Natural Language Understanding Systems Learn to Understand or to Find Shortcuts?

Naoya Inoue
Independent Researcher
About me: Naoya Inoue (井之上 直也)

• Current position
  – Independent researcher
  – Previously (until March):
    • Assistant professor at Tohoku University, Japan
    • Visiting researcher at RIKEN Center for Advanced Intelligence Project
  – Joining SBU-AI soon! (supposed to be from April, but...)

• Research interests
  – Discourse
  – Commonsense reasoning (abduction)
  – Knowledge base
  – Formal logic
  – Machine Reading Comprehension
Recent findings in NLU research

• NLU systems achieve superhuman performance
  – Causal commonsense reasoning
  – Reading comprehension
  – Part of the third AI summer (Kautz’20)

• Is it attributed to NLU systems’ ability of language understanding?
Choice of Plausible Alternatives (Roemmele+’11) (COPA)

• Benchmark for testing causal commonsense
  – Input: Premise, Question type (Cause/Effect), Two Alternatives
  – Output: The most likely cause or effect
    • Premise: The stain came out of the shirt. What was the CAUSE of this?
    • Alternative 1: I bleached the shirt. (Correct)
    • Alternative 2: I patched the shirt.

• Dataset
  – Dev: 500, Test: 500 (carefully authored by human)
  – Human performance: 100%

http://people.ict.usc.edu/~gordon/copa.html
Breakthrough in COPA

General-purpose Pretrained Language Models
(BERT, RoBERTa, T5)
Paragraph A: Mike Bryan
Michael Carl Bryan (born April 29, 1978) is an American professional tennis player. The right-hander turned professional in 1998. With his twin brother Bob, he has been world No. 1 doubles player for much of the last several years, first achieving top ranking in September 2003, and they became the second men's doubles team to complete the career golden slam by winning the 2012 Summer Olympics. He currently holds the records for the most ATP men's doubles titles (115) and final appearances (171), and is tied for fourth all-time in men's titles in the Open Era, all with his brother with the exception of one with Mahesh Bhupathi and another with Mark Knowles.

Paragraph B: Ray Ruffels
Raymond Owen "Ray" Ruffels (born 23 March 1946 in Sydney) is an Australian former professional tennis player and coach.

Q: Are Mike Bryan and Ray Ruffels both tennis players?
A: yes

- 100+ RC papers have been released in 4 years (Liu+’19)

https://hotpotqa.github.io/explorer.html
SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human Performance</td>
<td>86.831</td>
<td>89.452</td>
</tr>
<tr>
<td></td>
<td>Stanford University</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Rajpurkar &amp; Jia et al. ’18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ALBERT + DAAF + Verifier (ensemble)</td>
<td>90.002</td>
<td>92.425</td>
</tr>
<tr>
<td></td>
<td>PINGAN Omni-Sinitic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ALBERT (ensemble model)</td>
<td>89.731</td>
<td>92.215</td>
</tr>
<tr>
<td></td>
<td>Google Research &amp; TTIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>XLNet + DAAF + Verifier (ensemble)</td>
<td>88.592</td>
<td>90.859</td>
</tr>
<tr>
<td></td>
<td>PINGAN Omni-Sinitic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>albert+verifier (single model)</td>
<td>88.355</td>
<td>91.019</td>
</tr>
<tr>
<td></td>
<td>Ping An Life Insurance Company Al Team</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ALBERT (single model)</td>
<td>88.107</td>
<td>90.902</td>
</tr>
<tr>
<td></td>
<td>Google Research &amp; TTIC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

https://rajpurkar.github.io/SQuAD-explorer/  
(2019/12/23)
<table>
<thead>
<tr>
<th>Model</th>
<th>Code</th>
<th>Ans</th>
<th>Sup</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGN-large (single model)</td>
<td></td>
<td>69.22</td>
<td>82.19</td>
<td>62.76</td>
</tr>
<tr>
<td>Anonymous</td>
<td>Dec 1, 2019</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2F Reader (single model)</td>
<td></td>
<td>67.98</td>
<td>81.24</td>
<td>60.81</td>
</tr>
<tr>
<td>Joint Laboratory of HIT and iFLYTEK Research</td>
<td>Oct 18, 2019</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAE-large (single model)</td>
<td></td>
<td>66.92</td>
<td>79.62</td>
<td>61.53</td>
</tr>
<tr>
<td>JD AI Research</td>
<td>Nov 19, 2019</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tu, Huang et al., AAAI 2020</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HGN (single model)</td>
<td></td>
<td>66.07</td>
<td>79.36</td>
<td>60.33</td>
</tr>
<tr>
<td>Microsoft Dynamics 365 AI Research</td>
<td>Sep 27, 2019</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fang et al., 2019</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAP 2 (ensemble)</td>
<td></td>
<td>66.64</td>
<td>79.82</td>
<td>57.21</td>
</tr>
<tr>
<td>EPS + BERT(wwm) (single model)</td>
<td>Oct 1, 2019</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anonymous</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Human F$_1$: 91/90/83

https://hotpotqa.github.io/
(2019/12/23)
COPA’s superficial cues (Kavumba+’19)

• Causality is about between two events, but...

• RoBERTa that encodes only alternatives achieves an accuracy of 59.6 (±2.3) (c.f. chance rate: 50%)

• Found uneven frequency distribution of unigrams in correct alternatives and wrong alternatives (= superficial cues)
### went signals causality?

<table>
<thead>
<tr>
<th>Premise</th>
<th>Question</th>
<th>Correct Alternative</th>
<th>Wrong Alternative</th>
</tr>
</thead>
<tbody>
<tr>
<td>The woman's ring slipped off in the shower.</td>
<td>Effect?</td>
<td>The ring <strong>went</strong> down the drain.</td>
<td>The woman polished the ring.</td>
</tr>
<tr>
<td>The seamstress pushed the threaded needle into the fabric.</td>
<td>Effect?</td>
<td>The thread <strong>went</strong> through the fabric.</td>
<td>The thread wrapped around the needle.</td>
</tr>
<tr>
<td>The man's eye became infected.</td>
<td>Effect?</td>
<td>He <strong>went</strong> blind.</td>
<td>He put on glasses.</td>
</tr>
<tr>
<td>The police officer dropped the gun.</td>
<td>Effect?</td>
<td>The gun <strong>went</strong> off.</td>
<td>The gun recoiled.</td>
</tr>
<tr>
<td>I rubbed the soap between my hands.</td>
<td>Effect?</td>
<td>The soap foamed.</td>
<td>My hands <strong>went</strong> numb.</td>
</tr>
<tr>
<td>A burglar broke into the house.</td>
<td>Effect?</td>
<td>The security alarm <strong>went</strong> off.</td>
<td>The homeowners were asleep.</td>
</tr>
<tr>
<td>The driver pulled over to the side of the road.</td>
<td>Cause?</td>
<td>He noticed a stranded vehicle.</td>
<td>He <strong>went</strong> through an intersection.</td>
</tr>
<tr>
<td>The host served dinner to his guests.</td>
<td>Effect?</td>
<td>His guests were gracious.</td>
<td>His guests <strong>went</strong> hungry.</td>
</tr>
<tr>
<td>The man held his breath.</td>
<td>Cause?</td>
<td>He <strong>went</strong> underwater.</td>
<td>He treded water.</td>
</tr>
<tr>
<td>The boy felt homesick.</td>
<td>Cause?</td>
<td>He <strong>went</strong> away to camp.</td>
<td>He bickered with his sister.</td>
</tr>
<tr>
<td>The girl met her favorite actor.</td>
<td>Effect?</td>
<td>She asked him for his autograph.</td>
<td>She <strong>went</strong> to see his new film.</td>
</tr>
<tr>
<td>My mailbox was overflowing with letters.</td>
<td>Cause?</td>
<td>I <strong>went</strong> on vacation for two weeks.</td>
<td>The lid on the mailbox broke.</td>
</tr>
<tr>
<td>The woman wanted to be a doctor.</td>
<td>Effect?</td>
<td>She <strong>went</strong> to medical school.</td>
<td>She visited the hospital.</td>
</tr>
<tr>
<td>The man was bitten by mosquitoes.</td>
<td>Cause?</td>
<td>He <strong>went</strong> camping in the woods.</td>
<td>He fell asleep on his couch.</td>
</tr>
<tr>
<td>I needed to get cash.</td>
<td>Effect?</td>
<td>I <strong>went</strong> to the bank.</td>
<td>I bought a wallet.</td>
</tr>
<tr>
<td>The rain subsided.</td>
<td>Effect?</td>
<td>I <strong>went</strong> for a walk.</td>
<td>I browsed the internet.</td>
</tr>
<tr>
<td>The teenager grew taller than his father.</td>
<td>Cause?</td>
<td>He <strong>went</strong> through a growth spurt.</td>
<td>He joined the basketball team.</td>
</tr>
<tr>
<td>The woman hit a midlife crisis.</td>
<td>Effect?</td>
<td>She <strong>went</strong> sky diving.</td>
<td>She retired.</td>
</tr>
<tr>
<td>The driver got a flat tire.</td>
<td>Cause?</td>
<td>He ran over a nail.</td>
<td>He <strong>went</strong> over the speed limit.</td>
</tr>
<tr>
<td>The driver turned on the car's headlights.</td>
<td>Cause?</td>
<td>The sun <strong>went</strong> down.</td>
<td>He heard thunder.</td>
</tr>
<tr>
<td>The animal species became endangered.</td>
<td>Cause?</td>
<td>Their habitat was destroyed.</td>
<td>Their predators <strong>went</strong> extinct.</td>
</tr>
</tbody>
</table>
MRC datasets’ flaws

• Sugawara+ (‘18): in TriviaQA, 40% of problems can be solved by cheap heuristics (who => person!)
• Chen+ (‘19): in WikiHop, 60% of problems can be solved without looking at given paragraphs
• Multi-hop QA datasets do not actually need multi-hop reasoning (Min+’19, Wang+’19, Chen+’19, Jiang+’19)
• … followed by similar findings in other areas (e.g. SNLI (Gururangan+’18), Argument Mining (Niven+’19))
Article: Super Bowl 50
Paragraph: “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.

Question: “What is the name of the quarterback who was 38 in Super Bowl XXXIII?”
Original Prediction: John Elway
Why such superficial cues matter

- These cues are data-specific, not generalizable to other domains

- Prevent models from learning to solve the task using task-related cues

- **CAUTION:** Superficial cues are not always bad
  - If your problems can be solved with superficial cues *in general*, exploit them!
Our research questions

• When superficial cues are not available in test set, what will happen to the superhuman models? (Kavumba+COIN2019)

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conventional</strong></td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td><strong>Kavumba+’19</strong></td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td><strong>Kavubma+’19</strong></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

• How can we quantitatively measure correctness of models’ internal reasoning? (Inoue+ACL2020)
Categorizing test set

• RoBERTa that encodes only alternatives achieves an accuracy of 59.6 (±2.3) (c.f. chance rate: 50%)

• We divide the test set into:
  – **Cheatable subset**: 190 instances that are correctly solved by alternative-only RoBERTa (with 3 random seeds)
  – **Non-cheatable subset**: remaining 310 instances
Results

- Worse performance on non-cheatable subset: BERT and RoBERTa relied on superficial cues

<table>
<thead>
<tr>
<th>Superficial cues available?</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>Kavumba+’19</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td>Kavubma+’19</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
Mitigating superficial cues in train set

• Create new COPA instance so that wrong alternative becomes correct one

  - **Premise:** The stain came out of the shirt. What was the CAUSE of this?
  - **Alternative 1:** I bleached the shirt. (Correct)
  - **Alternative 2:** I patched the shirt.

  - **Premise:** *The shirt did not have a hole anymore.* What was the CAUSE of this?
  - **Alternative 1:** I bleached the shirt.
  - **Alternative 2:** I patched the shirt. (Correct)

  – Word freq. distribution in alternatives is not skewed anymore
  – Named Balanced COPA (B-COPA), 1,000 insts. in total

  https://balanced-copa.github.io/

• Use B-COPA for training, COPA for testing
Results

- Surprisingly: improvements on non-cheatable subset
  - When superficial cues are present, they try to exploit them
  - When they are eliminated, BERT/RoBERTa now seems to use task-related cues! … really?
Our research questions

• When superficial cues are not available in test set, what will happen to the superhuman models? (Kavumba+COIN2019)

<table>
<thead>
<tr>
<th></th>
<th>Superficial cues available?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
</tr>
<tr>
<td>Conventional</td>
<td>O</td>
</tr>
<tr>
<td>Kavumba+’19</td>
<td>O</td>
</tr>
<tr>
<td>Kavubma+’19</td>
<td>X</td>
</tr>
</tbody>
</table>

• How can we quantitatively measure correctness of models’ internal reasoning? (Inoue+ACL2020)
**R^4C: Right for Right Reasons RC**

- Benchmark for evaluating MRC models’ internal reasoning
- Requires to provide derivation (entity relations) – Explanation as to how the answer is derived

★★ Crowdsourced mid-scale dataset (4.6K)
★★ Quantitative evaluation measure

**Question**
What was the former band of the member of Mother Love Bone who died just before the release of “Apple”?

**Articles**

**Title:** Return to Olympus  [1] Return to Olympus is the only album by the alternative rock band Malfunkshun.  
[2] It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died...  
[3] Stone Gossard had compiled...

**Title:** Mother Love Bone  [4] Mother Love Bone was an American rock band that...  
[5] The band was active from...  
[6] Frontman Andrew Wood’s personality and compositions helped to catapult the group to...  
[7] Wood died only days before the scheduled release of the band’s debut album, “Apple”, thus ending the...

**Explanation**
Supporting facts (SFs):  
[1], [2], [4], [6], [7]

**Answer**
Malfunkshun

---

https://naoya-i.github.io/r4c/
Dataset statistics

• Annotated on top of HotpotQA (Yang+’18)
  – HotpotQA: Multi-hop QA dataset

• Each instance has 3 reference derivations
  – To account for linguistic variation

<table>
<thead>
<tr>
<th>Split</th>
<th># QA</th>
<th># derivations</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>2,379</td>
<td>7,137</td>
</tr>
<tr>
<td>dev</td>
<td>2,209</td>
<td>6,627</td>
</tr>
<tr>
<td>total</td>
<td>4,588</td>
<td>13,764</td>
</tr>
</tbody>
</table>
Overview

Artificial intelligence (AI) has become fast and accurate. However, it still falls short when it comes to human-level reasoning. In this task, you are helping us educate an AI on how to reason. Specifically, we will ask you to verify whether you can answer a question from articles and describe the steps you used to find the answer.

**NOTE**
High-quality workers will be invited to special HITs that offer a higher reward. To be eligible, complete at least five HITs.

Instructions

View instructions (for first time workers)

Task

1. Read

Read the following question and related articles carefully. Important sentences are highlighted.

**Question**

The 1988 American comedy film, *The Great Outdoors*, starred a four-time Academy Award nominee, who received a star on the Hollywood Walk of Fame in what year?

**Articles**

**Article 1: The Great Outdoors (Film)**

The Great Outdoors is a 1988 American comedy film directed by Howard Deutch, and written and produced by John Hughes. It stars Dan Aykroyd, John Candy, Stephanie Faracy and Annette Bening in her film debut.

**Article 2: Annette Bening**

Annette Carol Bening (born May 29, 1958) is an American actress. She began her career on stage with the Colorado Shakespeare Festival company in 1980, and played Lady Macbeth in 1984 at the American Conservatory Theatre. She was nominated for the 1987 Tony Award for Best Featured Actress in a Play for her Broadway debut in "Coastal Disturbances". She is a four-time Academy Award nominee, for "The Grifters" (1990), "American Beauty" (1999), "Being Julia" (2004) and "The Kids Are All Right" (2010). In 2009, she received a star on the Hollywood Walk of Fame.

2. Answer

Based upon the related articles, answer the question:

The 1988 American comedy film, *The Great Outdoors*, starred a four-time Academy Award nominee, who received a star on the Hollywood Walk of Fame in what year?

Choose your answer:

Kevin Bacon

3. Explain

Describe the step(s) you used to infer your answer from the above articles. Click on a sentence in the article first (on your left), and then describe your reasoning (on your right).

Write at least one reasoning step for each article.

**Articles**

Select sentences relevant to your reasoning.

**Article 1: The Great Outdoors (Film)**

The Great Outdoors is a 1988 American comedy film directed by Howard Deutch, and written and produced by John Hughes. It stars Dan Aykroyd, John Candy, Stephanie Faracy and Annette Bening in her film debut.

**Article 2: Annette Bening**

Annette Carol Bening (born May 29, 1958) is an American actress. She began her career on stage with the Colorado Shakespeare Festival company in 1980, and played Lady Macbeth in 1984 at the American Conservatory Theatre. She was nominated for the 1987 Tony Award for Best Featured Actress in a Play for her Broadway debut in "Coastal Disturbances". She is a four-time Academy Award nominee, for "The Grifters" (1990), "American Beauty" (1999), "Being Julia" (2004) and "The Kids Are All Right" (2010). In 2009, she received a star on the Hollywood Walk of Fame.

4. Self-check

Self-check your reasoning steps. If your reasoning step does not meet the following requirements, please go back to step 3.

1. Each reasoning step is grammatically correct. Typical mistake:
   - BAD: The hood office is [Delhi] — [Delhi]
   - BAD: The hood office is in [Delhi] — [Delhi]
   - BAD: [Delhi] is an author of [The Hobbit] — [Delhi]
   - BAD: [Delhi] received the award in [2002] — [Delhi]
   - BAD: [Delhi] is a student of [The Hobbit] — [Delhi]

2. Each reasoning step does not contain propositions that are not in the text. Typical mistake:
   - BAD: [Delhi] is an author of [The Hobbit] — [Delhi]
   - BAD: [Delhi] received the award in [2002] — [Delhi]

3. Each reasoning step does not contain complex expressions. Typical mistake:
   - BAD: [Delhi] is an author of [The Hobbit] — [Delhi]
   - BAD: [Delhi] received the award in [2002] — [Delhi]

4. Submit/Check does not contain propositions that are not in the text. Typical mistake:
   - BAD: [Delhi] is an author of [The Hobbit] — [Delhi]
   - BAD: [Delhi] received the award in [2002] — [Delhi]

5. Submit

We sincerely appreciate your cooperation! We will approve your submission within 24 hours if there is no problem. Feel free to give us your feedback (optional):
Derivation annotation interface

**Articles:**
Select sentences relevant to your reasoning.

**Article 1: The Great Outdoors (film)**
The Great Outdoors is a 1988 American comedy film directed by Howard Deutch, and written and produced by John Hughes.
It stars Dan Aykroyd, John Candy, Stephanie Faracy and Annette Bening in her film debut.

**Article 2: Annette Bening**
Annette Carol Bening (born May 29, 1958) is an American actress.
She began her career on stage with the Guthrie Theatre Festival company in 1980, and played Lady Macbeth in 1984 at the American Conservatory Theatre.
She was nominated for the 1987 Tony Award for Best Featured Actress in a Play for her Broadway debut in "Coastal Disturbances".
She is a four-time Academy Award nominee; for "The Grifters" (1990), "American Beauty" (1999), "Being Julia" (2004) and "The Kids Are All Right" (2010).
In 2006, she received a star on the Hollywood Walk of Fame.

**Reasoning:**
Write your reasoning steps in a simple form subject-verb-object. You may rely on words from “suggestions” generated automatically, but editing may be needed.

[The Great Outdoors (film)] star [Annette Bening] (Article 1)

She is a four-time Academy Award nominee; for "The Grifters" (1990), "American Beauty" (1999), "Being Julia" (2004) and "The Kids Are All Right" (2010).

From the above sentence, what information did you infer?

Who/what:
Annette Bening

Suggestions: Annette Bening, a four-time Academy Award nominee

Did what:

Suggestions: the kids are all right is has

To whom/which:

Suggestions: Annette Bening, a four-time Academy Award nominee

1) Click
2) Enter
3) Finalize

Automatically extracted candidates
Entity/relation separated input
Evaluation measure

• System output: derivation $D$
• Golden derivation: $G_1, G_2, G_3$

• Calculate the maximum number of correct entity relations in $D$:

$$c(D) = \max_i c(D; G_i)$$

Precision: $c(D) / |D|$  
Recall: $c(D) / |G_i|$

• $c(D; G_i)$ is goodness of $D$ w.r.t. $G_i$ in $[0, \min(|D|, |G_i|)]$
Evaluation measure: $c(D; G_i)$

- There exist several possible one-to-one alignments between derivation steps in $D$ and those in $G_i$.

![Diagram of alignments]

- Again, find best one: $c(D; G_i) = \max_j c(D; G_i, A_j)$
- Alignment score: $c(D; G_i, A_j) = \sum_{d_j, g_j \in A_j} a(d_j, g_j)$
- $a$: similarity function between derivation steps
Call for challengers!

https://naoya-i.github.io/r4c/
Summary

• NLU systems are competitive to human, relying on data-specific superficial cues (i.e. finding shortcuts rather than understanding)

• Eliminating superficial cues have a positive impact on performance on non-cheatable dataset

• Ready to evaluate NLU models’ internal reasoning (R⁴C)

• Future work
  – Developing interpretable MRC for solving R⁴C task with neural-symbolic approach (currently in survey phase)