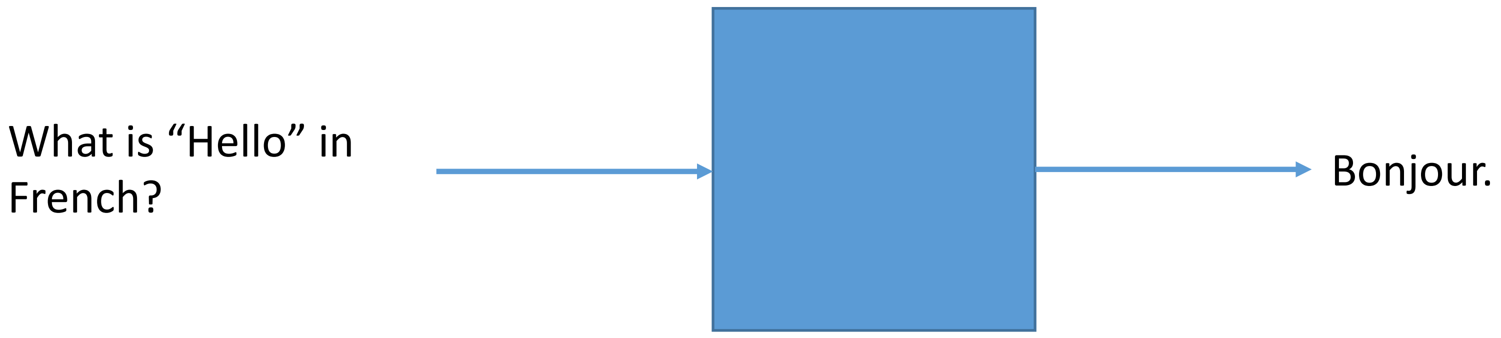
Towards End-to-End Reasoning for Question Answering

Minjoon Seo
University of Washington
September 21, 2016
@ SK T-Brain

What is reasoning?

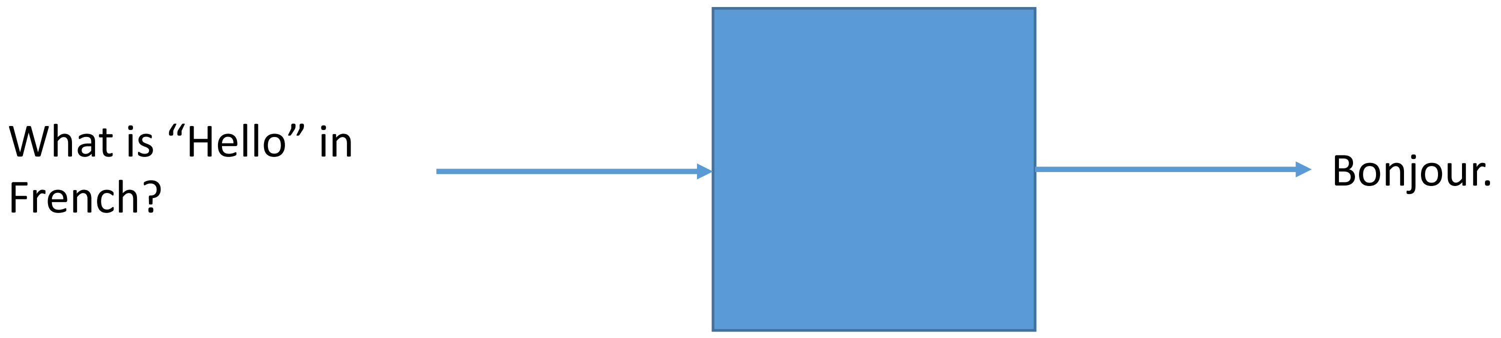
Simple Question Answering Model



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

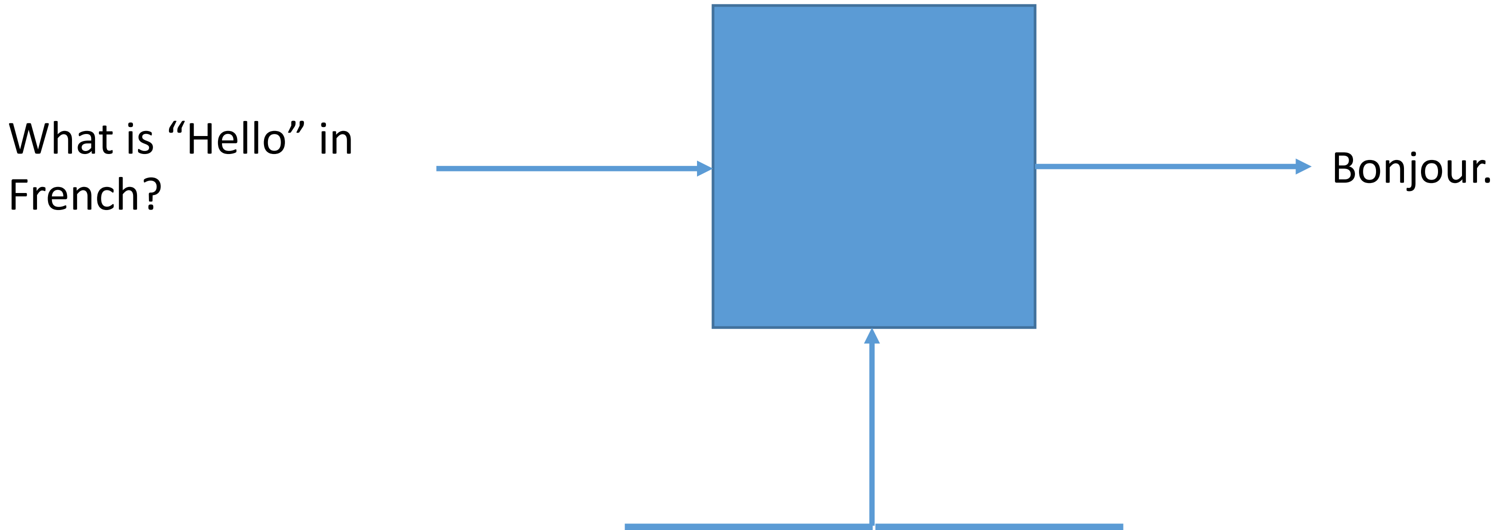
Simple Question Answering Model



Problem: parametric model has finite, pre-defined capacity.

“You can’t even fit in an entire sentence into a single small vector!”

QA Model with Context



English	French
Hello	Bonjour
Thank you	Merci

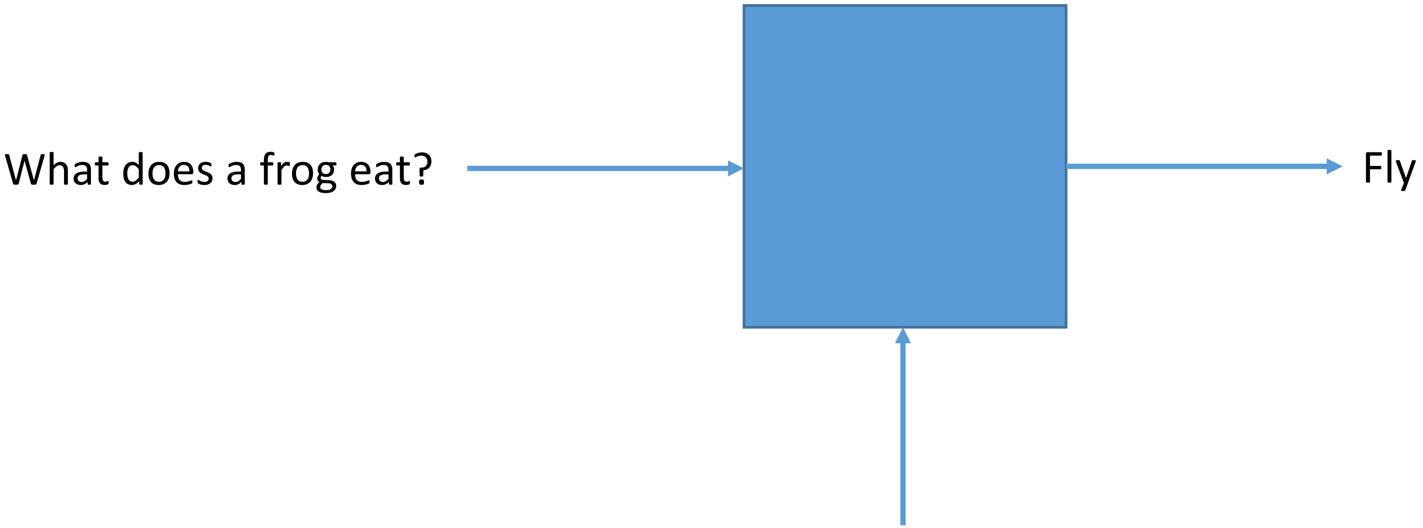
Context (Knowledge Base)

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

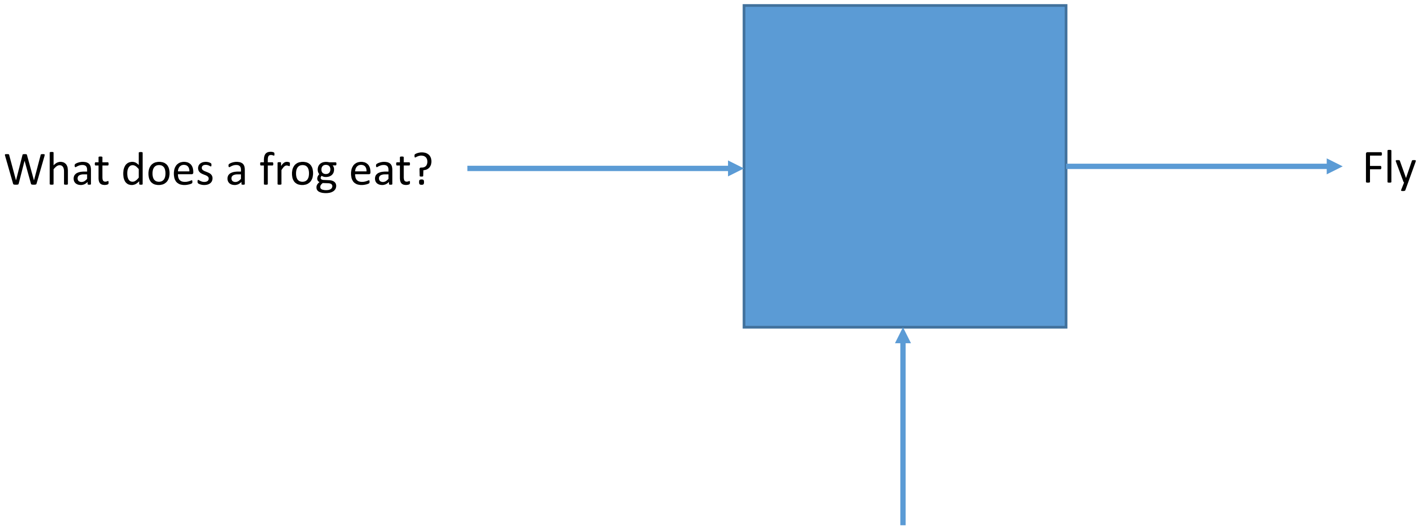


Eats	IsA
(Amphibian, insect)	(Frog, amphibian)
(insect, flower)	(Fly, insect)

Context (Knowledge Base)

Something is missing …

QA Model with Reasoning Capability



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

First Order Logic
IsA(A, B) ^ IsA(C, D) ^ Eats(B, D) → Eats(A, C)

Context (Knowledge Base)

Examples

- Semantic parsing
  - GeoQA (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 factoid QA and reasoning QA

- Factoid:
  - 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!

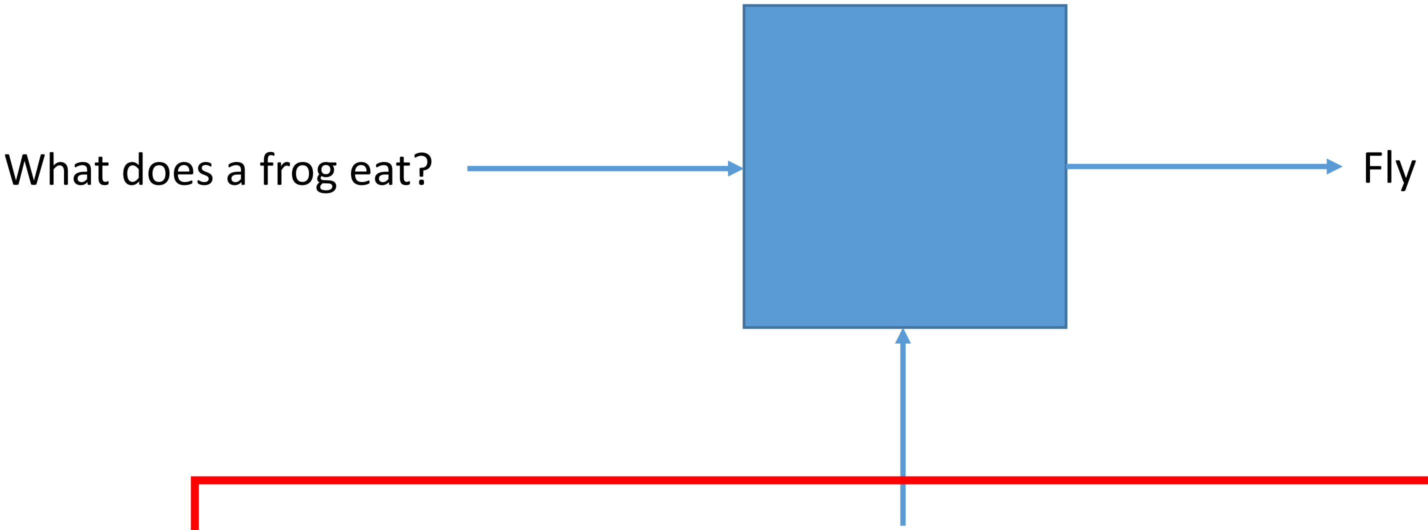
If our objective is to “answer” difficult questions …

- 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



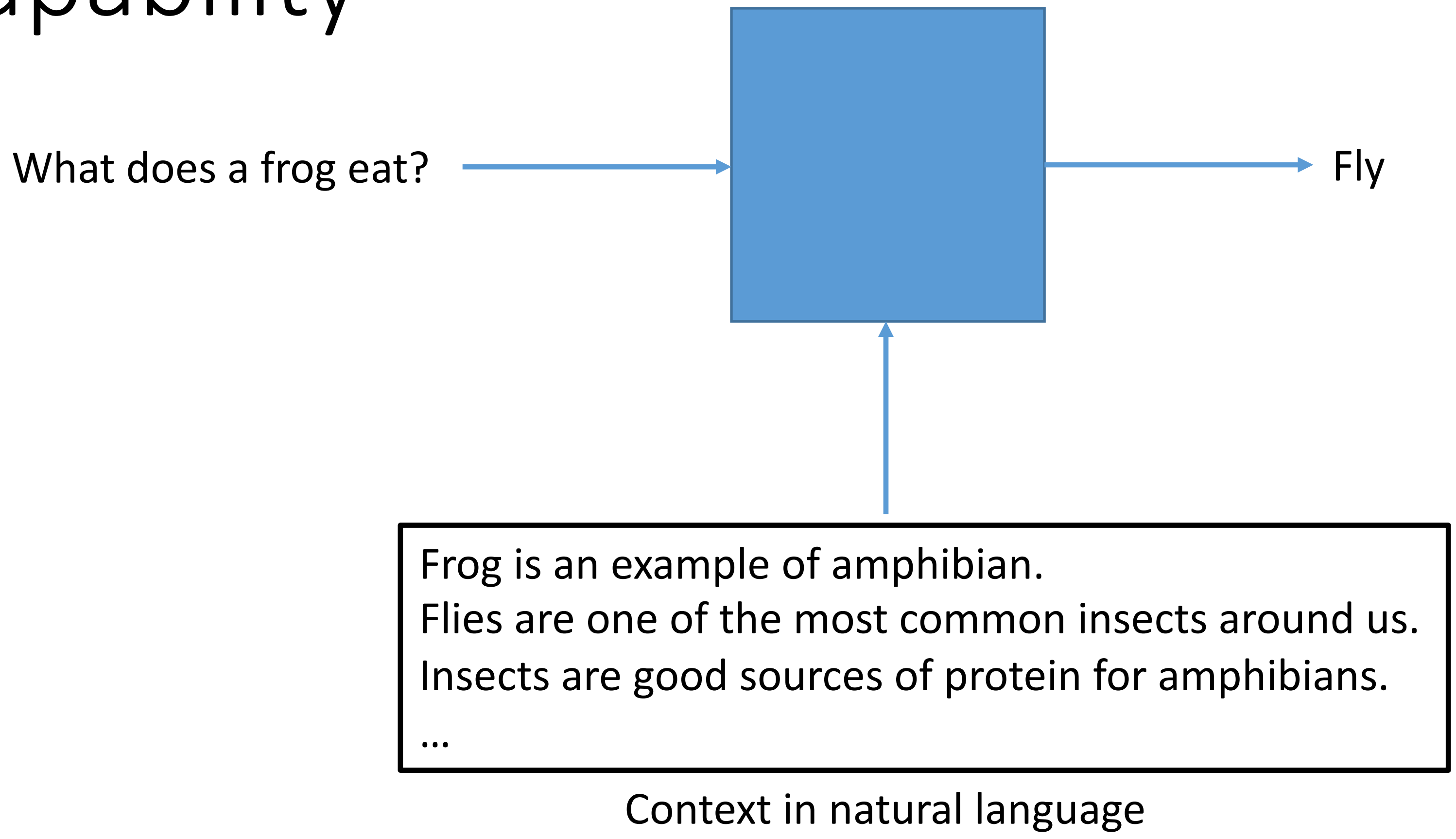
Eats	IsA	First Order Logic
(Amphibian, insect)	(Frog, amphibian)	IsA(A, B) ^ IsA(C, D) ^ Eats(B, D) → Eats(A, C)
(insect, flower)	(Fly, insect)	

Context (Knowledge Base)

Who makes this?

Tell me it's not me …

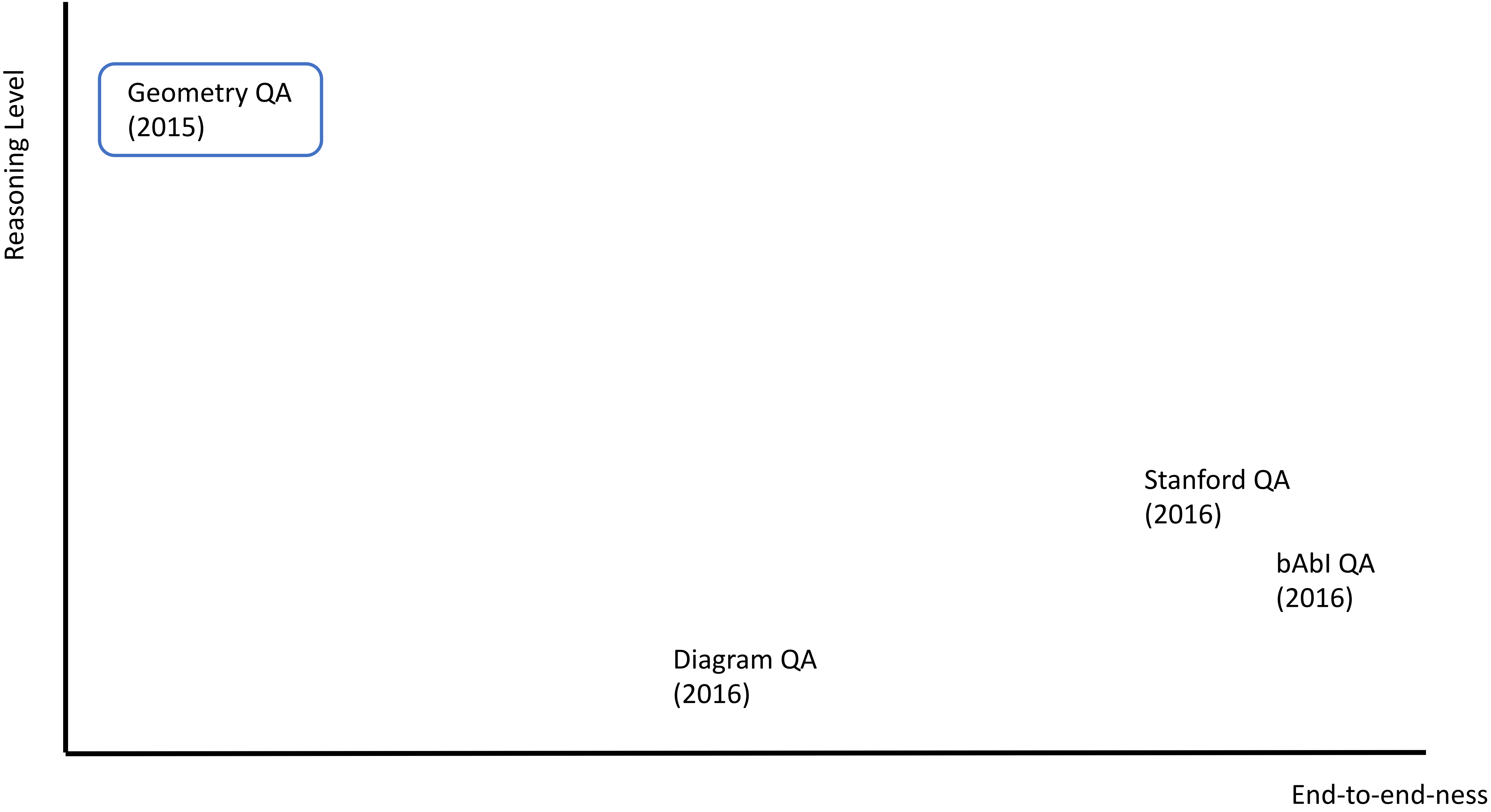
End-to-end QA Model with Reasoning Capability



Is end-to-end always feasible?

- **No**. End-to-end systems perform poorly if either:
  - Data is limited
  - Logic is super complicated

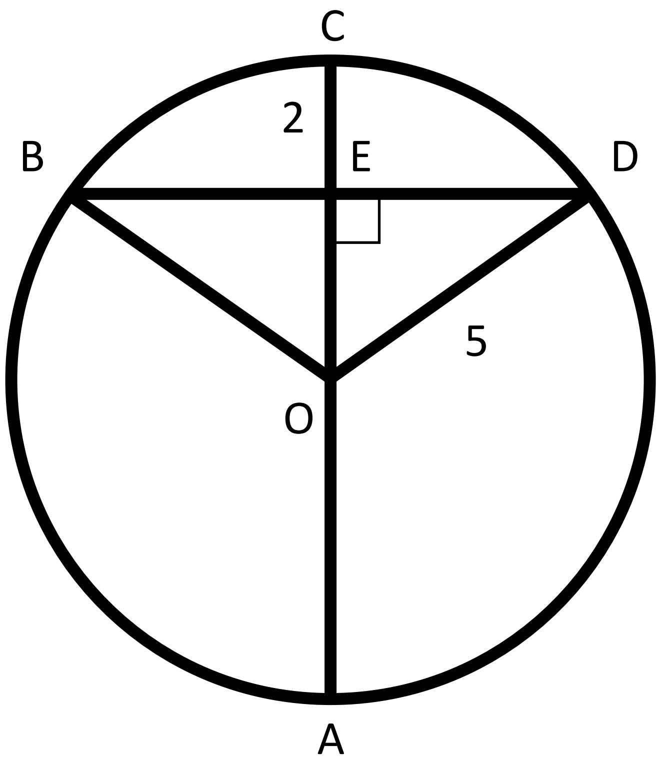
- But not hopeless.


Reasoning Level
Geometry QA (2015)
Stanford QA (2016)
bAbI QA (2016)
Diagram QA (2016)
End-to-end-ness

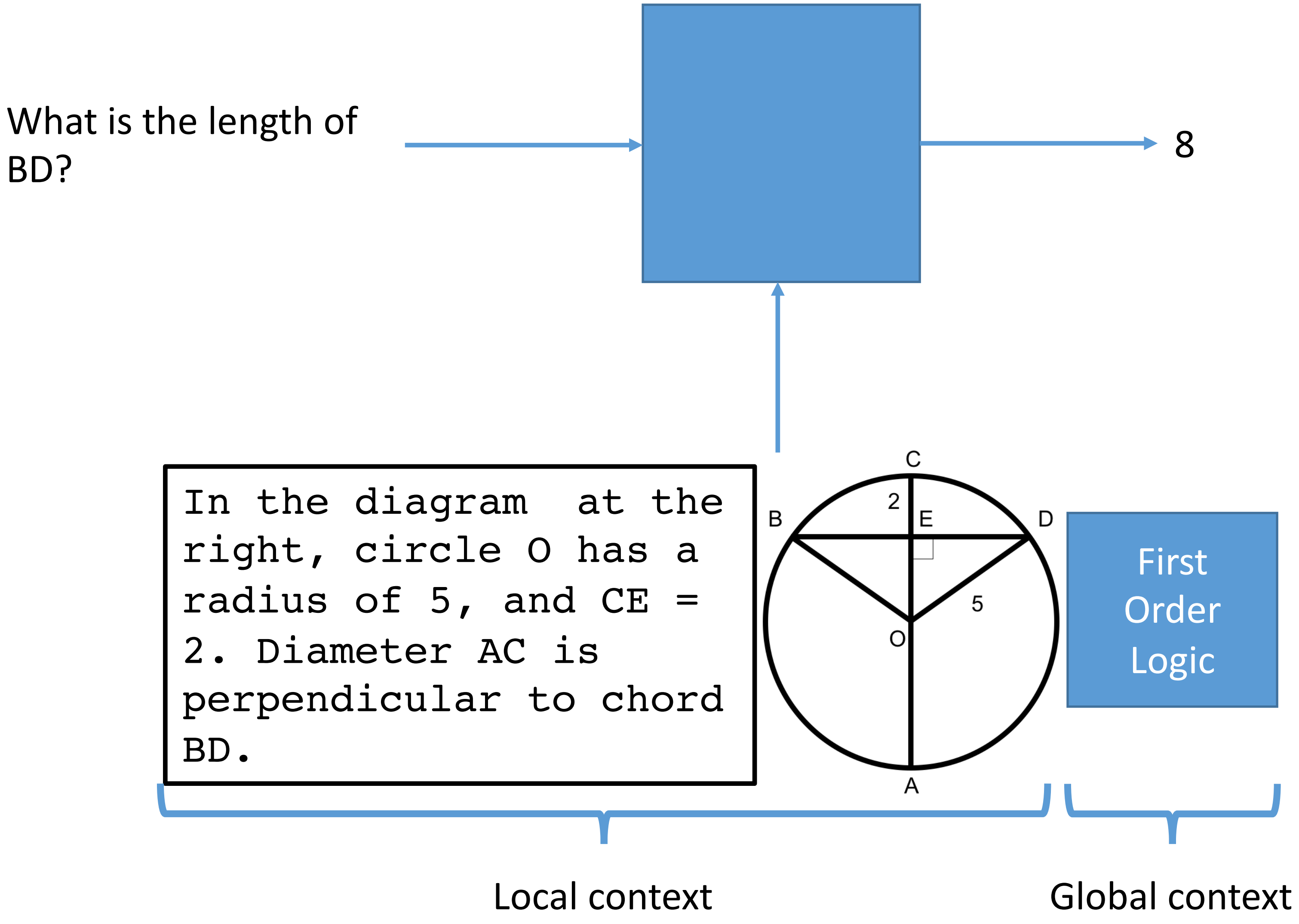
Geometry QA

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



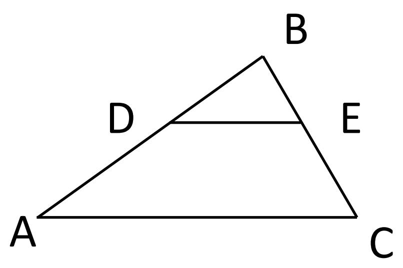
Method

- 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

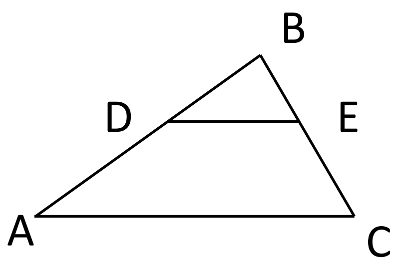
IsTriangle(ABC) ∧ Parallel(AC, DE) ∧ Equals(LengthOf(DB), 4) ∧ Equals(LengthOf(AD), 8) ∧ Equals(LengthOf(DE), 5) ∧ Find(LengthOf(AC))

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

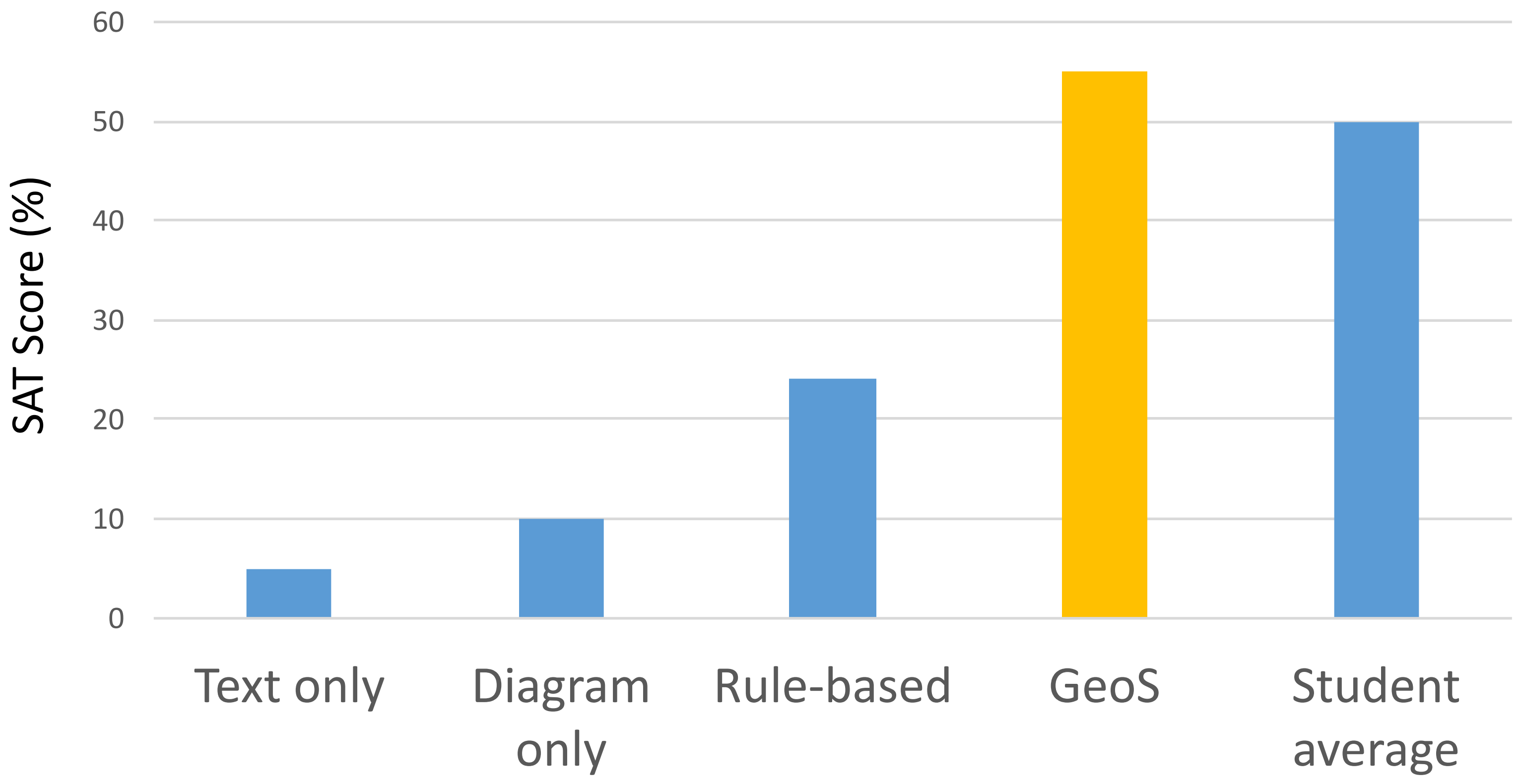
Literal	Equation
Equals(LengthOf(AB),d)	$(A_x-B_x)^2+(A_y-B_y)^2-d^2 = 0$
Parallel(AB, CD)	$(A_x-B_x)(C_y-D_y)-(A_y-B_y)(C_x-D_x) = 0$
PointLiesOnLine(B, AC)	$(A_x-B_x)(B_y-C_y)-(A_y-B_y)(B_x-C_x) = 0$
Perpendicular(AB,CD)	$(A_x-B_x)(C_x-D_x)+(A_y-B_y)(C_y-D_y) = 0$

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

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?)

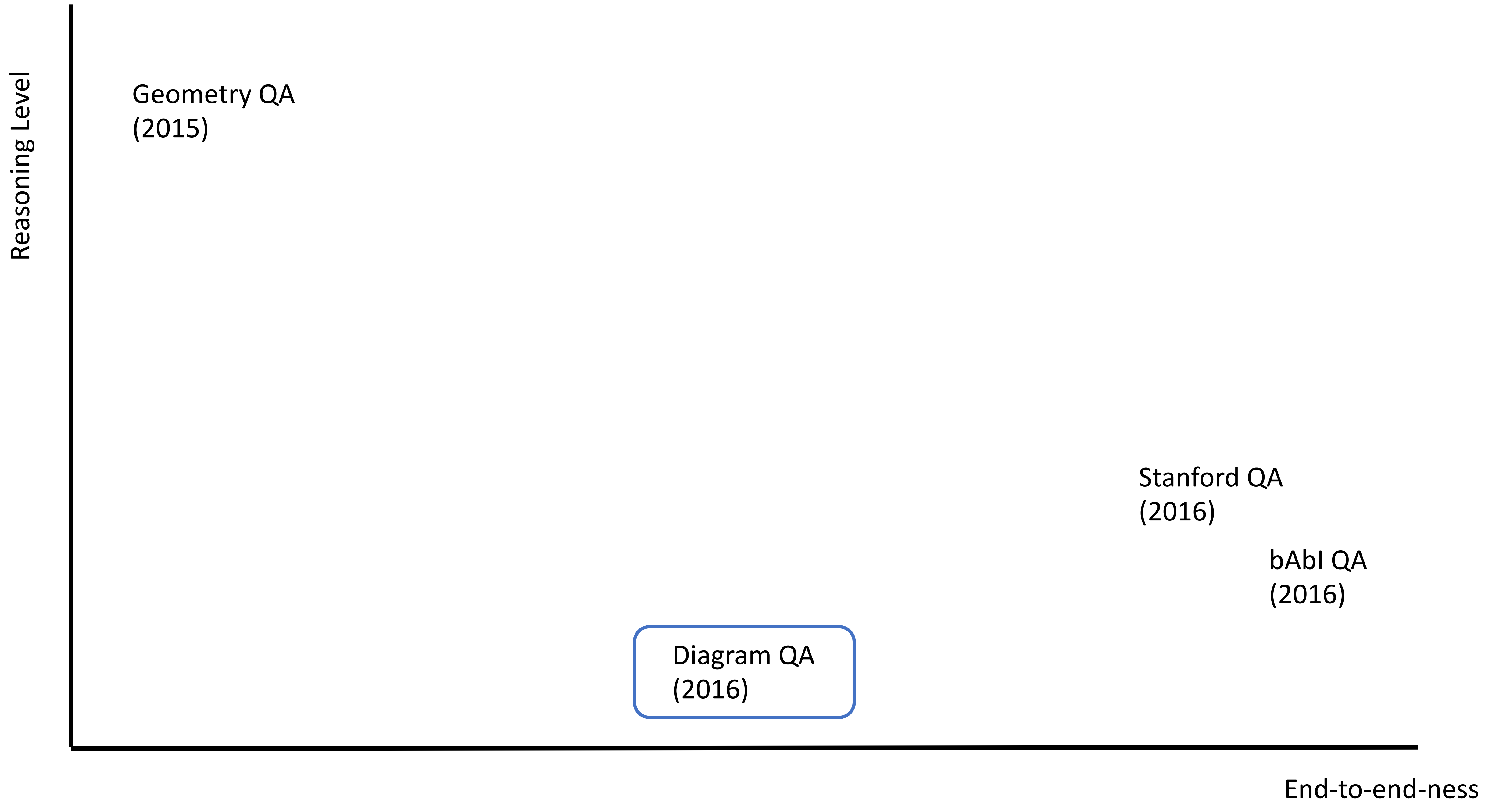
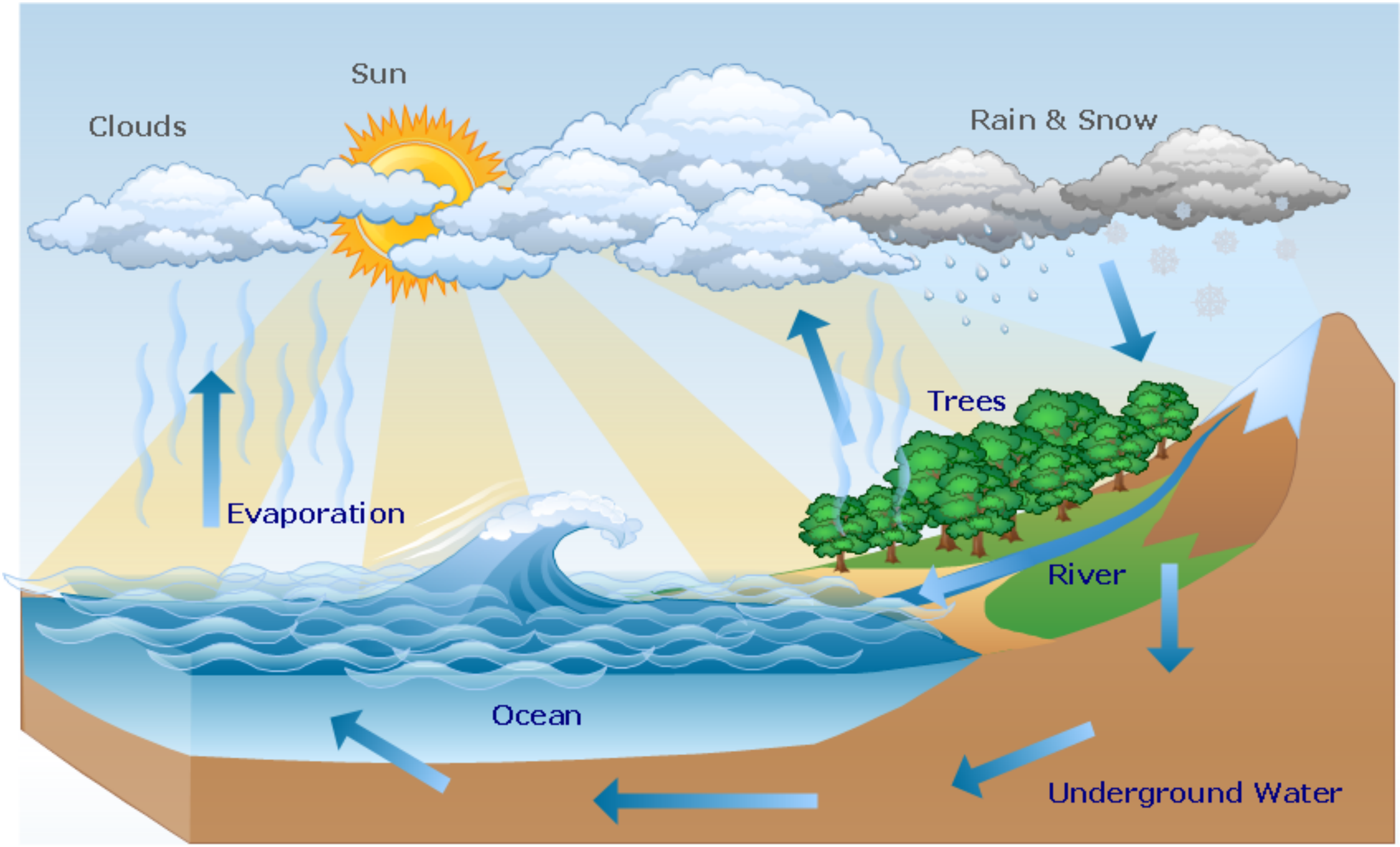

Reasoning Level
Geometry QA (2015)
Stanford QA (2016)
bAbI QA (2016)
Diagram QA (2016)
End-to-end-ness

Diagram QA



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

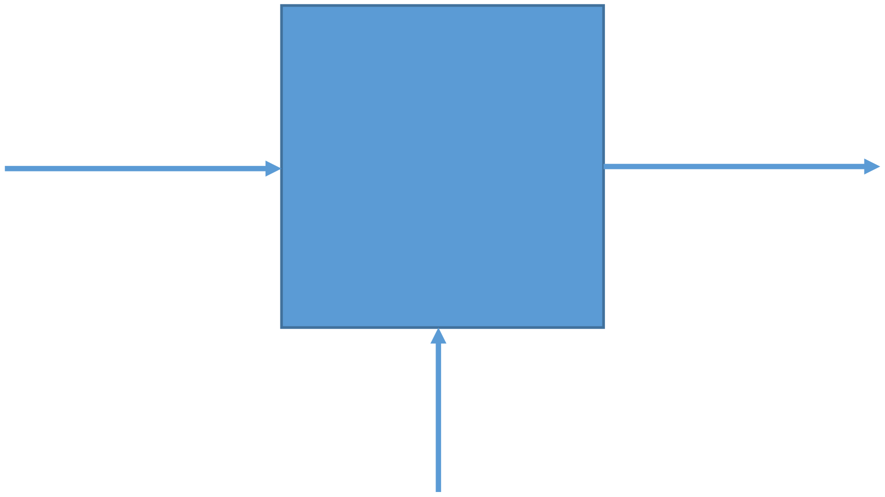
A: Evaporation

Is DQA subset of VQA?

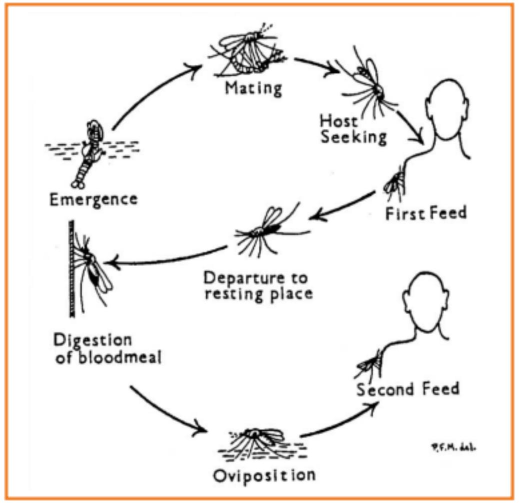
- Diagrams and real images are very different
- Diagram components are simpler than real images
- Diagram contains a lot of information in a single image
- Diagrams are few (whereas real images are almost infinitely many)

Problem

What comes before second feed?



8



Difficult to latently learn relationships

Strategy

What does a frog eat?

Fly

Diagram Graph

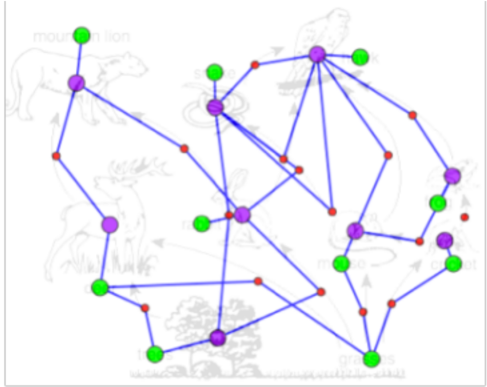
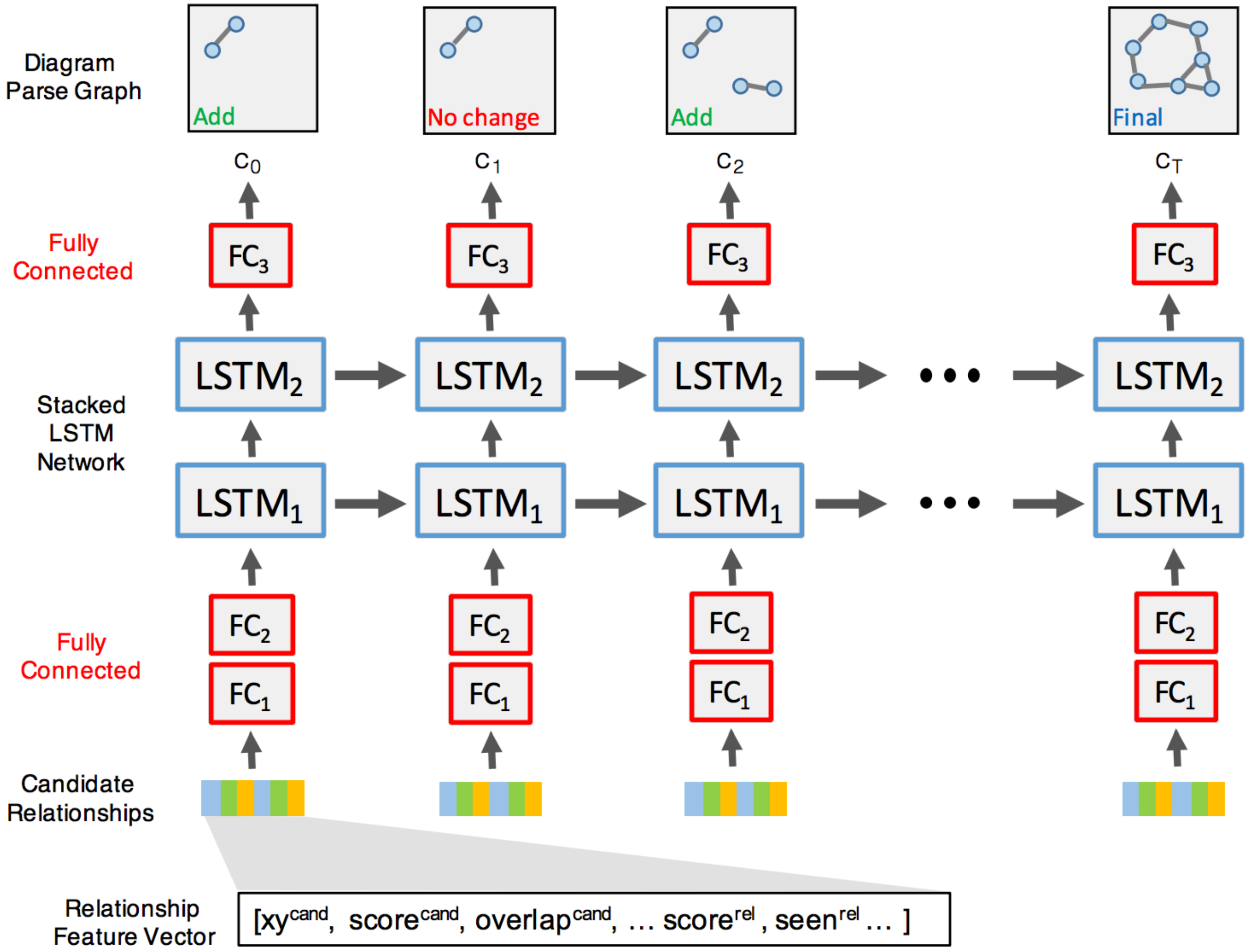
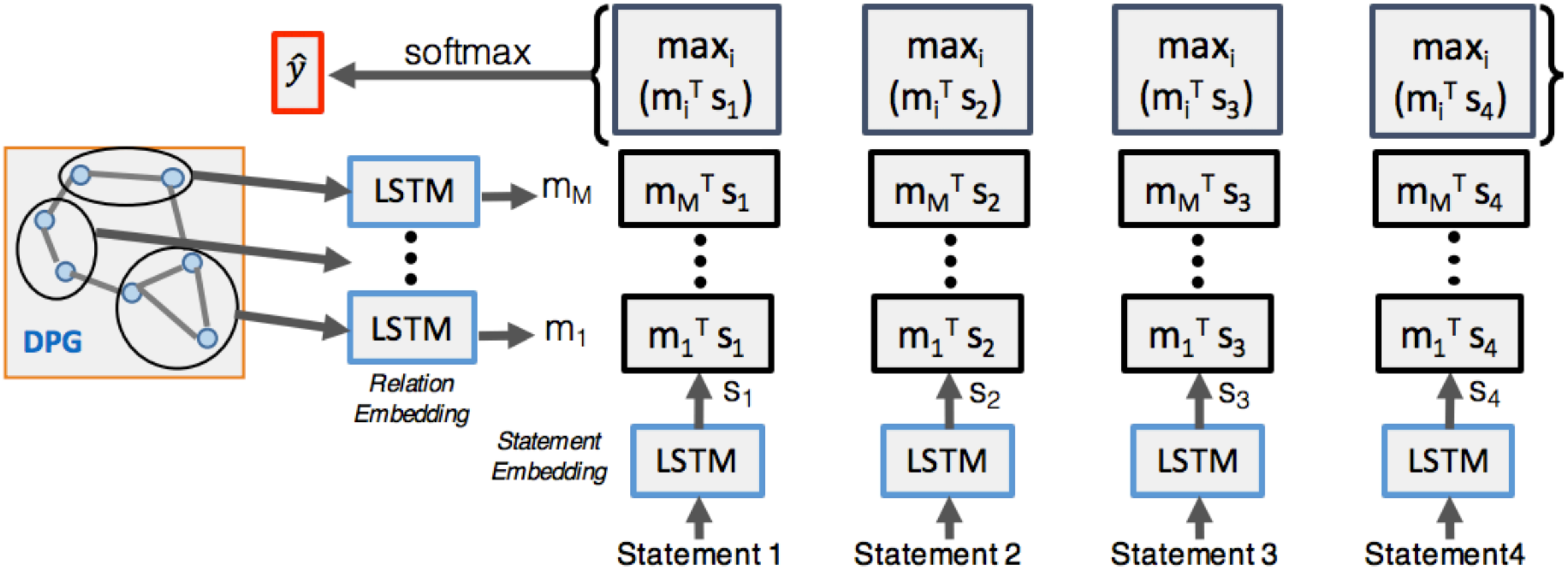


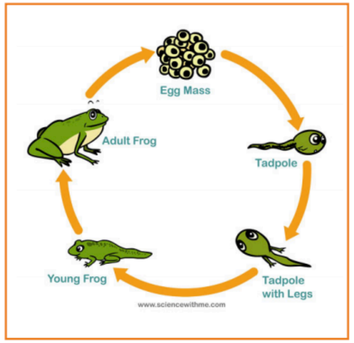
Diagram Parsing



Question Answering

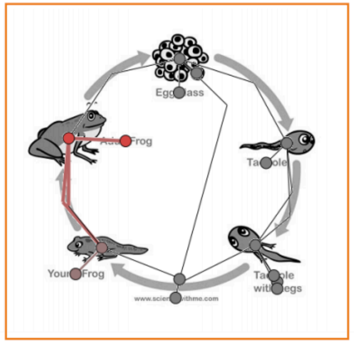


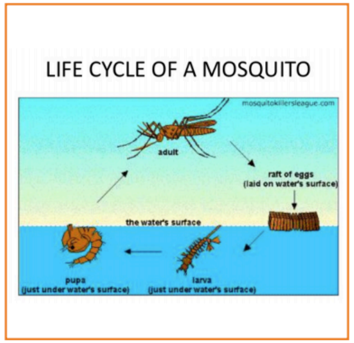
Attention visualization



The diagram depicts
The life cycle of

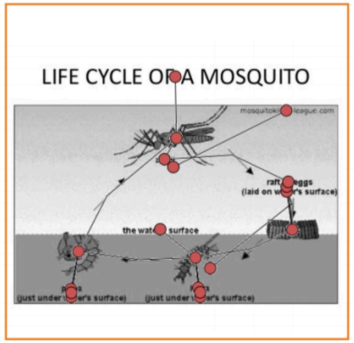
a) frog	**0.924**
b) bird	0.02
c) insecticide	0.054
d) insect	0.002

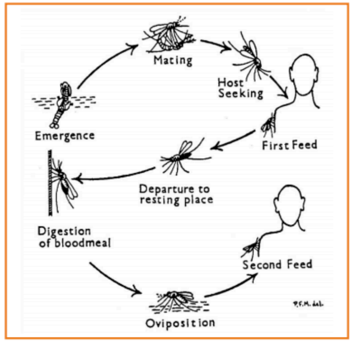




How many stages of
Growth does the diagram
Feature?

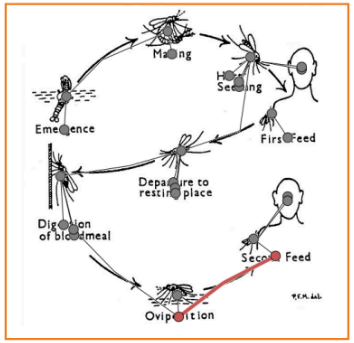
a) 4	**0.924**
b) 2	0.02
c) 3	0.054
d) 1	0.002





What comes before
Second feed?

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



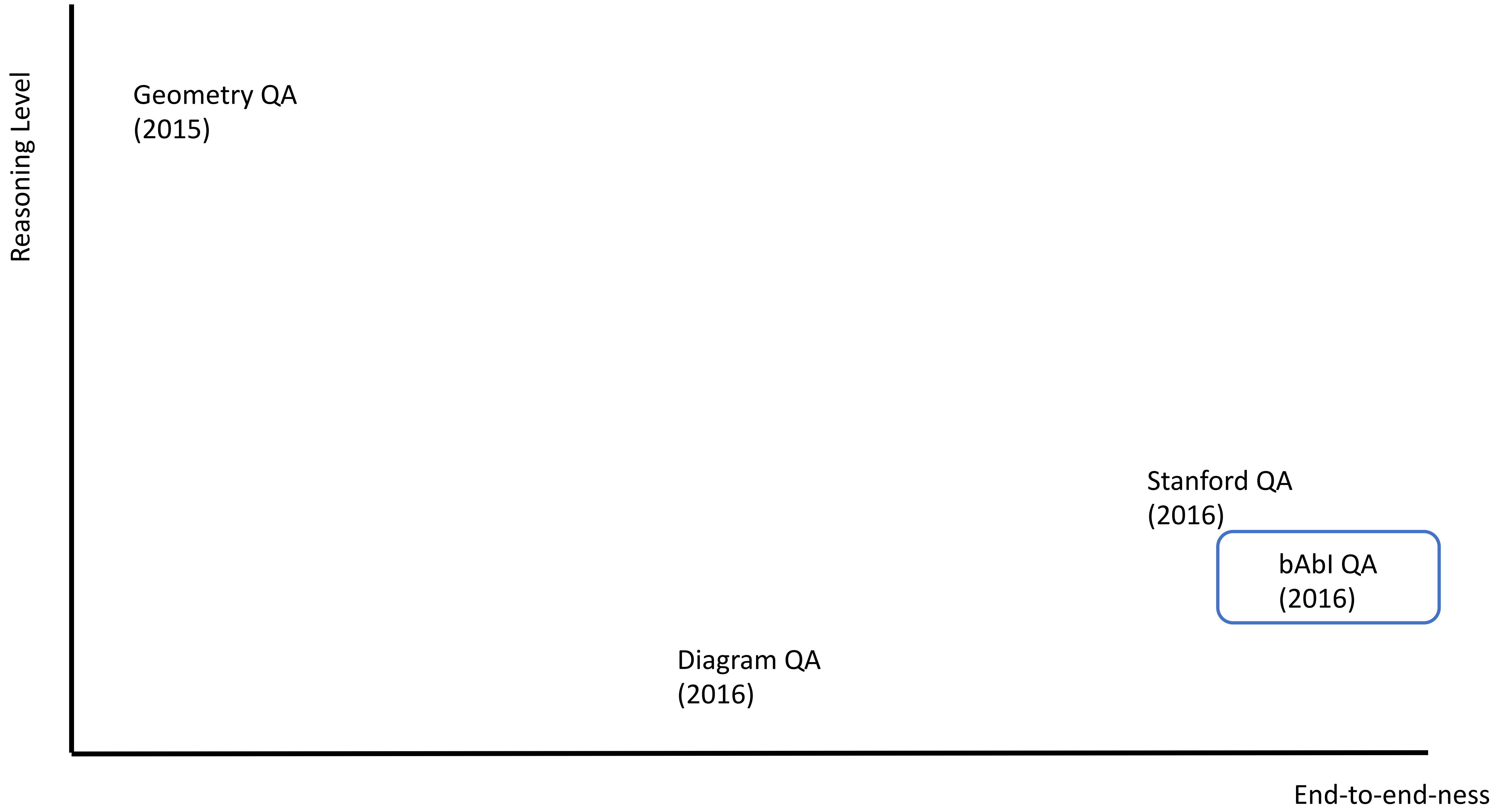
Results (ECCV 2016)

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

Limitations

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

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


Reasoning Level
Geometry QA (2015)
Stanford QA (2016)
bAbI QA (2016)
Diagram QA (2016)
End-to-end-ness

bAbI QA

- Weston et al., 2015 (Facebook)
- Synthetically generated reasoning story-question pairs
- 20 tasks, 1k questions in each task
- Each story can be as long as 200 sentences
- Requires reasoning over multiple sentences
- Should be trained end-to-end (no manual rules or external language resources)
- Passed a task if accuracy >= 95%

Tasks Examples

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 13: Compound Coreference
Daniel and Sandra journeyed to the office.
Then they went to the garden.
Sandra and John travelled to the kitchen.
After that they moved to the hallway.
Where is Daniel? A: garden

Task 7: Counting
Daniel picked up the football.
Daniel dropped the football.
Daniel got the milk.
Daniel took the apple.
How many objects is Daniel holding? A: two

Task 19: Path Finding
The kitchen is north of the hallway.
The bathroom is west of the bedroom.
The den is east of the hallway.
The office is south of the bedroom.
How do you go from den to kitchen? A: west, north
How do you go from office to bathroom? A: north, west

Previous work

- RNN: Tested as baseline by Weston et al. (2015)
  - Performs very poorly; hidden state is inherently unstable for long-term dependency

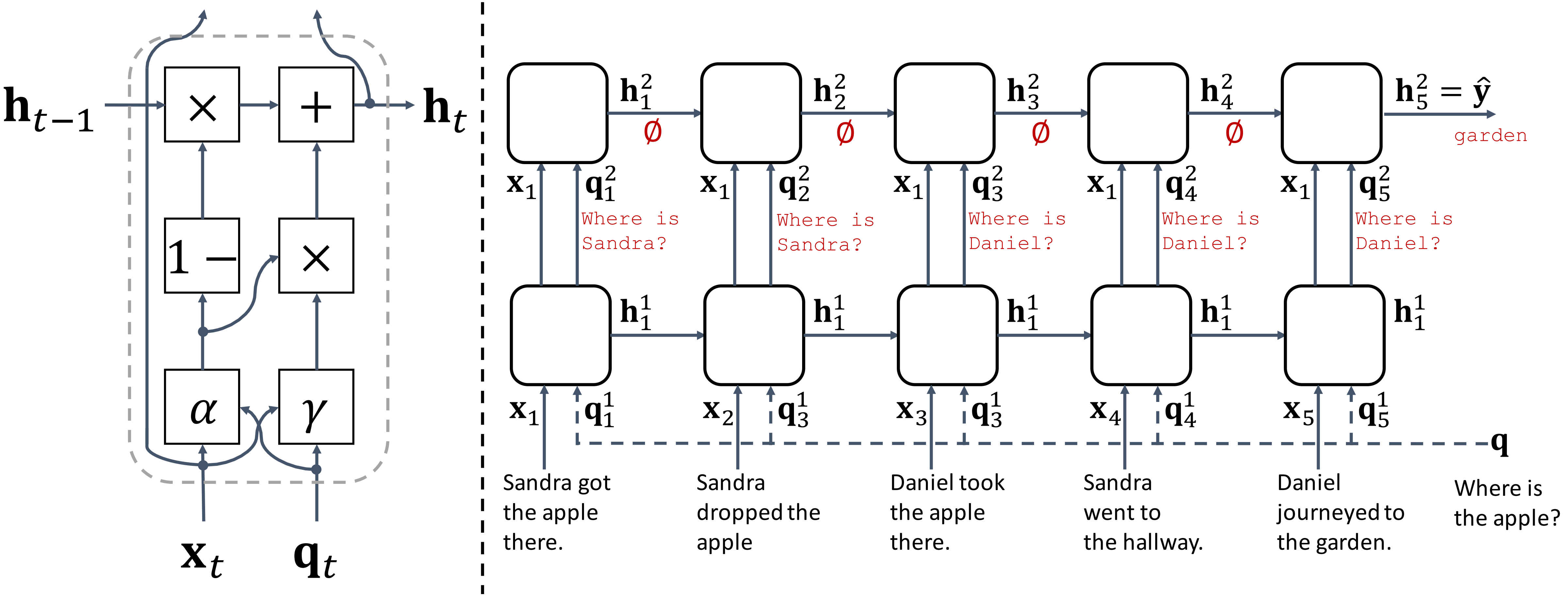
- Softmax attention mechanism (Sukhbaatar et al., 2015, Xiong et al., 2016)
  - Uses shared external memory with softmax attention mechanism
  - Attend on different facts over several layers
  - DMN: Combines RNN and attention mechanism
  - **Problem**:
    - vanilla softmax attention cannot distinguish between similar sentences at different time steps.
    - Cannot capture time locality information.

Query-regression networks

- Name comes from “Logic Regression” (not linear regression)
  - Transforming the original query to an easier-to-answer query, in vector space

- Pure RNN-based model
  - completely internal memory
  - Single unit recurring over time and layers (simple)
  - Although RNN, does not suffer from long-term dependency problem
  - Take full advantage of RNN’s capability to model sequential data
  - Can be considered as using “sigmoid attention”

Query-regression networks



Parallelization

$$\begin{pmatrix} \mathbf{h}_1^\top \\ \mathbf{h}_2^\top \\ \mathbf{h}_3^\top \\ \vdots \\ \mathbf{h}_T^\top \end{pmatrix} = \left[\exp\left\{ \begin{pmatrix} 0 & -\infty & -\infty & \dots & -\infty \\ b_2 & 0 & -\infty & \dots & -\infty \\ b_2 + b_3 & b_3 & 0 & \dots & -\infty \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{j=2}^T b_j & \sum_{j=3}^T b_j & \sum_{j=4}^T b_j & \dots & 0 \end{pmatrix} \right\} \right] \begin{pmatrix} z_1 \tilde{\mathbf{h}}_1^\top \\ z_2 \tilde{\mathbf{h}}_2^\top \\ z_3 \tilde{\mathbf{h}}_3^\top \\ \vdots \\ z_T \tilde{\mathbf{h}}_T^\top \end{pmatrix}$$

$$\mathbf{H} = [\mathbf{L} \circ \exp\left(\mathbf{L}\left[\mathbf{B} \circ \mathbf{L}'\right]\right)] \left[\mathbf{Z} \circ \tilde{\mathbf{H}}\right]$$

Results on bAbI QA 1k

	# of Tasks Passed	Average Accuracy (%)
LSTM (Weston et al., 2015)	0	48.7
End-to-end Memory Networks (Sukhbaatar et al., 2015)	10	84.8
QRN (2 layers)	13	**90.1**
QRN (3 layers)	**15**	88.7

Qualitative Results of QRN

	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 (hallway)			

	Layer 1			Layer 2
Task 3: Three Supporting Facts	z^1	$\overrightarrow{r}^1$	$\overleftarrow{r}^1$	z^2
Mary got the football there.	0.82	1.00	0.0	0.06
John went back to the bedroom.	0.01	0.00	0.72	0.57
Mary journeyed to the office.	0.01	0.04	0.06	0.88
Mary journeyed to the bathroom.	0.44	0.00	0.89	0.05
Mary dropped the football.	0.62	0.01	0.00	0.03
Where was the football before the bathroom?			office (office)	

	Layer 1			Layer 2
Task 7: Counting	z^1	$\overrightarrow{r}^1$	$\overleftarrow{r}^1$	z^2
Mary journeyed to the garden.	0.67	0.08	0.58	0.12
Mary journeyed to the office.	0.91	0.44	0.11	0.21
Sandra grabbed the apple there.	0.02	0.34	0.92	0.89
Sandra discarded the apple.	0.26	0.61	0.95	0.97
Daniel went to the bedroom.	0.70	0.44	0.99	0.03
How many objects is Sandra carrying?		none (none)		

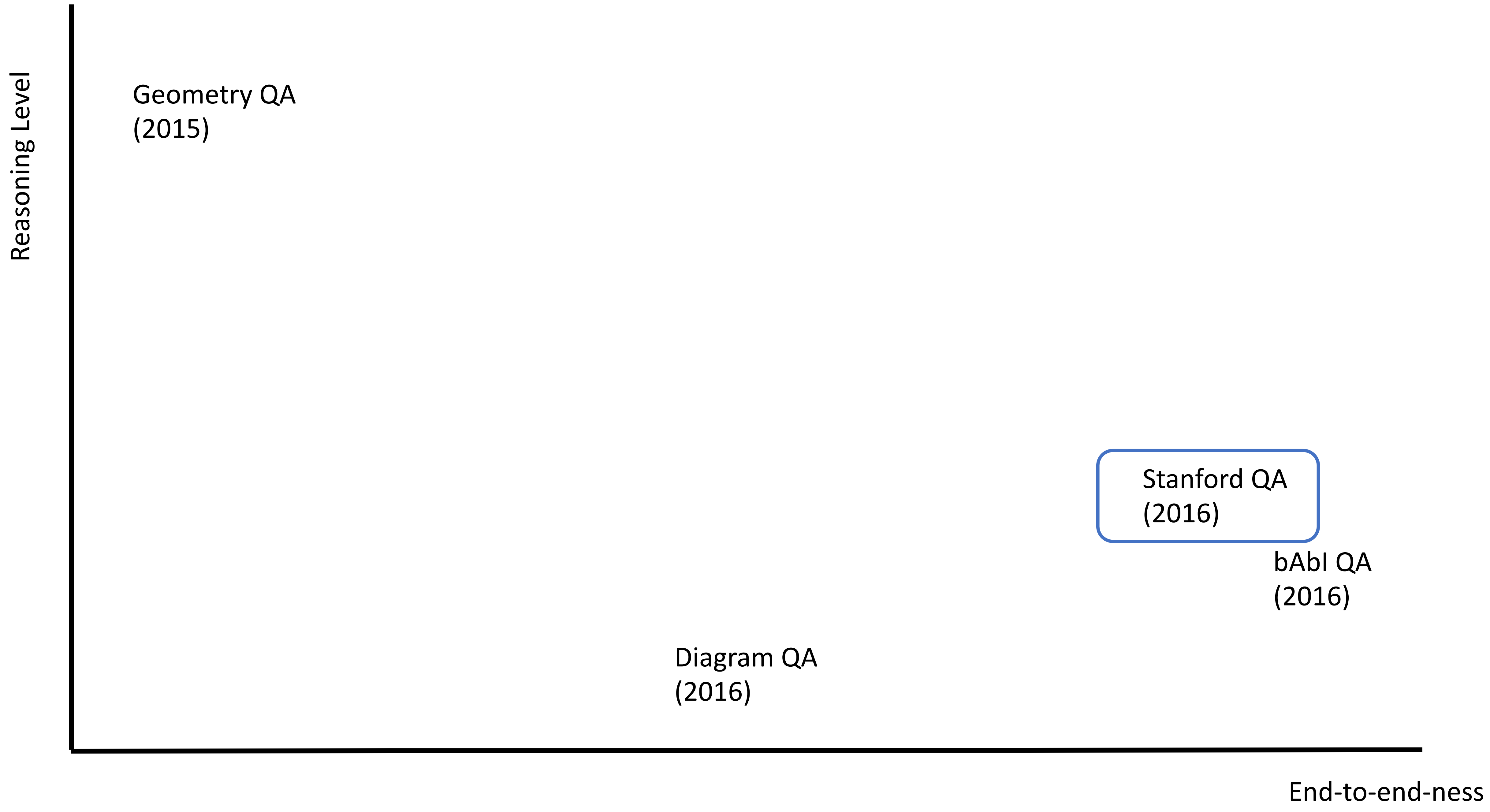
	Layer 1			Layer 2
Task 15: Deduction	z^1	$\overrightarrow{r}^1$	$\overleftarrow{r}^1$	z^2
Mice are afraid of wolves.	0.11	0.99	0.13	0.78
Gertrude is a mouse.	0.77	0.99	0.96	0.00
Cats are afraid of sheep.	0.01	0.99	0.07	0.03
Winona is a mouse.	0.14	0.85	0.77	0.05
Sheep are afraid of wolves.	0.02	0.98	0.27	0.05
What is Gertrude afraid of?		wolf (wolf)		

Results on bAbI QA 10k*

	# of Tasks Passed	Average Accuracy (%)
End-to-end Memory Networks (Sukhbaatar et al., 2015)	17	95.8
Dynamic Memory Networks Improved (Xiong et al., 2016)	19	97.2
QRN (2 layers)	18	96.8

Limitations

- Okay, the reasoning process is interesting …
- But “this is a fake dataset”! (by anonymous reviewers)


Reasoning Level
Geometry QA
(2015)
Stanford QA
(2016)
bAbI QA
(2016)
Diagram QA
(2016)
End-to-end-ness

SQuAD (Stanford QA)

Immune_system

The Stanford Question Answering Dataset

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

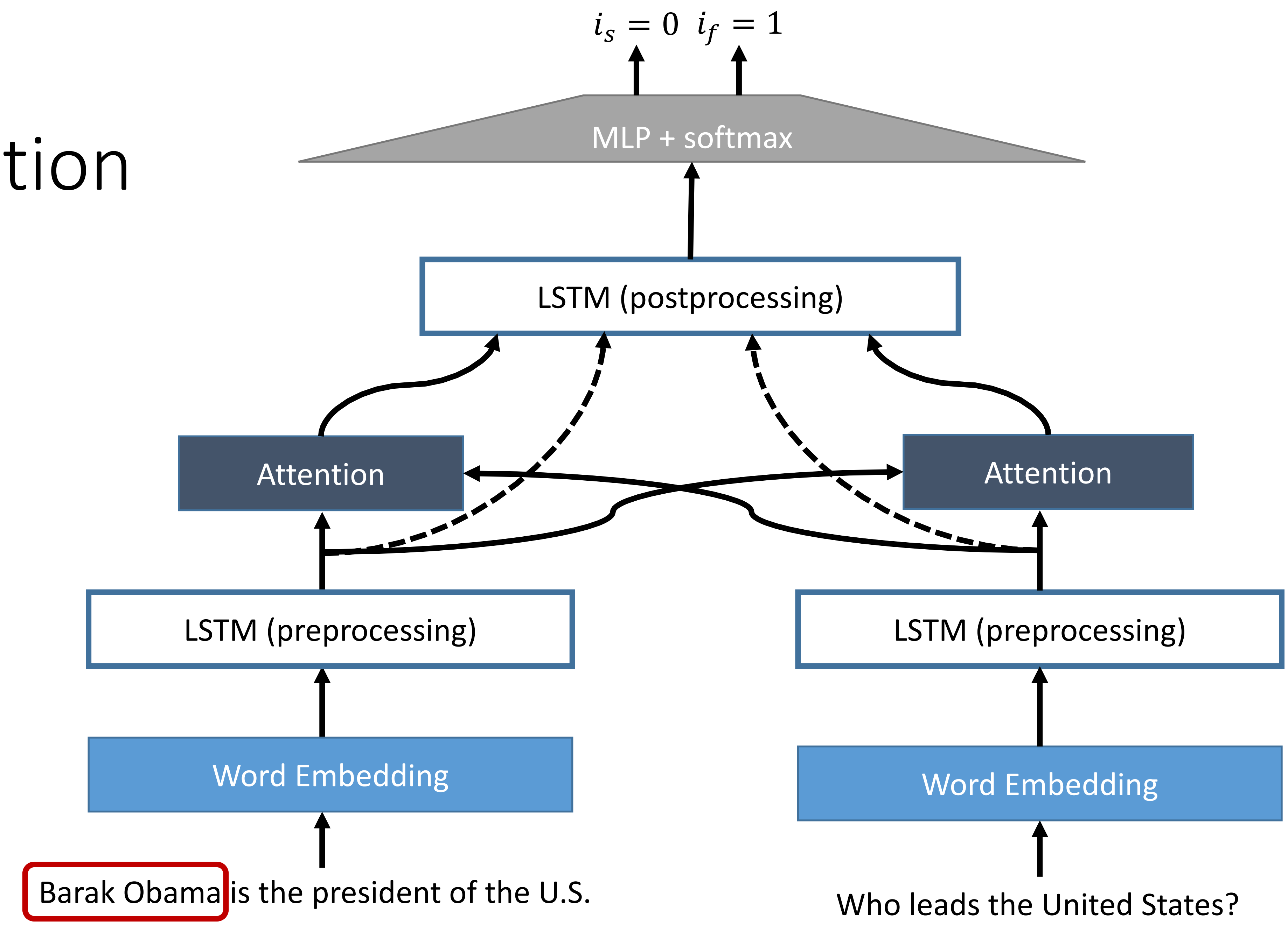
- Recently released: June 2016
- 100k+ paragraph-question-answer triples
- Paragraphs from most popular articles in Wikipedia
- Answer is the subphrase of the paragraph

Stanford QA vs Other "Big" QA Datasets

- CNN / Daily Mail (Hermann et al., 2015)
  - Google DeepMind
  - Document-Summary pairs from web
  - Cloze test on summary (fill in the blank)
- Children's Book Test (Hill et al., 2015)
  - Facebook AI Research
  - Project Gutenberg: Children's books
  - Cloze test on 21st sentence

- **Take away**: Cloze test, and crawled data
- Stanford QA is direct question, and carefully controlled (turked)

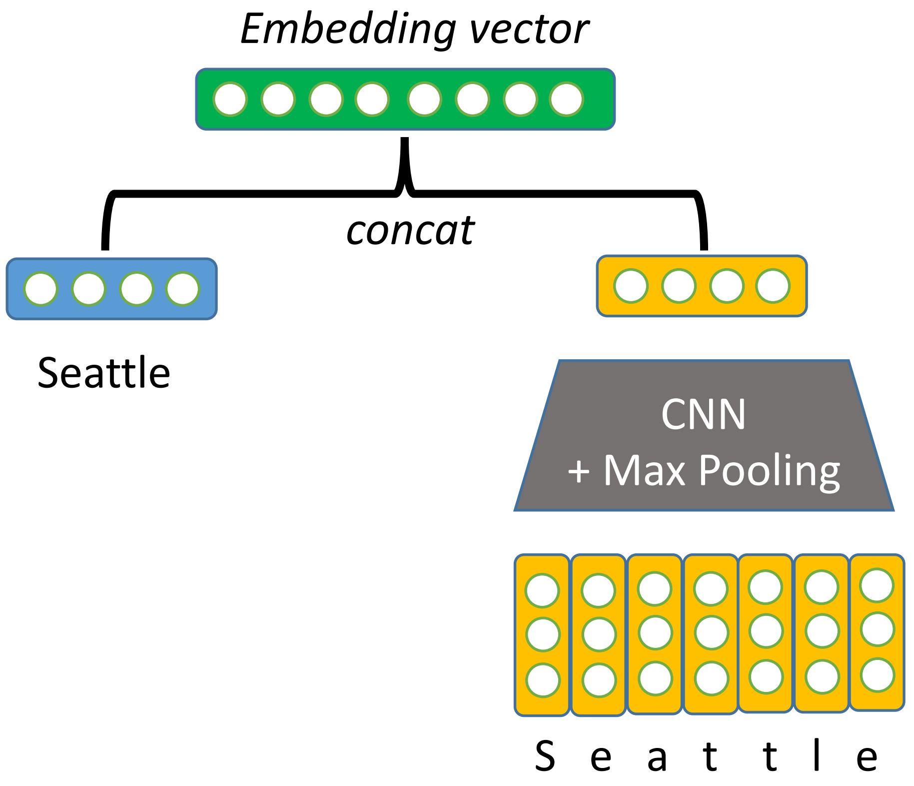
Model: Co-Attention



Embedding Module

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

- Char embedding as proposed by Yoon (2015)



Attention Mechanism: Motivation

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.

Q: Which city is gloomy in winter?

Attention Mechanism

- **Theoretically**, RNN can propagate information over a long distance through its recurrent state
- **Practically**, this is very difficult
  - Inherently unstable state, even using LSTM (Weston et al., 2014)
  - State size is fixed (Bahdanau et al., 2014)
- Attention provides *shortcut access* to distant information
- **Co-Attention**: question attends on context, and context attends on question. Similar in spirit to, but fundamentally different from, Lu et al. (2016).

Results: Metric

- Each question is answered by 2-5 different people (by indicating the answer phrase in the paragraph)
- **Exact Match**: the answer exactly matches one of the answers
- **F1 Score**: geometric average of precision and recall

- "The actors were paid $1.5 million on average."

- *Q: Who were paid more than $1 million on average?*

Results on Dev (Sept. 20, 2016)

	Exact Match (%)	F1 (%)
Baseline (June 2016)	39.0	51.0
Attention and Chunking (IBM)	48.0	64.5
Match LSTM v1 (Singapore)	54.8	68.0
Match LSTM v2 (Singapore)	59.4	70.0
Neural Chunker (IBM)	61.8	70.7
Co-Attention (Ours)	**62.2**	**72.6**

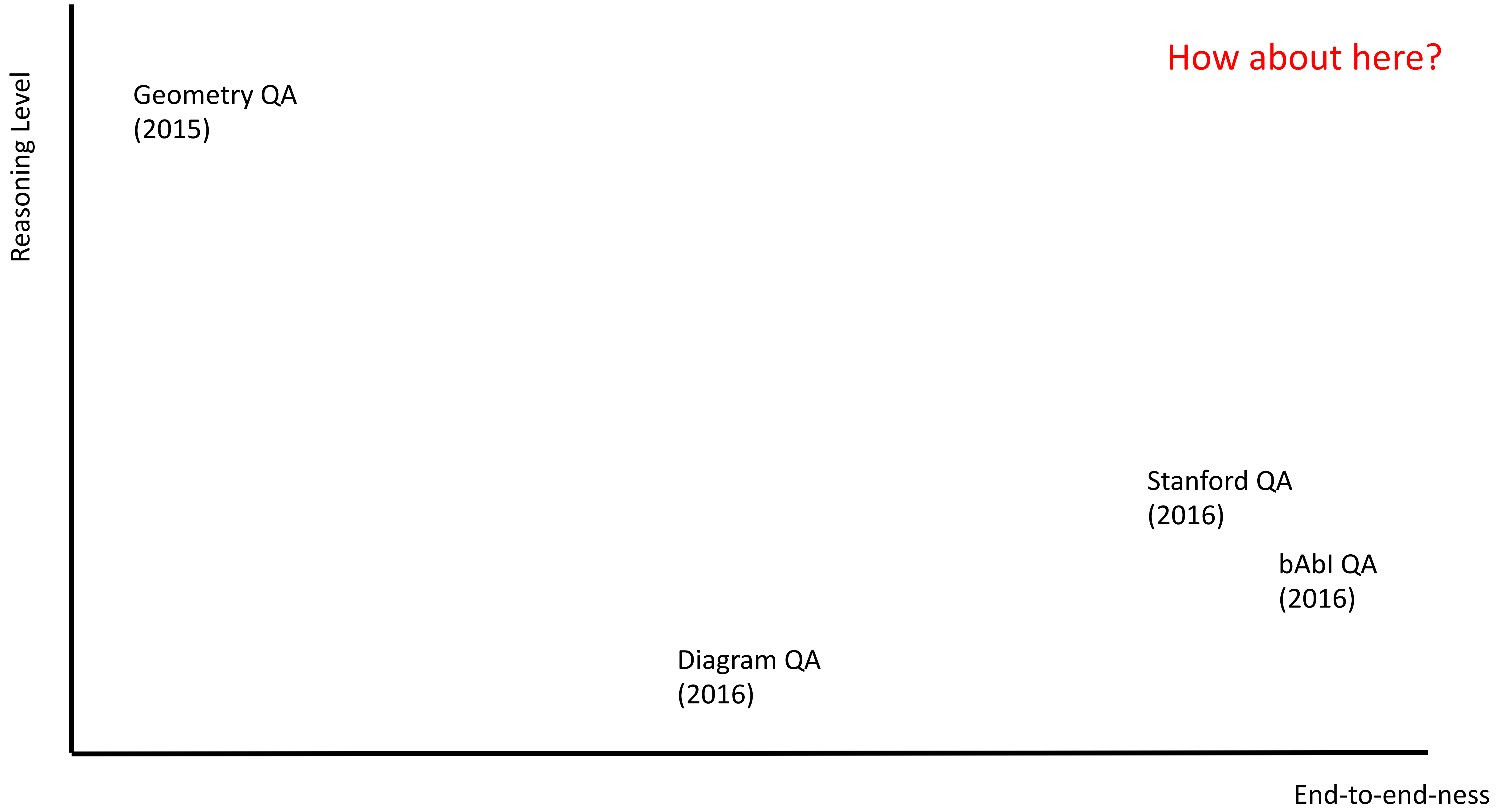
Attention Visualization

56e0bb9f7aa994140058e6cf	Other than his scientific achievements what was Tesla famous for ?	• showmanship • showmanship • showmanship	showmanship	0.637

Tesla	was	renowned	for	his	achievements	and	showmanship	,	eventually	earning	him	a	reputation	
-	-	-	-	-	-	-	-	-	-	-	-	-	-	
His	patents	earned	him	a	considerable	amount	of	money	,	much	of	which	was	
-	-	-	-	-	-	-	-	-	-	-	-	-	-	
Tesla	died	on	7	January	1943	.								
-	-	-	-	-	-	-								
His	work	fell	into	relative	obscurity	after	his	death	,	but	in	1960	the	
-	-	-	-	-	-	-	-	-	-	-	-	-	-	
There	has	been	a	resurgence	in	popular	interest	in	Tesla	since	the	1990s	.	
-	-	-	-	-	-	-	-	-	-	-	-	-	-	

56e0c0c7231d4119001ac376	What were Tesla 's sisters ' names ?	• Milka, Angelina and Marica • Milka, Angelina and Marica • Milka, Angelina and Marica	Milka , Angelina and Marica	0.555

Tesla	was	the	fourth	of	five	children	.									
-	-	-	-	-	-	-	-									
He	had	an	older	brother	named	Dane	and	three	sisters	,	Milka	,	Angelina	and	Marica	.
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Dane	was	killed	in	a	horse-riding	accident	when	Nikola	was	five	.					
-	-	-	-	-	-	-	-	-	-	-	-					
In	1861	,	Tesla	attended	the	"	Lower	"	or	"	Primary	"	School	in	Smiljan	where
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
In	1862	,	the	Tesla	family	moved	to	Gospić	,	Austrian	Empire	,	where	Tesla	's	father
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Nikola	completed	"	Lower	"	or	"	Primary	"	School	,	followed	by	the	"	Lower	Real
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-


How about here?
Reasoning Level
Geometry QA
(2015)
Stanford QA
(2016)
bAbI QA
(2016)
Diagram QA
(2016)
End-to-end-ness

Important questions

- Is fully end-to-end reasoning system feasible with reasonable amount of data? → Probably no
- How to balance between:
  - data size
  - model priors (manually defined rules, annotations, etc.)
- How to naturally incorporate model priors (which might be structured data) into the model?

Thank you!

- seominjoon@gmail.com
- http://seominjoon.github.io