How will we know when machines can read?

Matt Gardner, with many collaborators

MRQA workshop, November 4, 2019

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Look mom, I can read like a human!

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	**89.731**	**92.215**

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 May 21, 2019	XLNet (single model) *Google Brain & CMU*	**89.898**	**95.080**

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

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

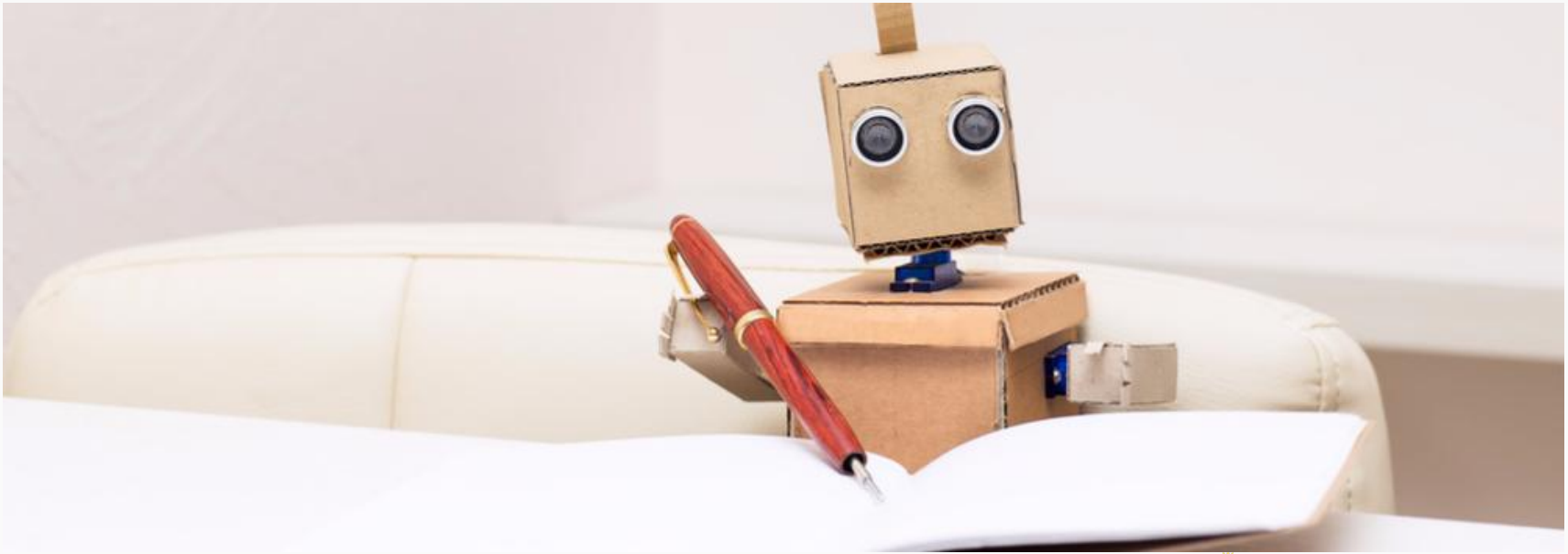
But...

Question

How many people star in The Matrix?

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

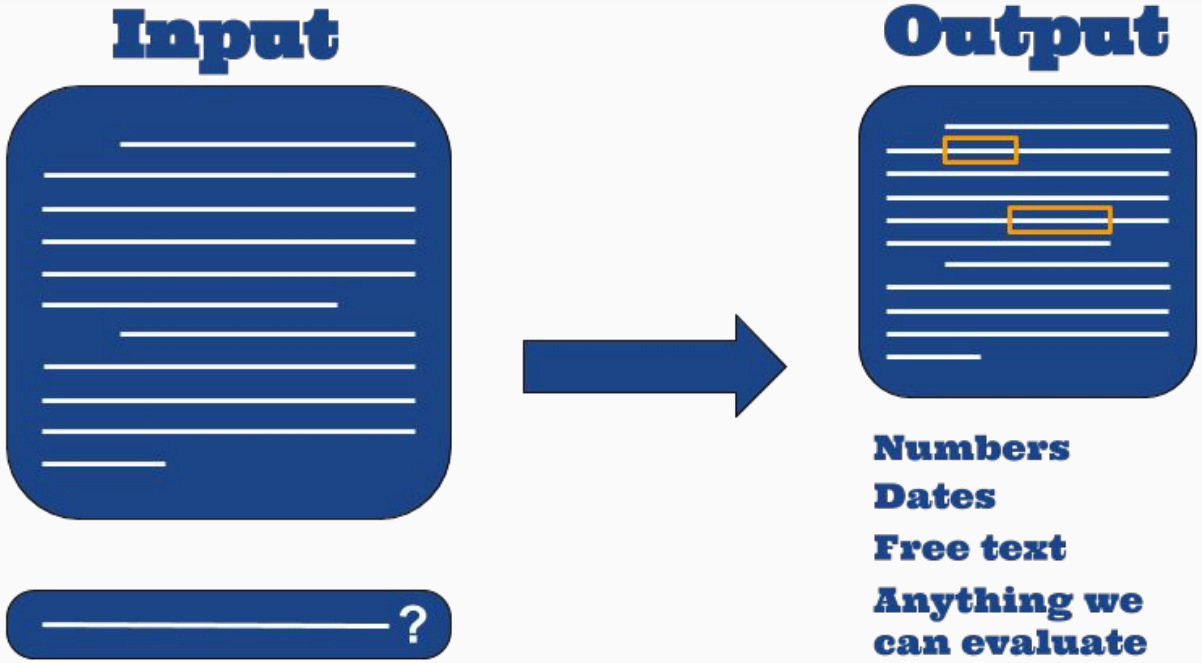
What is reading?

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?

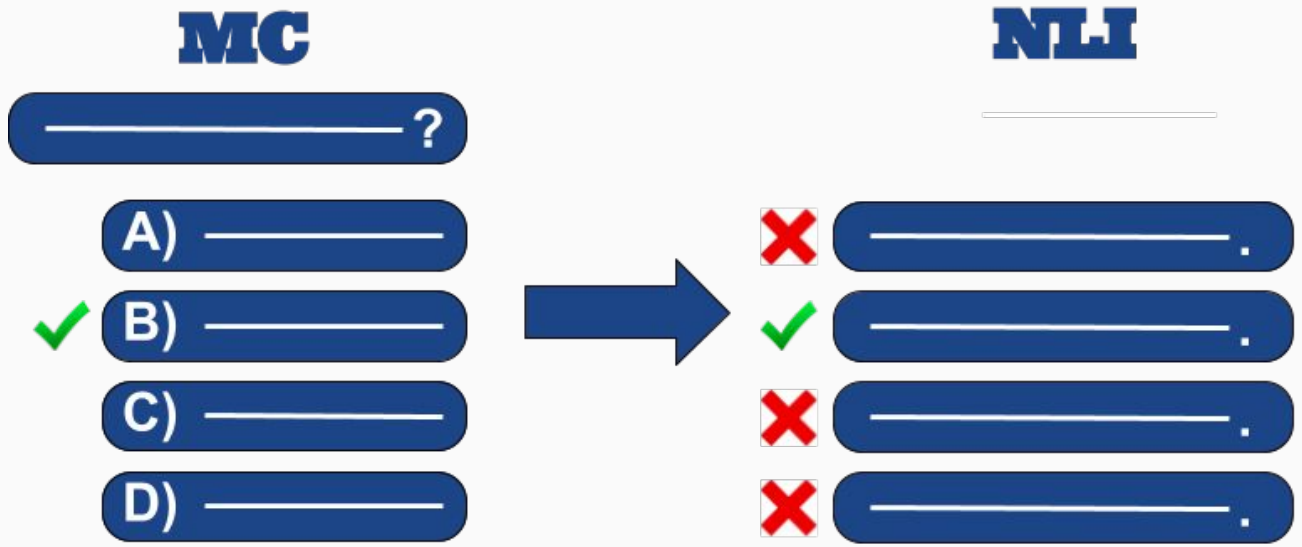
It has issues, but really, what other choice is there?
We don’t have a formalism for this.

What kind of QA?



ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

What about multiple choice, or NLI?



ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

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

Bottom line

I propose standardizing on SQuAD-style inputs, arbitrary (evaluable) outputs

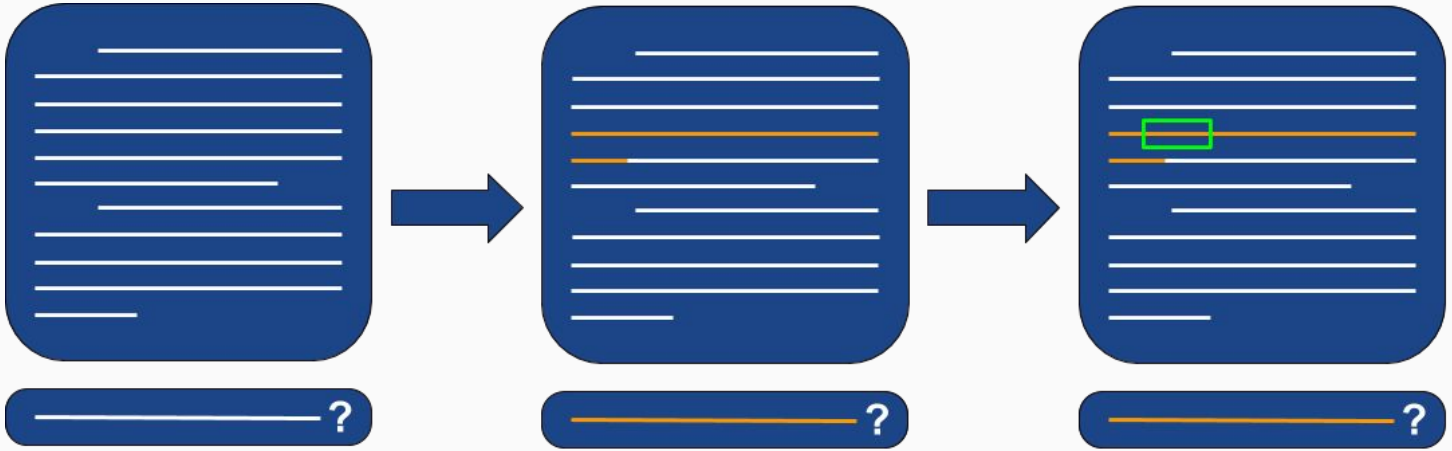
Test content

ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE

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



SQuAD2.0
The Stanford Question Answering Dataset

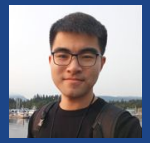
ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE

Sentence-level linguistic structure

But SQuAD just scratches the surface:

- Many other kinds of local structure
- Need to test coherence more broadly

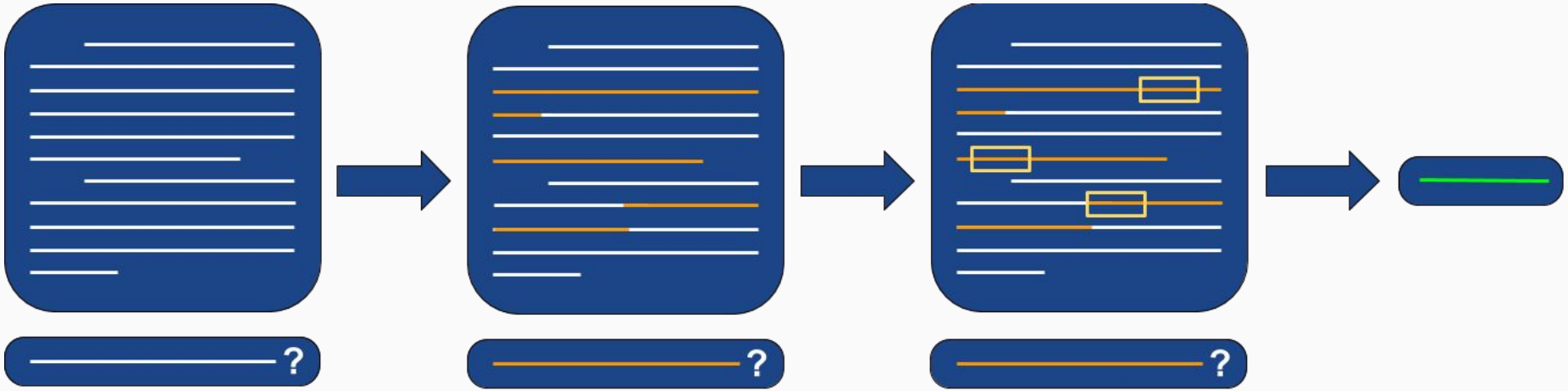
NAACL 2019



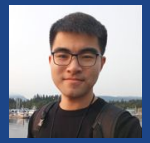
DROP:

Discrete Reasoning Over Paragraphs



ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

NAACL 2019



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

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Discourse structure

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

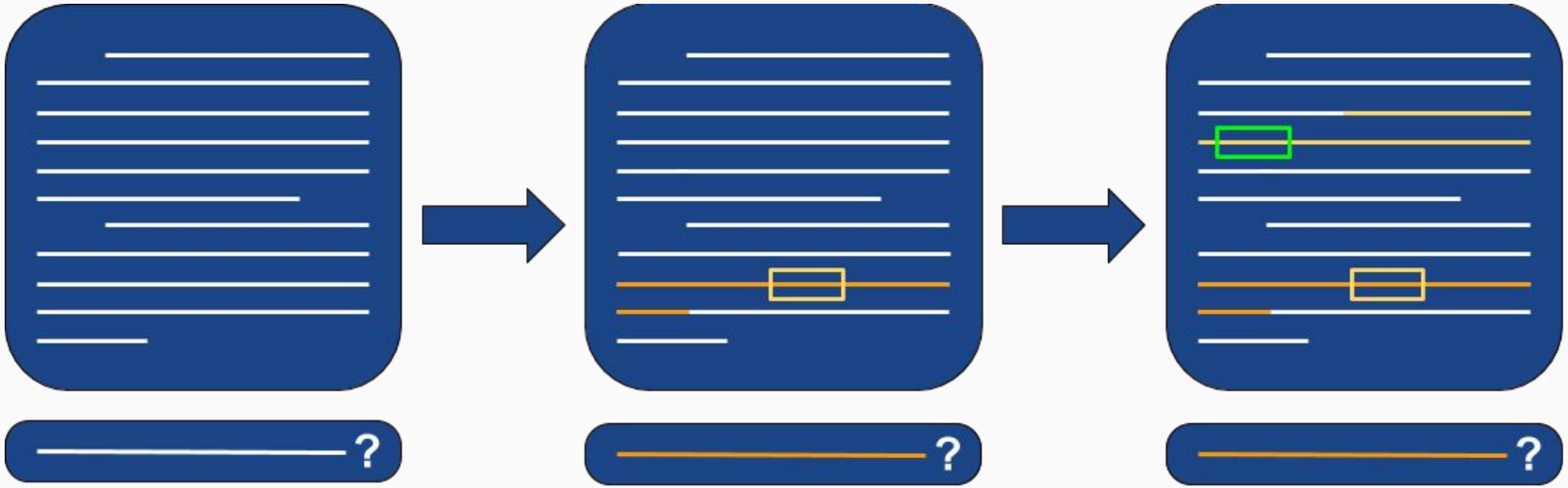
EMNLP 2019





Quoref:

Question-based coreference resolution



ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

EMNLP 2019





Quoref:

Question-based coreference resolution

Passage: Byzantines were avid players of tavli (Byzantine Greek:), 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

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

EMNLP 2019





Quoref:

Question-based coreference resolution

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

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

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?

MRQA 2019




ROPES:

Reasoning Over Paragraph Effects in Situations

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.

ROPES:

Reasoning Over Paragraph Effec

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**.

ROPES:

Reasoning Over Paragraph Effec

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.

Grounding

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

Grounding

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

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

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

ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE

MRQA 2019
Best paper




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

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

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

ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE




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

MRQA 2019




An Open Reading Benchmark

LEADERBOARD

Home Leaderboards Hello, dDua

ORB

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

+ Create Submission

Public Submissions | My Submissions | Getting Started | About

Rank	Submission	DROP F1	DuoRC F1	NewsQA F1	Quoref F1	Ropes EM	Narr. MET.	SQuAD1 F1	SQuAD2 F1
1	NABERT *AI2*	21.87	34.28	46.19	38.39	47.96	0.33	78.55	39.17

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

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!

We're hiring!

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE