How will we know when machines can read?

Matt Gardner, with many collaborators
MRQA workshop, November 4, 2019
Look mom, I can read like a human!

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Human Performance</td>
<td>86.831</td>
<td>89.452</td>
</tr>
<tr>
<td></td>
<td><em>Stanford University</em> (Rajpurkar &amp; Jia et al. '18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>ALBERT (ensemble model)</td>
<td>89.731</td>
<td>92.215</td>
</tr>
<tr>
<td></td>
<td><em>Google Research &amp; TTIC</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Human Performance</td>
<td>82.304</td>
<td>91.221</td>
</tr>
<tr>
<td></td>
<td><em>Stanford University</em> (Rajpurkar et al. '16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>XLNet (single model)</td>
<td>89.898</td>
<td>95.080</td>
</tr>
<tr>
<td></td>
<td><em>Google Brain &amp; CMU</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*ALBERT was updated on Sep 18, 2019.*

*XLNet was updated on May 21, 2019.*
Look mom, I can read like a human!

ALIBABA AI MODEL TOPS HUMANS IN READING COMPREHENSION

ADAM NAJBERG | JANUARY 15, 2018

ROBOTS CAN NOW READ BETTER THAN HUMANS, PUTTING MILLIONS OF JOBS AT RISK

BY ANTHONY CUTHBERTSON ON 1/15/18 AT 8:00 AM EST
The Matrix is a 1999 science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world."
So what’s the right evaluation?
Building the right test

- What format should the test be?
- What should be on the test?
- How do we evaluate the test?
Test format
What is reading?

Postulate: an entity *understands* a passage of text if it is able to answer *arbitrary questions* about that text.
Why is QA the right format?

It has issues, but really, what other choice is there? We don’t have a formalism for this.
What kind of QA?

**Input**

**Output**

- Numbers
- Dates
- Free text
- Anything we can evaluate

?
What about multiple choice, or NLI?

**MC**

___ ?

A) ___

B) ___ ✔

C) ___

D) ___

**NLI**

___

___ ❌

___ ✔

___ ❌

___ ❌

___ ❌
What about multiple choice, or NLI?

Both have same problems:

1. Distractors have biases
2. Low entropy output space
3. Machines (and people!) use different models for this
I propose standardizing on SQuAD-style inputs, arbitrary (evaluable) outputs
Test content
I really meant arbitrary

- The test won’t be convincing unless it has all kinds of questions, about every aspect of reading you can think of.
- So what are those aspects?
Sentence-level linguistic structure

SQuAD 2.0
The Stanford Question Answering Dataset
Sentence-level linguistic structure

But SQuAD just scratches the surface:

- Many other kinds of local structure
- Need to test coherence more broadly
DROP:
Discrete Reasoning Over Paragraphs
Denver would retake the lead with kicker Matt Prater nailing a 43-yard field goal, yet Carolina answered as kicker John Kasay ties the game with a 39-yard field goal. ... Carolina closed out the half with Kasay nailing a 44-yard field goal. ... In the fourth quarter, Carolina sealed the win with Kasay’s 42-yard field goal.

Which kicker kicked the most field goals? John Kasay
Discourse structure

- Tracking entities across a discourse
- Understanding discourse connectives and discourse coherence
- ...

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE
Quoref:
Question-based coreference resolution
Passage: Byzantines were avid players of tavli (Byzantine Greek: ταβλί), a game known in English as backgammon, which is still popular in former Byzantine realms, and still known by the name tavli in Greece. Byzantine nobles were devoted to horsemanship, particularly tzykanion, now known as polo. The game came from Sassanid Persia in the early period and a Tzykanisterion (stadium for playing the game) was built by Theodosius II (r. 408450) inside the Great Palace of Constantinople. Emperor Basil I (r. 867886) excelled at it; Emperor Alexander (r. 912913) died from exhaustion while playing, Emperor Alexios I Komnenos (r. 10811118) was injured while playing with Tatikios, and John I of Trebizond (r. 12351238) died from a fatal injury during a game. Aside from Constantinople and Trebizond, other Byzantine cities also featured tzykanisteria, most notably Sparta, Ephesus, and Athens, an indication of a thriving urban aristocracy.

Question: What is the Byzantine name of the game that Emperor Basil I excelled at?  
Answer: tzykanion

Question: What are the names of the sport that is played in a Tzykanisterion?  
Answer: tzykanion, polo

Question: What cities had tzykanisteria?  
Answer: Constantinople, Trebizond, Sparta, Ephesus, Athens
## Quoref:
### Question-based coreference resolution

<table>
<thead>
<tr>
<th>Reasoning</th>
<th>Paragraph Snippet</th>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronominal resolution (69%)</td>
<td>Anna and Declan eventually make their way on foot to a roadside pub, where they discover <strong>the three van thieves</strong> going through Anna’s luggage. Declan fights them, displaying unexpected strength for a man of his size, and retrieves Anna’s bag.</td>
<td>Who does Declan <strong>get into a fight</strong> with?</td>
<td>the three van thieves</td>
</tr>
<tr>
<td>Nominal resolution (54%)</td>
<td>Later, <strong>Hippolyta</strong> was granted a daughter, Princess Diana, ... <strong>Diana defies her mother</strong> and ...</td>
<td>What is the name of the person who is <strong>defied by her daughter</strong>?</td>
<td>Hippolyta</td>
</tr>
<tr>
<td>Multiple resolutions (32%)</td>
<td>The now upbeat collective keep <strong>the toucan</strong>, nick-naming it “Amigo” ... When authorities show up to catch the <strong>bird</strong>, Pete and Liz spirit <strong>him</strong> away by Liz hiding <strong>him</strong> in her dress ...</td>
<td>What is the name of the character who <strong>hides in Liz’s dress</strong>?</td>
<td>Amigo</td>
</tr>
<tr>
<td>Commonsense reasoning (10%)</td>
<td>Amos reflects back on his early childhood ... with <strong>his mother Fania and father Arieh</strong>. ... One of <strong>his mother</strong>’s friends is killed while hanging up laundry during the war. ... <strong>Fania</strong> falls into a depression. ... <strong>she</strong> ... goes to ... Tel Aviv, where <strong>she</strong> kills <strong>herself</strong> by overdose ...</td>
<td>How does Arieh’s wife die?</td>
<td>kills herself by overdose</td>
</tr>
</tbody>
</table>
Implicative meaning

- What do the propositions in the text imply about other propositions I might see in other text?
- E.g., “Bill loves Mary”, “Mary gets sick” → “Bill is sad”
- Where do these implications come from?
**Background:** Scientists think that the earliest flowers attracted insects and other animals, which spread pollen from flower to flower. This greatly increased the efficiency of fertilization over wind-spread pollen, which might or might not actually land on another flower. To take better advantage of this animal labor, plants evolved traits such as brightly colored petals to attract pollinators. In exchange for pollination, flowers gave the pollinators nectar.
Situation: Last week, John visited the national park near his city. He saw many flowers. His guide explained him that there are two categories of flowers, category A and category B. **Category A flowers spread pollen via wind, and category B flowers spread pollen via animals.**
Background: Scientists think that the earliest flowers attracted insects and other animals, which spread pollen from flower to flower. This greatly increased the efficiency of fertilization over wind-spread pollen, which might or might not actually land on another flower. To take better advantage of this animal labor, plants evolved traits such as brightly colored petals to attract pollinators. In exchange for pollination, flowers gave the pollinators nectar.

Situation: Last week, John visited the national park near his city. He saw many flowers. His guide explained him that there are two categories of flowers, category A and category B. Category A flowers spread pollen via wind, and category B flowers spread pollen via animals.

Question: Would category B flower have more or less efficient fertilization than category A flower?

Answer: more
Time

- Temporal ordering of events
- Duration of events
- Which things are events in the first place?
Grounding

- Common sense
- Factual knowledge
- More broadly: speaker is trying to communicate world state, and in a person it induces a mental model of that world state. We need to figure out ways to probe these mental models.
Passage: While Mr. Mueller found insufficient evidence to bring charges against President Trump for conspiring or colluding with Russia to influence the 2016 elections, he cited at least 10 specific instances in which Mr. Trump may have obstructed his investigation. After the release of the Mueller report, Ms. Pelosi promised a series of hearings and investigations that would allow the American people to see the facts for themselves and decide whether impeachment was warranted.

Question: What did the special counsel cite?  
Answer: 10 specific instances.

Question: What did the speaker of the house cite?  
Answer: no answer

Question: What did speaker of the house promise?  
Answer: a series of hearings.

Question: What did the president cite?  
Answer: no answer
Passage: I’m afraid to sit in case I wrinkle the fabric or stain it with something I don’t even know is on my pants. The couch is cream but inlaid with a fine green silk. The white curtains are linen, the kind of white that is untouched by hands and devoid of dust. There is no television, no bookshelf, no dining table, only the chairs arranged around the bespoke fireplace which leaps with a gas flame. The photographs are black and white, not casual family snaps, but arranged to look like such by a professional. The floor is a high polished wood, dark and free of either dust or clutter.

Question: Which things could you sit on?
Answer: couch, chairs, floor

Question: Which things show pictures on them?
Answer: television, photographs

Question: Has the room been cleaned recently?
Answer: yes
Many, many, many, more...

- Pragmatics, factuality
- Coordination, distributive vs. non-distributive
- Deixis
- Aspectual verbs
- Bridging and other elided elements
- Negation and quantifier scoping
- Distribution of quantifiers
- Preposition senses
- Noun compounds
- ...
Test evaluation
How do we evaluate generative QA?

- This is a serious problem that severely limits our test
- No solution yet, but we’re working on it
- See Anthony’s talk for more detail
What about reasoning shortcuts?

- It’s easy to write questions that don’t test what you think they’re testing
- See our MRQA paper for more on how to combat this
What about generalization?

- There is growing realization that the traditional supervised learning paradigm is broken in high level, large-dataset NLP - we’re fitting artifacts
- The test should include not just hidden test data, but hidden test data from a different distribution than the training data
- MRQA has the right idea here
- That is, we should explicitly make test sets without training sets (as long as they are close enough to training that it should be possible to generalize)
A beginning, and a call for help
An Open Reading Benchmark

- Evaluate one model on all of these questions at the same time
- Standardized (SQuAD-like) input, arbitrary output
- Will grow over time, as more datasets are built
An Open Reading Benchmark

ORB
ORB is an evaluation server which tests a single reading comprehension model's performance on... (More)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Submission</th>
<th>DROP F1</th>
<th>DuoRC F1</th>
<th>NewsQA F1</th>
<th>Quoref F1</th>
<th>Ropes EM</th>
<th>Narr. MET.</th>
<th>SQuAD1 F1</th>
<th>SQuAD2 F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NABERT AI2</td>
<td>21.87</td>
<td>34.28</td>
<td>46.19</td>
<td>38.39</td>
<td>47.96</td>
<td>0.33</td>
<td>78.55</td>
<td>39.17</td>
</tr>
</tbody>
</table>
An *Open* Reading Benchmark

- Making a good test is a bigger problem than any one group can solve
- We need to work together to make this happen
- We will add any good dataset that matches the input format
To conclude:

- Current reading comprehension benchmarks are insufficient to convince a reasonable researcher that machines can read.
- There are a lot of things that need to be tested before we will be convinced.
- We need to work together to make a sufficient test - there’s too much for anyone to do on their own.

Thanks!