Commonsense Reasoning in the Wild

Xiang Ren

Joint work w/ Bill Yuchen Lin, Pei Zhou, Qinyuan Ye, Jay Pujara, Yejin Choi, Chandra Bhagavatula, William Cohen

Department of Computer Science & Information Science Institute
University of Southern California
http://inklab.usc.edu
NLP Models on Research Benchmarks

How are you?

Fine, thanks!

<table>
<thead>
<tr>
<th>Superhuman Performance</th>
<th>Human Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>90.6 91.0 98.6/99.2 97.4 88.6/83.2 94.7/94.2 92.6 77.4 97.3 88.6 92.7/94.7</td>
<td></td>
</tr>
<tr>
<td>90.4 91.4 95.8/97.6 96.0 88.3/83.0 94.2/93.5 93.0 77.9 96.6 69.1 92.7/91.9</td>
<td></td>
</tr>
<tr>
<td>90.3 90.4 95.7/97.6 98.4 86.2/83.7 94.5/94.1 93.2 77.5 95.9 66.7 93.3/93.8</td>
<td></td>
</tr>
<tr>
<td>89.8 89.0 95.8/98.9 100.0 81.6/85.1 91.7/91.3 93.6 80.0 100.0 76.6 99.3/99.7</td>
<td></td>
</tr>
<tr>
<td>89.3 91.2 93.5/96.8 94.8 88.1/83.3 94.1/93.4 92.5 76.9 93.8 85.6 92.7/91.9</td>
<td></td>
</tr>
<tr>
<td>86.7 87.8 94.4/96.0 93.6 84.6/85.1 90.1/89.6 89.1 74.6 93.2 58.0 87.1/74.4</td>
<td></td>
</tr>
<tr>
<td>86.1 88.1 92.4/95.4 91.8 84.6/84.7 85.6/88.3 88.8 74.1 93.2 75.6 98.3/96.3</td>
<td></td>
</tr>
</tbody>
</table>
Our NLP model achieved superhuman performance on the XYZ benchmark. So, that means it will perform well on real-world data, right? Right?
I don’t know.
Narrow AI

Performs well on a specific benchmark

- Highly customized for narrow tasks
- Hard to deal with unseen situations
- Struggles with under-specified inputs

Training

Testing

Testing

Collected dataset

i.i.d.

Train

Evaluation

⇒ 97% Acc.
Narrow AI

Performs well **on a specific benchmark**

- Highly customized for narrow tasks
- Hard to deal with unseen situations
- Struggles with under-specified inputs

General AI

Performs well **in the real world (in the wild)**

- Applicable to a wide range of tasks
- Generalizes well to novel settings
- Can handle noisy/ambiguous inputs

---

**Training**

![Training](image)

**Testing**

![Testing](image)

![Testing](image)

![Testing](image)

![Testing](image)
**Narrow AI**

Performs well **on a specific benchmark**

- Highly customized for narrow tasks
- Hard to deal with unseen situations
- Struggles with under-specified inputs

**General AI**

Performs well **in the real world (in the wild)**

- Applicable to a wide range of tasks
- Generalizes well to novel settings
- Can handle noisy/ambiguous inputs

---

**Collected dataset**

Train

\[ \text{i.i.d.} \rightarrow 97\% \text{ Acc.} \]

Evaluation

---

**Generalize to unseen cases**

- Wikipedia
- News
- Books

---

**Robust to perturbations**

- When is the time change?
- Do you mean **when is the time change?**

---

**Training/data efficiency**

---

**Trustworthy**

- 😊 (100 years later...)
  - When was Tokyo 2020 Olympics?
  - July 2021
- 😞 What?? Why???
Narrow AI
Performs well **on a specific benchmark**
- Highly customized for narrow tasks
- Hard to deal with unseen situations
- Struggles with under-specified inputs

General AI
Performs well **in the real world (in the wild)**
- Applicable to a wide range of tasks
- Generalizes well to novel settings
- Can handle noisy/ambiguous inputs

Commonsense Reasoning!
Why teaching machines common sense?

The human-like ability to understand and generate everyday scenarios (situations, events)

Image source: WikiHow
Why teaching machines common sense?

Common sense -> desirable *inductive bias* for machines to generalize to real-world settings.

Image source: WikiHow
Solving a Commonsense Reasoning Dataset

**Goal:** Perform well on a test set?

**Commonsense Question Answering**

![Graph showing accuracy over time for Commonsense QA models]

**Paper With Code:** CommonsenseQA 1.1
Solving a Commonsense Reasoning Dataset

Goal: Perform well on a test set?

Commonsense Question Answering

- **discriminative** (closed-ended) reasoning

Paper With Code: CommonsenseQA 1.1
Solving a Commonsense Reasoning Dataset

Goal: Perform well on a test set?

Commonsense Question Answering

- discriminative (closed-ended) reasoning
- logically robust to linguistic variations

Apples and oranges grow on trees
Oranges and apples grow on trees
Fruits grow on trees
Apples and oranges grow on plants
Trees grow on apples
Apples and trees grow on oranges

Paper With Code: CommonsenseQA 1.1
Solving a Commonsense Reasoning Dataset

**Goal:** Perform well on a test set?

**Commonsense Question Answering**

- **Discriminative** (closed-ended) reasoning
- **Not logically robust** to linguistic variations
- **Not** quickly adapt to unseen tasks

**Paper With Code:** CommonsenseQA 1.1

![Diagram showing accuracy over time with different methods and model names](image)
This talk
This talk

• New ways of formulating CSR challenges:
  • Capable of **open-ended** reasoning

In the school play, Robin played a hero in the struggle to the death with the angry villain.

Q: How would others feel afterwards?
A: (a) sorry for the villain  (b) hope Robin will win  (c) like Robin should lose
This talk

• New ways of formulating CSR challenges:
  • Capable of open-ended reasoning
  • Reason in a logically consistent manner

Apples and oranges grow on trees
Oranges and apples grow on trees
Fruits grow on trees
Apples and oranges grow on plants
Trees grow on apples
Apples and trees grow on oranges
This talk

• **New ways of formulating CSR challenges:**
  - Capable of *open-ended* reasoning
  - Reason in a *logically consistent* manner
  - Better at *cross-task generalization*

Directly Learning Calculus...

Machine: ❓❓❓
Baby: ❓❓❓

Learning Math, History, Geography, Chemistry, ...
in high school

Learning Calculus in undergrad
Student: *Yes I can do it!*
CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning

Bill Yuchen Lin†  Wangchunshu Zhou†  Ming Shen†  Pei Zhou†
Chandra Bhagavatula‡  Yejin Choi‡‡  Xiang Ren‡‡
†University of Southern California  ‡Allen Institute for Artificial Intelligence
‡Paul G. Allen School of Computer Science & Engineering, University of Washington

(Findings of EMNLP 2020)
What is CommonGen?

- Most current tasks for machine commonsense focus on **discriminative** reasoning.
  - CommonsenseQA, SWAG.

- Humans not only use **commonsense knowledge** for understanding text, but also for **generating sentences**.

---

Concept-Set: a collection of objects/actions.
- dog, frisbee, catch, throw

**Generative Commonsense Reasoning**

Expected Output: everyday scenarios covering all given concepts.
- A dog leaps to catch a thrown frisbee.
- The dog catches the frisbee when the boy throws it.
- A man throws away his dog’s favorite frisbee expecting him to catch it in the air.

Input:
- A set of common concepts (actions & objects)

Output:
- A sentence that **describes an everyday scenario** the given concepts.
Why is generative CSR hard?

(1) Relational knowledge are latent and compositional.

Underlying Relational Commonsense Knowledge
(exercise, HasSubEvent, releasing energy)
(rope, UsedFor, tying something)
(releasing energy, HasPrerequisite, motion)
(wave, IsA, motion) ; (rope, UsedFor, waving)
The motion costs more energy if ropes are tied to a wall.

Relational Reasoning for Generation
A woman in a gym exercises by waving ropes tied to a wall.

<table>
<thead>
<tr>
<th>Category</th>
<th>Relations</th>
<th>1-hop</th>
<th>2-hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial knowledge</td>
<td>AtLocation, LocatedNear</td>
<td>9.40%</td>
<td>39.31%</td>
</tr>
<tr>
<td>Object properties</td>
<td>UsedFor, CapableOf, PartOf, ReceivesAction, MadeOf, FormOf, HasProperty, HasA</td>
<td>9.60%</td>
<td>44.04%</td>
</tr>
<tr>
<td>Human behaviors</td>
<td>CausesDesire, MotivatedBy, Desires, NotDesires, Manner</td>
<td>4.60%</td>
<td>19.59%</td>
</tr>
<tr>
<td>Temporal knowledge</td>
<td>Subevent, Prerequisite, First/Last-Subevent</td>
<td>1.50%</td>
<td>24.03%</td>
</tr>
<tr>
<td>General</td>
<td>RelatedTo, Synonym, DistinctFrom, Isa, HasContext, SimilarTo</td>
<td>74.89%</td>
<td>69.65%</td>
</tr>
</tbody>
</table>
Why is generative CSR hard?

(2) Compositional Generalization for unseen concept compounds.

\[ x_1 = \{ \text{apple, bag, put} \} \quad \text{Training} \]
\[ y_1 = \text{a girl puts an apple in her bag} \]
\[ x_2 = \{ \text{apple, tree, pick} \} \]
\[ y_2 = \text{a man picks some apples from a tree} \]
\[ x_3 = \{ \text{apple, basket, wash} \} \]
\[ y_3 = \text{a boy takes an apple from a basket and washes it.} \]

Compositional Generalization

\[ x = \{ \text{pear, basket, pick, put, tree} \}, \quad y = ? \]

Reference: “a girl picks some pear from a tree and put them in her basket.”

→ Unseen Concept in Training
Case Study

Concept-Set: \{ hand, sink, wash, soap \}

[brNN-CopyNet]: a hand works in the sink.

[MeanPooling-CopyNet]: the hand of a sink being washed up

[ConstLeven]: a hand strikes a sink to wash from his soap.

[GPT-2]: hands washing soap on the sink.

[BERT-Gen]: a woman washes her hands with a sink of soaps.

[UniLM]: hands washing soap in the sink

[BART]: a man is washing his hands in a sink with soap and washing them with hand soap.

[T5]: hand washed with soap in a sink.

1. A girl is washing her hands with soap in the bathroom sink.
2. I will wash each hand thoroughly with soap while at the sink.
3. The child washed his hands in the sink with soap.
4. A woman washes her hands with hand soap in a sink.
5. The girl uses soap to wash her hands at the sink.
Experimental Results

<table>
<thead>
<tr>
<th>Model \ Metrics</th>
<th>ROUGE-2/L</th>
<th>BLEU-3/4</th>
<th>METEOR</th>
<th>CIDEr</th>
<th>SPICE</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>bRNN-CopyNet (Gu et al., 2016)</td>
<td>7.61</td>
<td>27.79</td>
<td>10.70</td>
<td>5.70</td>
<td>15.80</td>
<td>4.79</td>
</tr>
<tr>
<td>Trans-CopyNet</td>
<td>8.78</td>
<td>28.08</td>
<td>11.90</td>
<td>7.10</td>
<td>15.50</td>
<td>4.61</td>
</tr>
<tr>
<td>MeanPooling-CopyNet</td>
<td>9.66</td>
<td>31.14</td>
<td>10.70</td>
<td>6.10</td>
<td>16.40</td>
<td>5.06</td>
</tr>
<tr>
<td>LevenTrans. (Gu et al., 2019)</td>
<td>10.58</td>
<td>32.23</td>
<td>19.70</td>
<td>11.60</td>
<td>20.10</td>
<td>7.54</td>
</tr>
<tr>
<td>ConstLeven. (Susanto et al., 2020)</td>
<td>11.82</td>
<td>33.04</td>
<td>18.90</td>
<td>10.10</td>
<td>24.20</td>
<td>10.51</td>
</tr>
<tr>
<td>GPT-2 (Radford et al., 2019)</td>
<td>17.18</td>
<td>39.28</td>
<td>30.70</td>
<td>21.10</td>
<td>26.20</td>
<td>12.15</td>
</tr>
<tr>
<td>BERT-Gen (Bao et al., 2020)</td>
<td>18.05</td>
<td>40.49</td>
<td>30.40</td>
<td>21.10</td>
<td>27.30</td>
<td>12.49</td>
</tr>
<tr>
<td>UniLM (Dong et al., 2019)</td>
<td>21.48</td>
<td>43.87</td>
<td>38.30</td>
<td>27.70</td>
<td>29.70</td>
<td>14.85</td>
</tr>
<tr>
<td>UniLM-v2 (Bao et al., 2020)</td>
<td>18.24</td>
<td>40.62</td>
<td>31.30</td>
<td>22.10</td>
<td>28.10</td>
<td>13.10</td>
</tr>
<tr>
<td>BART (Lewis et al., 2019)</td>
<td>22.23</td>
<td>41.98</td>
<td>36.30</td>
<td>26.30</td>
<td>30.90</td>
<td>13.92</td>
</tr>
<tr>
<td>T5-Base (Raffel et al., 2019)</td>
<td>14.57</td>
<td>34.55</td>
<td>26.00</td>
<td>16.40</td>
<td>23.00</td>
<td>9.16</td>
</tr>
<tr>
<td>T5-Large (Raffel et al., 2019)</td>
<td>22.01</td>
<td>42.97</td>
<td>39.00</td>
<td>28.60</td>
<td>30.10</td>
<td>14.96</td>
</tr>
</tbody>
</table>

Our analysis shows that SPICE has the best correlation with human judgments
<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>BLEU-4</th>
<th>CIDEr</th>
<th>SPICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KFCNet</td>
<td>43.619</td>
<td>18.845</td>
<td>33.911</td>
</tr>
<tr>
<td></td>
<td>MSRA and Microsoft Ads</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Email Paper (EMNLP'21)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>KGR^4</td>
<td>42.818</td>
<td>18.423</td>
<td>33.564</td>
</tr>
<tr>
<td></td>
<td>Anonymous (under review)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Email Document (placeholder)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>KFC (v1)</td>
<td>42.453</td>
<td>18.376</td>
<td>33.277</td>
</tr>
<tr>
<td></td>
<td>MSRA and Microsoft Ads</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Email Paper (EMNLP'21)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>R^3-BART</td>
<td>41.954</td>
<td>17.706</td>
<td>32.981</td>
</tr>
<tr>
<td></td>
<td>Anonymous (under review)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Email Document (placeholder)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>WittGEN + T5-large</td>
<td>38.233</td>
<td>18.036</td>
<td>31.682</td>
</tr>
<tr>
<td></td>
<td>Anonymous (under review)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T5-Base (Raffel et al., 2019)</td>
<td>14.57</td>
<td>34.55</td>
<td>26.00</td>
</tr>
<tr>
<td></td>
<td>T5-Large (Raffel et al., 2019)</td>
<td>22.01</td>
<td>42.97</td>
<td>39.00</td>
</tr>
<tr>
<td></td>
<td>Human Performance</td>
<td>48.88</td>
<td>63.79</td>
<td>48.20</td>
</tr>
</tbody>
</table>
Q: What can help alleviate global warming?

**Open-Ended CSR**
Input: a question only

A large text corpus of commonsense facts

*Carbon dioxide* is the major greenhouse gas contributing to *global warming*.

*Trees* remove *carbon dioxide* from the atmosphere through photosynthesis.

renewable energy, *tree*, solar battery, ...

Output: a ranked list of concepts as answers.

**Multiple-Choice/Closed CSR**

Input: a question + a few choices

A) air conditioner  B) fossil fuel  C) renewable energy  D) carbon dioxide

Can machines learn to reason without answer candidates?

(Lin et al., with Google Research, NAACL 2021)  https://open-csr.github.io/
Smooth communication requires common sense

Text Message:
“I’m going to perform in front of thousands tomorrow...”

Explicit Knowledge:
Friend is going to perform in front of many people tomorrow

Commonsense Axiom:
Performing in front of people can cause anxiety
Smooth communication requires common sense

Text Message:
“I’m going to perform in front of thousands tomorrow…”

Explicit Knowledge:
Friend is going to perform in front of many people tomorrow

Commonsense Axiom:
Performing in front of people can cause anxiety
Smooth communication requires common sense

Text Message:
“I’m going to perform in front of thousands tomorrow…”

Explicit Knowledge:
Friend is going to perform in front of many people tomorrow

Commonsense Axiom:
Performing in front of people can cause anxiety
Smooth communication requires common sense

Explicit Knowledge:
Friend is going to perform in front of many people tomorrow

Commonsense Axiom:
Performing in front of people can cause anxiety

Text Message:
“I’m going to perform in front of thousands tomorrow...”

Inference Made:
My friend might be anxious, let me try to calm them

Text Message:
“Deep breaths, you’ll do great!”
Smooth communication requires common sense.

**Text Message:**
“I’m going to perform in front of thousands tomorrow…”

**Explicit Knowledge:**
Friend is going to perform in front of many people tomorrow

**Commonsense Axiom:**
Performing in front of people can cause anxiety

**Inference Made:**
My friend might be anxious, let me try to calm them

**Linguistically-Variated Statements of the same Commonsense Axiom**
- A person performing in front of people might be nervous
- People performing in front of people find it harder to be relaxed
- It can be hard for someone to be calm when they’re about to perform
Two key challenges

Inference making requires *implicit* commonsense reasoning

Humans fluidly adapt to *diverse* linguistic expressions
RICA: Evaluating Robust Inference Capabilities Based on Commonsense Axioms

Pei Zhou, Rahul Khanna, Seyeon Lee, Bill Yuchen Lin, Daniel Ho, Jay Pujara, Xiang Ren

EMNLP 2021
The RICA Challenge

Define logical primitives

Mine common sense

Represent commonsense in logic

Create commonsense statements that can be used to probe language models

Perturb and convert logic to text

Results: random guessing, heavy bias, and not robust
Results: random guessing, heavy bias, and not robust

- More data helps on human-verified set

Human Performance: 91.7%
Cross-task generalization in NLP

- Humans can learn a new task **efficiently** with only few examples, by leveraging their knowledge obtained when learning prior tasks.

- We refer to this ability as **cross-task generalization**.

- How such ability can be **acquired**, and further **applied** to build better few-shot learners across **diverse NLP tasks**.
CrossFit 🏋️‍♂️: A Few-shot Learning Challenge for Cross-task Generalization

Qinyuan Ye  Bill Yuchen Lin  Xiang Ren

University of Southern California - Information Sciences Institution
INK Lab @ USC-ISI
inklab.usc.edu

(Ye et al., EMNLP 2021)
Problem Setting

Prevalent Pipeline

- Large-scale Pre-training
- *Downstream Fine-tuning*

In our CrossFit Setting

- Large-scale Pre-training
- *Upstream Learning on a set of seen tasks*
- *Downstream Fine-tuning on an unseen target task*
Problem Setting

- To instantiate different settings in **CrossFit** and facilitate in-depth analysis.
- We present **NLP Few-shot Gym 🏋️‍♂️**, a repository of 160 diverse few-shot NLP tasks.
- We introduce 8 different seen/unseen tasks partitions of these few-shot tasks.

<table>
<thead>
<tr>
<th>No.</th>
<th>Shorthand</th>
<th>( T_{\text{train}} )</th>
<th>( T_{\text{dev}} )</th>
<th>( T_{\text{test}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Random</td>
<td>120</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>2.1</td>
<td>45cls</td>
<td>45 cls.</td>
<td>10 cls.</td>
<td>10 cls.</td>
</tr>
<tr>
<td>2.2</td>
<td>23cls+22non-cls</td>
<td>23 cls. + 22 non-cls</td>
<td>10 cls.</td>
<td>10 cls.</td>
</tr>
<tr>
<td>2.3</td>
<td>45non-cls</td>
<td>45 non-cls.</td>
<td>10 cls.</td>
<td>10 cls.</td>
</tr>
<tr>
<td>3.1</td>
<td>Held-out-NLI</td>
<td>57 non-NLI cls.</td>
<td>/</td>
<td>8 NLI</td>
</tr>
<tr>
<td>3.2</td>
<td>Held-out-Para</td>
<td>61 non-Paraphrase cls.</td>
<td>/</td>
<td>4 Para. Iden.</td>
</tr>
<tr>
<td>4.1</td>
<td>Held-out-MRC</td>
<td>42 non-MRC QA</td>
<td>/</td>
<td>9 MRC</td>
</tr>
<tr>
<td>4.2</td>
<td>Held-out-MCQA</td>
<td>29 non-MC QA</td>
<td>/</td>
<td>22 MC QA</td>
</tr>
</tbody>
</table>
Key Findings

- Q1. Does upstream learning help cross-task generalization?
Key Findings

- Q1. Does upstream learning help cross-task generalization?
  - We tried applying **multi-task learning** and **meta-learning** methods during the upstream learning stage.

  *Yes!* These methods do help pre-trained LMs to acquired cross-task generalization.
Key Findings

- Q2. “Well-rounded” or “specialized”? How to select tasks during upstream learning?

- Controlled experiments by fixing the downstream tasks to be 10 classification tasks.

- The upstream tasks are
  - 100% classification tasks
  - 50% classification + 50% non-classification tasks
  - 100% non-classification tasks

- Classification tasks and non-classification tasks seem to be equivalently helpful.

- Our understanding of tasks may not align with how models learn transferable skills.
Takeaways

Solving a Commonsense Reasoning Dataset

Goal: Perform well on a test set

Commonsense Question Answering

Paper With Code: CommonsenseQA 1.1

Solving Commonsense Reasoning

Goal: Satisfy the real-world needs

Generalize to unseen cases

Wikipedia News Books

Robust to perturbations

Do you mean when is the time change?

Trustworthy

(100 years later...) When was Tokyo 2020 Olympics?

July 2021

What??? Why???

And more...