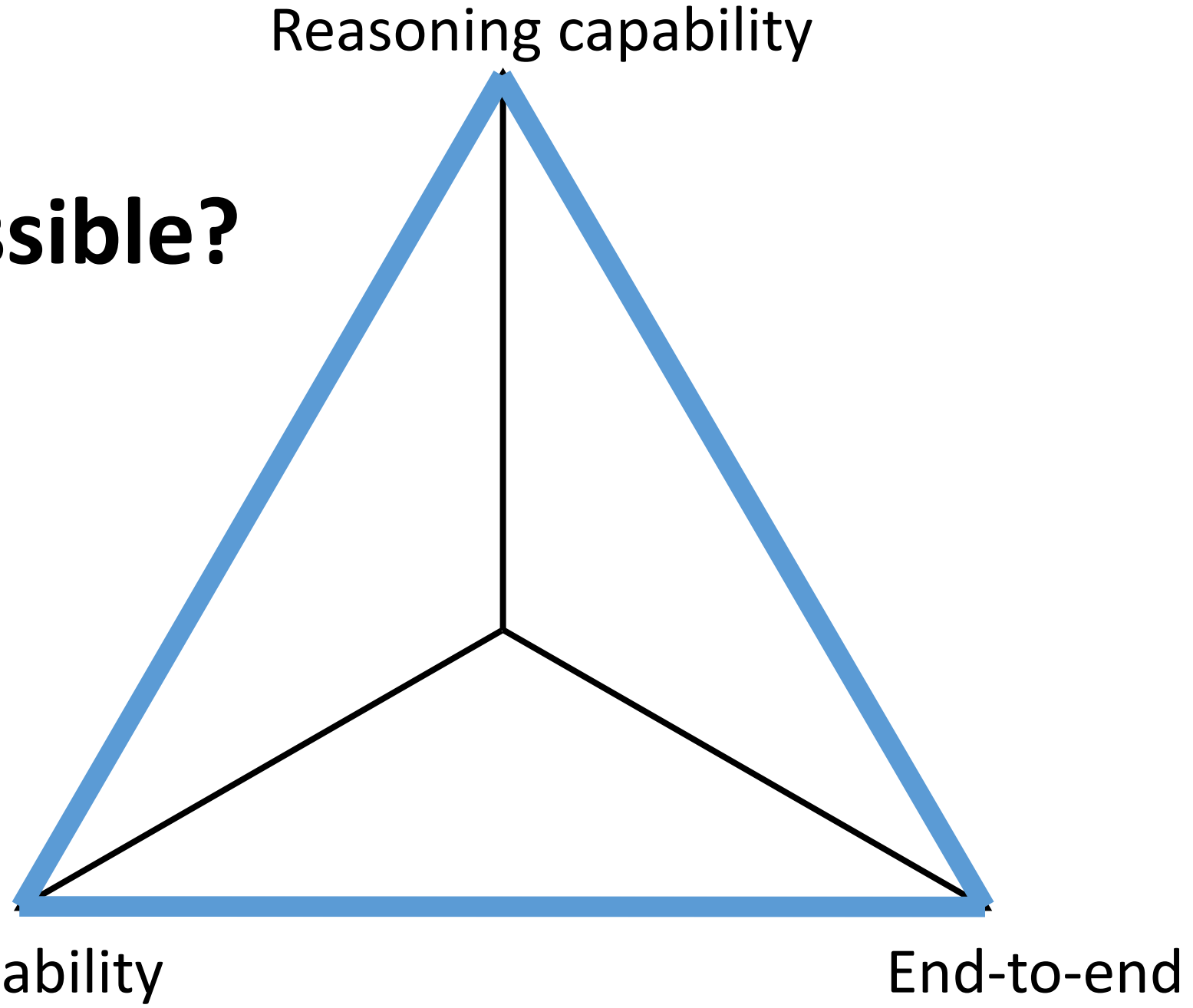
DSTC2 (Dialog) Visualization

	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				

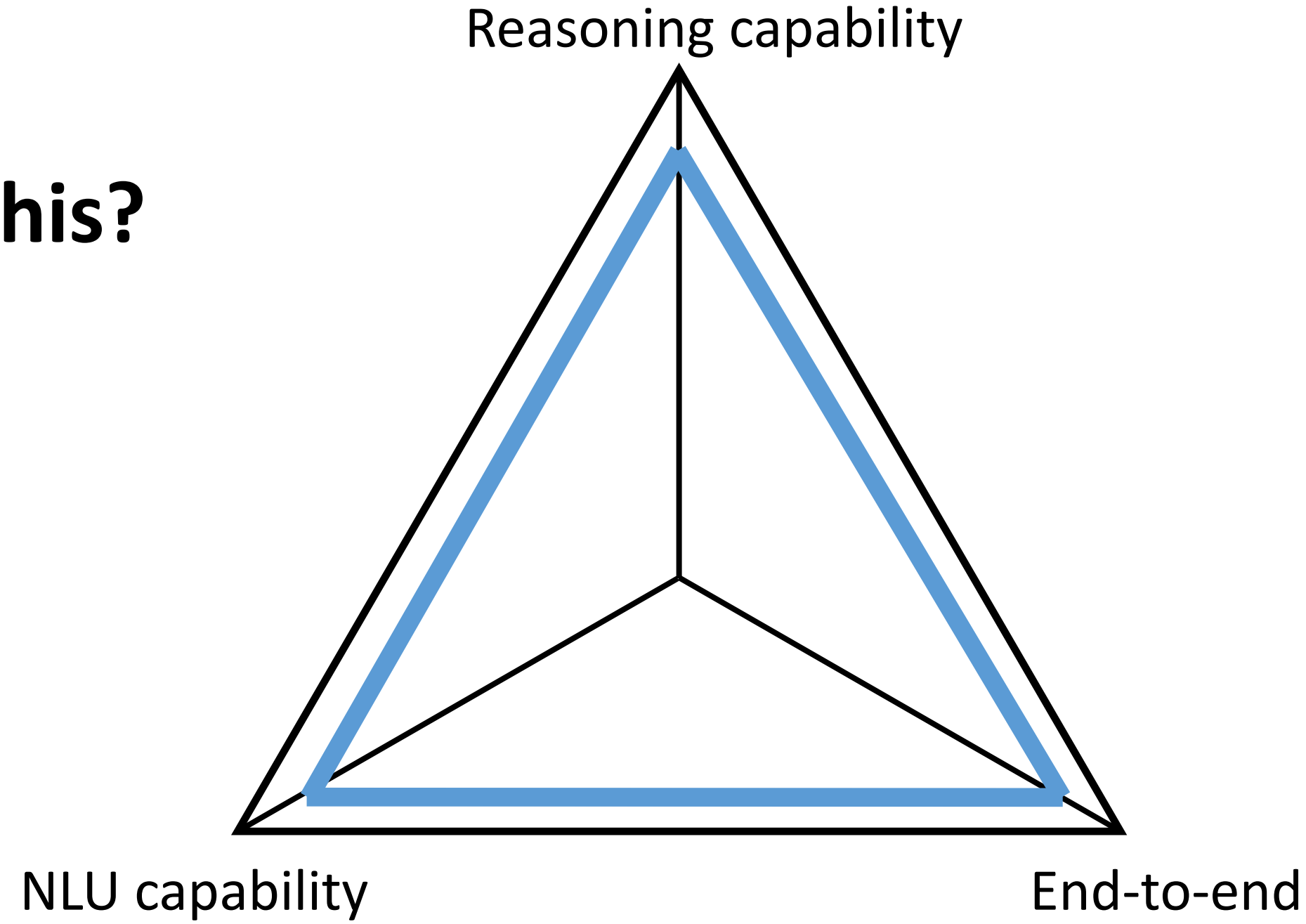
$z^l =$ Local attention (update gate) at layer *l*

So…

Is this possible?



Or this?



So… What should we do?

- **Disclaimer**: completely subjective!

- **Logic (reasoning) is discrete**
- **Modeling logic with differentiable model is hard**
  - *Relaxation*: either hard to optimize or converge to bad optimum (low generalization error)
  - *Estimation*: Low-bias or low-variance methods are proposed (Williams, 1992; Jang et al., 2017), but improvements are not substantial.
  - *Big data*: how much do we need? Exponentially many?
  - Perhaps new paradigm is needed…

ASK ME ANYTHING

“If you got a billion dollars to spend on a huge research project, what would you like to do?”

“I'd use the billion dollars to build a NASA-size program focusing on *natural language processing* (NLP), in all of its glory (semantics, pragmatics, etc).”

Michael Jordan
Professor of Computer Science
UC Berkeley

Towards Artificial General Intelligence…

Natural language is the best tool to describe and communicate “thoughts”

Asking and answering questions is an effective way to develop deeper “thoughts”

Thank you!

- minjoon@cs.uw.edu
- http://seominjoon.github.io