Learning to reason by reading text and answering questions

Minjoon Seo
Natural Language Processing Group
University of Washington
May 30, 2017
@ Qualcomm
What is reasoning?
One-to-one model

“Hello” \rightarrow \text{A lot of parameters (to learn)} \rightarrow “Bonjour”
Examples

• Most neural machine translation systems (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
One-to-one Model

**Problem:** parametric model has finite capacity.

“*You can’t even fit a sentence into a single vector*” - Dan Roth
Model with explicit knowledge

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hello</td>
<td>Bonjour</td>
</tr>
<tr>
<td>Thank you</td>
<td>Merci</td>
</tr>
</tbody>
</table>

“Hello” —> Knowledge Base —> “Bonjour”
Examples

- Phrase-based Statistical Machine Translation (Chiang, 2005)
- 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 probabilistic models
- Deep learning models with external memory (e.g. Memory Networks)
Model with explicit knowledge

What does a frog eat? 🔴

Context (Knowledge Base)

<table>
<thead>
<tr>
<th>Eats</th>
<th>IsA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Amphibian, insect)</td>
<td>(Frog, amphibian)</td>
</tr>
<tr>
<td>(insect, flower)</td>
<td>(Fly, insect)</td>
</tr>
</tbody>
</table>

Something is missing ...
Explicit knowledge and reasoning capability

What does a frog eat?

Fly

<table>
<thead>
<tr>
<th>Eats</th>
<th>IsA</th>
<th>First Order Logic</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Amphibian, insect)</td>
<td>(Frog, amphibian)</td>
<td>IsA(A, B) ^ IsA(C, D) ^ Eats(B, D) → Eats(A, C)</td>
</tr>
<tr>
<td>(insect, flower)</td>
<td>(Fly, insect)</td>
<td></td>
</tr>
</tbody>
</table>

Context (Knowledge Base)
Examples

• Semantic parsing
  • SAIL (Chen & Mooney, 2011; Artzi & Zettlemoyer, 2013)

• 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)
Explicit knowledge and reasoning capability

What does a frog eat?

Fly

Who makes this? Tell me it’s not me ...

<table>
<thead>
<tr>
<th>Eats</th>
<th>IsA</th>
<th>First Order Logic</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Amphibian, insect)</td>
<td>(Frog, amphibian)</td>
<td>IsA(A, B) ^ IsA(C, D) ^ Eats(B, D) ⇒ Eats(A, C)</td>
</tr>
<tr>
<td>(insect, flower)</td>
<td>(Fly, insect)</td>
<td></td>
</tr>
</tbody>
</table>

Context (Knowledge Base)
What does a frog eat?

Fly

Frog is an example of amphibian. Flies are one of the most common insects around us. Insects are good sources of protein for amphibians. ...
Let’s define:
‘reasoning’ = “using existing knowledge (or context) to produce new knowledge”
How to learn to reason?

• Question-driven
• Read text (unstructured data)

• That is, learning to reason by reading text and answering questions
Three aspects of “reasoning system”

• **Natural language understanding**
  • How to retrieve relevant knowledge (formulas)?
  • Natural language has diverse surface forms (lexically, syntactically)

• **Reasoning**
  • Deriving new knowledge from the retrieved knowledge

• **End-to-end training**
  • Minimizing human efforts
  • Using only unstructured data
What we want...

Reasoning capability

NLU capability

End-to-end
Geometry QA

Reasoning capability

NLU capability

End-to-end
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  
\textcolor{green}{d) 8}  e) 10
What is the length of BD?

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

Difficult to directly map text to a long logical form!
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

Selected subset

Logical form

Over-generated literals

Text scores

Diagram scores

<table>
<thead>
<tr>
<th>Literal</th>
<th>Score</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>IsTriangle(ABC)</td>
<td>0.96</td>
<td>1.00</td>
</tr>
<tr>
<td>Parallel(AC, DE)</td>
<td>0.91</td>
<td>0.99</td>
</tr>
<tr>
<td>Parallel(AC, DB)</td>
<td>0.74</td>
<td>0.02</td>
</tr>
<tr>
<td>Equals(LengthOf(DB), 4)</td>
<td>0.97</td>
<td>n/a</td>
</tr>
<tr>
<td>Equals(LengthOf(AD), 8)</td>
<td>0.94</td>
<td>n/a</td>
</tr>
<tr>
<td>Equals(LengthOf(DE), 5)</td>
<td>0.94</td>
<td>n/a</td>
</tr>
<tr>
<td>Equals(4, LengthOf(AD))</td>
<td>0.31</td>
<td>n/a</td>
</tr>
</tbody>
</table>

…

…”

Text / input

Logical form

IsTriangle(ABC) \text{∧} Parallel(AC, DE) \text{∧} \quad \text{Find}(\text{LengthOf}(AC))

Equals(LengthOf(DB), 4) \text{∧} \quad \text{Equals}(\text{LengthOf}(AD), 8) \text{∧} \quad \text{Equals}(\text{LengthOf}(DE), 5) \text{∧}
Numerical solver

- Translate literals to numeric equations

<table>
<thead>
<tr>
<th>Literal</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equals(LengthOf(AB),d)</td>
<td>$(A_x-B_x)^2 + (A_y-B_y)^2 - d^2 = 0$</td>
</tr>
<tr>
<td>Parallel(AB, CD)</td>
<td>$(A_x-B_x)(C_y-D_y) - (A_y-B_y)(C_x-D_x) = 0$</td>
</tr>
<tr>
<td>PointLiesOnLine(B, AC)</td>
<td>$(A_x-B_x)(B_y-C_y) - (A_y-B_y)(B_x-C_x) = 0$</td>
</tr>
<tr>
<td>Perpendicular(AB,CD)</td>
<td>$(A_x-B_x)(C_x-D_x) + (A_y-B_y)(C_y-D_y) = 0$</td>
</tr>
</tbody>
</table>

- 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
In the figure to the left, triangle ABC is inscribed in the circle with center O and diameter AC. If AB = AO, what is the degree measure of angle ABO?

(A) 15°  
(B) 30°  
(C) 45°  
(D) 60°  
(E) 90°
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
Diagram QA

Reasoning capability

NLU capability

End-to-end
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?

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)

<table>
<thead>
<tr>
<th>Method</th>
<th>Training data</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random (expected)</td>
<td>-</td>
<td>25.00</td>
</tr>
<tr>
<td>LSTM + CNN</td>
<td>VQA</td>
<td>29.06</td>
</tr>
<tr>
<td>LSTM + CNN</td>
<td>AI2D</td>
<td>32.90</td>
</tr>
<tr>
<td>Ours</td>
<td>AI2D</td>
<td><strong>38.47</strong></td>
</tr>
</tbody>
</table>
Limitations

• You can’t really call this reasoning...
  • Rather matching 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”
Cell Structures

Introduction
In some ways, a cell resembles a plastic bag full of Jell-O. Its basic structure is a cell membrane with a cytoplasmic interior. The cytoplasm of a eukaryotic cell is like Jell-O containing mixed fruit, but 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.

Cytoplasm
Cytoplasm is a single-celled organism that lacks a nucleus. It is the material outside the cell membrane containing all the cell's organelles. The cytoplasm is composed of water, proteins, and other organelles. In some eukaryotic cells, the cytoplasm is the entire cell.

Organelles
- Cell Wall
- Rigid layer that surrounds the cell membrane and supports and protects the cell.
- Cytoplasm
- Cell containing all the cell's organelles and other parts.
- Central Vacuole
- Large storage sac found in the cells of plants.

Cell Membrane Structure

Lesson Summary
- The cell membrane consists of two layers of phospholipids.
- The cell membrane consists of various cellular structures.
- Eukaryotic cells contain a nucleus and other organelles.

Instructional Diagrams

Questions
What is the outer surrounding part of the Nucleus?

a. Nuclear Membrane
b. Cell Body
c. Cell Membrane
d. Nucleus

Which statement about the cell membrane is false?

a. It regulates the cytoplasm.
b. It protects and supports the cell.
c. It keeps all external substances out of the cell.
d. None of the above.
Reasoning capability

Machine Comprehension

NLU capability

End-to-end
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)

Barak Obama is the president of the U.S.

Who leads the United States?
(Bidirectional) Attention Flow

- **Output Layer**
- **Modeling Layer**
- **Attention Flow Layer**
- **Phrase Embed Layer**
- **Word Embed Layer**
- **Character Embed Layer**

```
(x_1, x_2, x_3, x_T) \rightarrow \text{Context} \rightarrow \text{Query2Context and Context2Query Attention} \rightarrow (u_1, u_J) 
```

```
(h_1, h_2, h_T) \rightarrow \text{Dense + Softmax} \rightarrow \text{LSTM + Softmax} 
```

```
(q_1, q_J) \rightarrow \text{Softmax} \rightarrow \text{Max} 
```

```
Start \rightarrow (m_1, m_2, m_T) \rightarrow \text{LSTM} \rightarrow (g_1, g_2, g_T) 
```

```
End 
```

```
(g_1, g_2, g_T) \rightarrow \text{LSTM + Softmax} \rightarrow \text{Dense + Softmax} 
```

```
(h_1, h_2, h_T) \rightarrow \text{Softmax} \rightarrow \text{Max} 
```

```
(m_1, m_2, m_T) \rightarrow \text{LSTM} \rightarrow (g_1, g_2, g_T) 
```

```
(u_1, u_J) \rightarrow \text{Softmax} \rightarrow \text{Max} 
```

```
(h_1, h_2, h_T) \rightarrow \text{Softmax} \rightarrow \text{Max} 
```

```
(g_1, g_2, g_T) \rightarrow \text{LSTM + Softmax} \rightarrow \text{Dense + Softmax} 
```

```
(h_1, h_2, h_T) \rightarrow \text{Softmax} \rightarrow \text{Max} 
```

```
(u_1, u_J) \rightarrow \text{Softmax} \rightarrow \text{Max} 
```

```
(m_1, m_2, m_T) \rightarrow \text{LSTM} \rightarrow (g_1, g_2, g_T) 
```

```
(g_1, g_2, g_T) \rightarrow \text{LSTM + Softmax} \rightarrow \text{Dense + Softmax} 
```

```
(h_1, h_2, h_T) \rightarrow \text{Softmax} \rightarrow \text{Max} 
```

```
(u_1, u_J) \rightarrow \text{Softmax} \rightarrow \text{Max} 
```
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

- **Output Layer**
- **Modeling Layer**
- **Attention Flow Layer**
- **Phrase Embed Layer**
- **Word Embed Layer**
- **Character Embed 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

Output Layer

Modeling Layer

Attention Flow Layer

Phrase Embed Layer

Word Embed Layer

Character Embed Layer

Start

End

Dense + Softmax

LSTM + Softmax

Query2Context and Context2Query

Attention

Context

Query

LSTM

LSTM

LSTM

LSTM

Softmax

Max

Softmax

x_1

x_2

x_3

x_T

q_1

q_J

h_1

h_2

h_T

u_1

u_J

g_1

g_2

g_T

m_1

m_2

m_T

Attention Flow

Phrase Embed Layer

Word Embed Layer

Character Embed Layer

Output 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.
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

\[
L(\theta) = -\frac{1}{N} \sum_{i}^{N} \log(p_{y_i}^1) + \log(p_{y_i}^2)
\]

\(y_i^1\) True start index of example i

\(y_i^2\) True end index of example i

\(p^1\) Probability distribution of start index

\(p^2\) 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
Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016)

- 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)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>Test EM</th>
<th>Test F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BiDAF (ensemble)</td>
<td>73.3</td>
<td>81.1</td>
</tr>
<tr>
<td></td>
<td>Allen Institute for AI &amp; University of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Washington</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Seo et al. ‘16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Dynamic Coattention Networks (ensemble)</td>
<td>71.6</td>
<td>80.4</td>
</tr>
<tr>
<td></td>
<td>Salesforce Research</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Xiong &amp; Zhong et al. ‘16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>r-net (ensemble)</td>
<td>72.1</td>
<td>79.7</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>r-net (single model)</td>
<td>68.4</td>
<td>77.5</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>BiDAF (single model)</td>
<td>68.0</td>
<td>77.3</td>
</tr>
<tr>
<td></td>
<td>Allen Institute for AI &amp; University of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Washington</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Seo et al. ‘16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Multi-Perspective Matching (ensemble)</td>
<td>68.2</td>
<td>77.2</td>
</tr>
<tr>
<td></td>
<td>IBM Research</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rank</td>
<td>Model</td>
<td>EM</td>
<td>F1</td>
</tr>
<tr>
<td>------</td>
<td>--------------------------------------------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>1</td>
<td>r-net (ensemble)</td>
<td>76.922</td>
<td>84.006</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="http://aka.ms/rnet">http://aka.ms/rnet</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>MEMEN (ensemble)</td>
<td>75.37</td>
<td>82.658</td>
</tr>
<tr>
<td></td>
<td>Eigen Technology &amp; Zhejiang University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ReasoNet (ensemble)</td>
<td>75.034</td>
<td>82.552</td>
</tr>
<tr>
<td></td>
<td>MSR Redmond</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>r-net (single model)</td>
<td>74.614</td>
<td>82.458</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="http://aka.ms/rnet">http://aka.ms/rnet</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Mnemonic Reader (ensemble)</td>
<td>73.754</td>
<td>81.863</td>
</tr>
<tr>
<td></td>
<td>NUDT &amp; Fudan University</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="https://arxiv.org/abs/1705.02798">https://arxiv.org/abs/1705.02798</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>SEDT+BiDAF (ensemble)</td>
<td>73.723</td>
<td>81.53</td>
</tr>
<tr>
<td></td>
<td>CMU</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="https://arxiv.org/abs/1703.00572">https://arxiv.org/abs/1703.00572</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>BiDAF (ensemble)</td>
<td>73.744</td>
<td>81.525</td>
</tr>
<tr>
<td></td>
<td>Allen Institute for AI &amp; University of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Washington</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="https://arxiv.org/abs/1611.01603">https://arxiv.org/abs/1611.01603</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ablations on dev data

- No Char Embedding
- No Word Embedding
- No C2Q Attention
- No Q2C Attention
- Dynamic Attention
- Full Model

EM | F1
---|---
No Char Embedding | 70 | 80
No Word Embedding | 55 | 65
No C2Q Attention | 60 | 70
No Q2C Attention | 65 | 70
Dynamic Attention | 70 | 75
Full Model | 75 | 80
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

Opening in May 1852 at the state of these may be more effect and may result in
from 28 January to 25 August may not aid but by September had been
debut on May 5 ,
How does it compare with feature-based models?

Questions answered correctly by our BiDAF model and the more traditional baseline model.

- BiDAF: 3734 questions
- Baseline: 3585 questions
- BiDAF (Ours): 509 questions

The chart shows the percentage of questions with correct answers for each category.
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
<table>
<thead>
<tr>
<th>Model</th>
<th>CNN val</th>
<th>CNN test</th>
<th>DailyMail val</th>
<th>DailyMail test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attentive Reader (Hermann et al., 2015)</td>
<td>61.6</td>
<td>63.0</td>
<td>70.5</td>
<td>69.0</td>
</tr>
<tr>
<td>MemNN (Hill et al., 2016)</td>
<td>63.4</td>
<td>6.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AS Reader (Kadlec et al., 2016)</td>
<td>68.6</td>
<td>69.5</td>
<td>75.0</td>
<td>73.9</td>
</tr>
<tr>
<td>Stanford AR (Chen et al., 2016)</td>
<td>68.6</td>
<td>69.5</td>
<td>75.0</td>
<td>73.9</td>
</tr>
<tr>
<td>DER Network (Kobayashi et al., 2016)</td>
<td>71.3</td>
<td>72.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Iterative Attention (Sordoni et al., 2016)</td>
<td>72.6</td>
<td>73.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EpiReader (Trischler et al., 2016)</td>
<td>73.4</td>
<td>74.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GAIReader (Dhingra et al., 2016)</td>
<td>73.0</td>
<td>73.8</td>
<td>76.7</td>
<td>75.7</td>
</tr>
<tr>
<td>AoA Reader (Cui et al., 2016)</td>
<td>73.1</td>
<td>74.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ReasoNet (Shen et al., 2016)</td>
<td>72.9</td>
<td>74.7</td>
<td>77.6</td>
<td>76.6</td>
</tr>
<tr>
<td>BiDAF (Ours)</td>
<td>76.3</td>
<td>76.9</td>
<td>80.3</td>
<td>79.6</td>
</tr>
<tr>
<td>MemNN* (Hill et al., 2016)</td>
<td>66.2</td>
<td>69.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ASReader* (Kadlec et al., 2016)</td>
<td>73.9</td>
<td>75.4</td>
<td>78.7</td>
<td>77.7</td>
</tr>
<tr>
<td>Iterative Attention* (Sordoni et al., 2016)</td>
<td>74.5</td>
<td>75.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GA Reader* (Dhingra et al., 2016)</td>
<td>76.4</td>
<td>77.4</td>
<td>79.1</td>
<td>78.1</td>
</tr>
</tbody>
</table>
## Transfer Learning (ACL 2017)

<table>
<thead>
<tr>
<th>Pretrained dataset</th>
<th>Fine-tuned</th>
<th>WikiQA</th>
<th>SemEval-2016</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAP</td>
<td>MRR</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>62.96</td>
<td>64.47</td>
</tr>
<tr>
<td>SQuAD-T</td>
<td>No</td>
<td>75.22</td>
<td>76.40</td>
</tr>
<tr>
<td>SQuAD</td>
<td>No</td>
<td>75.19</td>
<td>76.31</td>
</tr>
<tr>
<td>SQuAD-T</td>
<td>Yes</td>
<td>76.44</td>
<td>77.85</td>
</tr>
<tr>
<td>SQuAD</td>
<td>Yes</td>
<td>79.90</td>
<td>82.01</td>
</tr>
<tr>
<td>SQuAD*</td>
<td>Yes</td>
<td><strong>83.20</strong></td>
<td><strong>84.58</strong></td>
</tr>
<tr>
<td>Rank 1</td>
<td></td>
<td>74.33</td>
<td>75.45</td>
</tr>
<tr>
<td>Rank 2</td>
<td></td>
<td>74.17</td>
<td>75.88</td>
</tr>
<tr>
<td>Rank 3</td>
<td></td>
<td>70.69</td>
<td>72.65</td>
</tr>
</tbody>
</table>
Some limitations of SQuAD

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

Reasoning capability

NLU capability

End-to-end
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**
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:

Where is the apple?
Where is Sandra?
Where is Sandra?
Where is Daniel?
Where is Daniel?
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), \textit{in vector space}
QRN Cell

\[ z_t = \alpha(x_t, q_t) \]
\[ \tilde{h}_t = \rho(x_t, q_t) \]
\[ h_t = z_t \tilde{h}_t + (1 - z_t) h_{t-1} \]
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

\[ z_t = \alpha(x_t, q_t) \]

\[ \hat{h}_t = \rho(x_t, q_t) \]

\[ h_t = z_t \hat{h}_t + (1 - z_t)h_{t-1} \]

\[ h_t = \sum_{i=1}^{t} \left[ \prod_{j=i+1}^{t} 1 - z_j \right] z_i \hat{h}_i \]

computed from inputs only, so can be trivially parallelized

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

Where is Sandra? Where is Sandra? Where is Daniel? Where is Daniel? Where is Daniel?

Sandra got the apple there. Sandra dropped the apple. Daniel took the apple there. Sandra went to the hallway. Daniel journeyed to the garden.

- q
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)

Avg Error (%)

- LSTM
- DMN+
- MemN2N
- GMemN2N
- QRN (Ours)
bAbI QA Results (10k)

![Average Error (%) Bar Chart]

- MemN2N
- DNC
- GMemN2N
- DMN+
- QRN (Ours)
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)

![Bar chart showing average error (%) for MemN2N, GMemN2N, and QRN (Ours).]
DSTC2* Dialog Results

 Avg Error (%)

MemN2N  GMemN2N  QRN (Ours)

Avg Error (%)

0  10  20  30  40  50  60  70
### bAbI QA Visualization

<table>
<thead>
<tr>
<th>Task 2: Two Supporting Facts</th>
<th>Layer 1</th>
<th>Layer 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sandra picked up the apple there.</td>
<td>0.95</td>
<td>0.89</td>
</tr>
<tr>
<td>Sandra dropped the apple.</td>
<td>0.83</td>
<td>0.05</td>
</tr>
<tr>
<td>Daniel grabbed the apple there.</td>
<td>0.88</td>
<td>0.93</td>
</tr>
<tr>
<td>Sandra travelled to the bathroom.</td>
<td>0.01</td>
<td>0.18</td>
</tr>
<tr>
<td>Daniel went to the hallway.</td>
<td>0.01</td>
<td>0.24</td>
</tr>
<tr>
<td>Where is the apple?</td>
<td>hallway</td>
<td></td>
</tr>
</tbody>
</table>

$z^l = \text{Local attention (update gate) at layer } l$
DSTC2 (Dialog) Visualization

\[ z^l = \text{Local attention (update gate) at layer } l \]
So...
Is this possible?
Reasoning capability

Or this?

NLU capability

End-to-end
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...
“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