Welcome
2017 Faculty Summit Attendees

Faculty Summit 2017
microsoftfacultysummit.com

Microsoft Research
Microsoft.com/research

Facebook
@microsoftresearch

Twitter
@MSFTResearch
#FacSumm
#EdgeofAI
Questioning the Question
Answering Dataset

Percy Liang
Questioning the Question Answering Dataset

Percy Liang

Microsoft Faculty Summit — July 17, 2017
Datasets drive progress

<table>
<thead>
<tr>
<th>Input sentence</th>
<th>Translation (PBMT):</th>
<th>Translation (GATM):</th>
<th>Translation (Human):</th>
</tr>
</thead>
</table>
| 麥克繪出將於今年內赴加
總理年度訪華之行，與
加拿大總理杜魯多舉行
兩國首腦首次年度對
話。 | Li Keqiang premier
added this line to start
the annual dialogue
 mechanism with the
Canadian Prime Minister
Trudeau two prime
ministers held its first
annual session. | Li Keqiang will start the
annual dialogue
 mechanism with Prime
Minister Trudeau of
Canada and hold the first
annual dialogue between
the two premiers. | Li Keqiang will initiate the
annual dialogue
 mechanism between
 premiers of China and
 Canada during this visit,
and hold the first annual
dialogue with Premier
Trudeau of Canada. |
What about question answering?
What about question answering?

Desiderata: datasets that can’t be ”solved” by ”cheap tricks”
Executable semantic parsing

What is the largest city in Europe by population?
Executable semantic parsing

*What is the largest city in Europe by population?*

semantic parsing

Cities
Executable semantic parsing

What is the largest city in Europe by population?

semantic parsing

Cities Europe
Executable semantic parsing

What is the largest city in Europe by population?

Cities ContainedBy(Europe)
Executable semantic parsing

What is the largest city in Europe by population?

semantic parsing

Cities ∩ ContainedBy(Europe)
Executable semantic parsing

What is the largest city in Europe by population?

semantic parsing

Cities \cap ContainedBy(Europe) \quad Population
Executable semantic parsing

What is the largest city in Europe by population?

semantic parsing

argmax(Cities ∩ ContainedBy(Europe), Population)
Executable semantic parsing

What is the largest city in Europe by population?

semantic parsing

argmax(Cities ∩ ContainedBy(Europe), Population)

execute

Istanbul
WebQuestions (2013)

- **Realistic**: user queries from search engine
- **Most questions can’t be answered by Freebase**
- **Small dataset**: only 6K questions (started with 100K)
- **Noisy**: ceiling is 60-70% (current best is 55%)
- **Not much compositionality** (can be solved without semantic parsing)
In what city did Piotr’s last 1st place finish occur?
WikiTableQuestions (2015)

- Compositionality: requires reasoning
- Clean: crowdworkers made sure questions were answerable
- Gave up on realism
- State-of-the-art is 45%, ceiling is around 90%
- Dataset is too small given the complexity (22K questions)?
- A bit niche for the NLP community

In what city did Piotr’s last 1st place finish occur?
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud
SQuAD (2016)

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail. Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall? gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail? graupel

Where do water droplets collide with ice crystals to form precipitation? within a cloud

- Desiderata: large and clean
- 100K examples from 536 articles
- Answer is span of paragraph
- Train and test have disjoint articles
- Gave up on guarding against cheap tricks
- Questions don’t require much reasoning
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupe

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

- Desiderata: large and clean
- 100K examples from 536 articles
- Answer is span of paragraph
- Train and test have disjoint articles
- Gave up on guarding against cheap tricks
- Questions don’t require much reasoning
- ”necessary but not sufficient” dataset
submit code to run on test set ⇒ public models
<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>r-net (ensemble)</td>
<td>77.688</td>
<td>84.666</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>Interactive AoA Reader (ensemble)</td>
<td>76.492</td>
<td>83.745</td>
</tr>
<tr>
<td></td>
<td>Joint Laboratory of HIT and iFLYTEK Research</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>smarnet (ensemble)</td>
<td>75.989</td>
<td>83.475</td>
</tr>
<tr>
<td></td>
<td>Eigen Technology &amp; Zhejiang University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>DCN+ (single model)</td>
<td>74.866</td>
<td>82.806</td>
</tr>
<tr>
<td></td>
<td>Salesforce Research</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>MEMEN (ensemble)</td>
<td>75.370</td>
<td>82.658</td>
</tr>
<tr>
<td></td>
<td>Eigen Technology &amp; Zhejiang University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</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></td>
<td><a href="https://arxiv.org/abs/1609.05284">https://arxiv.org/abs/1609.05284</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</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>7</td>
<td>Mnemonic Reader (ensemble)</td>
<td>74.268</td>
<td>82.371</td>
</tr>
<tr>
<td></td>
<td>NUDT and Fudan University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SSAE (ensemble)</td>
<td>74.080</td>
<td>81.665</td>
</tr>
<tr>
<td></td>
<td>Tsinghua University</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- over 1 year
- 30 teams
- best system: 84.7%
- humans: 91.2%
- Advances in attention mechanisms
Other datasets

Fill in the blank:

- CNN/DailyMail (1.4M): cloze-style predict entity
- Children’s Book Test (688K), BookTest (14M): cloze-style predict word
Other datasets

Fill in the blank:

• CNN/DailyMail (1.4M): cloze-style predict entity
• Children’s Book Test (688K), BookTest (14M): cloze-style predict word
• LAMBADA: predict word (dataset constructed to need longer context)
• ROCStories (100K): predict wrong/right ending, model discourse
Other datasets

Fill in the blank:
• CNN/DailyMail (1.4M): cloze-style predict entity
• Children’s Book Test (688K), BookTest (14M): cloze-style predict word
• LAMBADA: predict word (dataset constructed to need longer context)
• ROCStories (100K): predict wrong/right ending, model discourse

Question answering via machine reading:
• MCTest: 2640 questions, multiple choice, reasoning
• AI2 science questions: 855 questions, knowledge, reasoning
Other datasets

Fill in the blank:
- CNN/DailyMail (1.4M): cloze-style predict entity
- Children’s Book Test (688K), BookTest (14M): cloze-style predict word
- LAMBADA: predict word (dataset constructed to need longer context)
- ROCStories (100K): predict wrong/right ending, model discourse

Question answering via machine reading:
- MCTest: 2640 questions, multiple choice, reasoning
- AI2 science questions: 855 questions, knowledge, reasoning
- NewsQA (100K): questions crowdsourced based on CNN summaries
- MS-MARCO (100K): real questions from users of search engines
- TriviaQA (650K): trivia questions made by enthusiasts
- Quasar (80K): cloze-style from StackOverflow, trivia questions
Are these datasets enough?
Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager.

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

BiDAF

John Elway
Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Jeff Dean is the name of the quarterback who was 37 in Champ Bowl XXXIV.

What is the name of the quarterback who was 38 in Super Bowl XXXIII?
Adversarial evaluation

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager.

Jeff Dean is the name of the quarterback who was 37 in Champ Bowl XXXIV.

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

BiDAF

Jeff Dean
## Results on public models

<table>
<thead>
<tr>
<th>Model</th>
<th>Original F1</th>
<th>Adversarial F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReasoNet-E</td>
<td>81.1</td>
<td>49.8</td>
</tr>
<tr>
<td>SEDT-E</td>
<td>80.1</td>
<td>46.5</td>
</tr>
<tr>
<td>BiDAF-E</td>
<td>80.0</td>
<td>46.9</td>
</tr>
<tr>
<td>Mnemonic-E</td>
<td>79.1</td>
<td>55.3</td>
</tr>
<tr>
<td>Ruminating</td>
<td>78.8</td>
<td>47.7</td>
</tr>
<tr>
<td>jNet</td>
<td>78.6</td>
<td>47.0</td>
</tr>
<tr>
<td>Mnemonic-S</td>
<td>78.5</td>
<td>56.0</td>
</tr>
<tr>
<td>ReasoNet-S</td>
<td>78.2</td>
<td>50.3</td>
</tr>
<tr>
<td>MPCM-S</td>
<td>77.0</td>
<td>50.0</td>
</tr>
<tr>
<td>RaSOR</td>
<td>76.2</td>
<td>49.5</td>
</tr>
<tr>
<td>BiDAF-S</td>
<td>75.5</td>
<td>45.7</td>
</tr>
</tbody>
</table>
## Results on public models

<table>
<thead>
<tr>
<th>Model</th>
<th>Original F1</th>
<th>Adversarial F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReasoNet-E</td>
<td>81.1</td>
<td>49.8</td>
</tr>
<tr>
<td>SEDT-E</td>
<td>80.1</td>
<td>46.5</td>
</tr>
<tr>
<td>BiDAF-E</td>
<td>80.0</td>
<td>46.9</td>
</tr>
<tr>
<td>Mnemonic-E</td>
<td>79.1</td>
<td>55.3</td>
</tr>
<tr>
<td>Ruminating</td>
<td>78.8</td>
<td>47.7</td>
</tr>
<tr>
<td>jNet</td>
<td>78.6</td>
<td>47.0</td>
</tr>
<tr>
<td>Mnemonic-S</td>
<td>78.5</td>
<td><strong>56.0</strong></td>
</tr>
<tr>
<td>ReasoNet-S</td>
<td>78.2</td>
<td>50.3</td>
</tr>
<tr>
<td>MPCM-S</td>
<td>77.0</td>
<td>50.0</td>
</tr>
<tr>
<td>RaSOR</td>
<td>76.2</td>
<td>49.5</td>
</tr>
<tr>
<td>BiDAF-S</td>
<td>75.5</td>
<td>45.7</td>
</tr>
<tr>
<td>Humans</td>
<td>92.6</td>
<td>89.2</td>
</tr>
</tbody>
</table>
Training versus testing

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Jeff Dean is the name of the quarterback who was 37 in Champ Bowl XXXIV.

- If retrain model on adversarial examples, 74.3 ⇒ 70.0
- If test this model on prepended sentences, 70.0 ⇒ 36.9
- Think of this as test-set evaluation only
Proposal: testing properties

Properties:

• Should be invariant under adding distracting sentences
Proposal: testing properties

Properties:

• Should be invariant under adding distracting sentences
• Should be invariant under paraphrase
• If replace 'John Elway' with entity X, should return X

![Diagram showing test examples and implied test examples]
Proposal: testing properties

Properties:

- Should be invariant under adding distracting sentences
- Should be invariant under paraphrase
- If replace 'John Elway' with entity X, should return X

• Ongoing: automatically generate much wider space of adversarial examples (adversaries = teachers)
Summary

• Tradeoff between clean dataset and realistic dataset

• Progression of datasets, each at the cusp of what’s solvable

• Stronger testing of models based on adversaries and properties
Thank you!
Thank you