Commonsense Knowledge Representation in NLP

Xiang Lorraine Li
PhD Student at UMass Amherst,
Advised by Prof Andrew McCallum
Common sense is sound practical judgement

- **Concerning everyday matters**
- **Basic ability to perceive, understand, and judge**
- **Shared by ("common to") nearly all people.** ----Wikipedia

**Disclaimer:** we mean, roughly, what a typical five year old knows about the world, including fundamental categories like time and space, and specific domains such as physical objects and substances; plants, animals, and other natural entities; humans, their psychology, and their interactions; and society at large. We will not attempt to be precise about this, but let us indicate roughly which issues we are considering and which we are ignoring. Obviously, this body of knowledge in fact depends on place, time, culture, social standing, personal characteristics (e.g. unusual cognitive or physical abilities or disabilities), schooling, perhaps on language. We ignore all that; without embarrassment, we have in mind a 21st-century, first-world, urban, with an appropriate level of schooling.

Commonsense: Essential in AI [1]

- Natural Language Processing
  - Machine Translation: reading and hearing.
  - Smart Home Assistant

Logical Formalization of Commonsense

In order for a program to be capable of learning something it must first be capable of being told it — John McCarthy [1]

Programs with Common Sense was probably the first paper on logical AI, i.e. AI in which logic is the method of representing information in computer memory and not just the subject matter of the program. The paper was given in the Teddington Conference on the Mechanization of Thought Processes in December 1958 and printed in the proceedings of that conference. It may also be the first paper to propose common sense reasoning ability as the key to AI.

1. First, we have a predicate “at”. “at(x, y)” is a formalization of “is at”. Under this heading we have the premises

   \[ \text{at}(I, \text{desk}) \quad (1) \]
   \[ \text{at}(\text{desk}, \text{home}) \quad (2) \]
   \[ \text{at}(\text{car}, \text{home}) \quad (3) \]
   \[ \text{at}(\text{home}, \text{county}) \quad (4) \]
   \[ \text{at}(\text{airport}, \text{county}) \quad (5) \]
Logical Formalization of Commonsense

**Deductive**

- There are three types of non-logical symbols: Constants, functions, and predicates. A constant symbol denotes an individual entity; a function denotes a mapping; and a predicate symbol denotes a relation.
- The logical symbols include the Boolean operators: the two quantifiers ∀ (for all) and ∃ (there exists); and variable symbols. Optionally, the equality sign = may also be included.

**Propositional Logic**

- Modal sentences are generally closer to a natural language expression of the fact.
- Modal sentences may be easier for a knowledge engineer to use.
- Propositional modal logics — that is, logics that have modal operators but no explicit quantifiers — are often both expressive enough for the purpose at hand and reasonably tractable, or at least decidable (Yardi, 1998).

**First-Order Logic**

**Modal Logic**

**Plausible Reasoning**

- Non-monotonic Logic
- Probabilistic Logic
- Fuzzy Logic

---

Logical Formalization of Commonsense

But … it did not end up well.

Past failures (in 70s – 80s) are inconclusive

-- weak computing power

-- not much data

-- not as strong computational models

-- not ideal conceptualization / representations

“I was told not to speak the word commonsense…” — Yejin Choi [1]
Outline

• Commonsense Knowledge.

• Learn the Right Representation.

• Commonsense Knowledge in Pre-trained LMs.

• Benchmark Datasets for Evaluation.
Vector Representation
Vector Representation

❌ Region ❌ Asymmetry
Gaussian Representation

✓ Region ✓ Asymmetry

[Vilnis and McCallum 2014]
Gaussian Representation

✓ Region  ✓ Asymmetry
Gaussian Representation

✓ Region  ✓ Asymmetry  ✓ Disjointness
Gaussian Representation

✓ Region ✓ Asymmetry ✓ Disjointness ❌ Closed under intersection
**Cone Representation**

- ✓ Region
- ✓ Asymmetry
- ✗ Disjointness
- ✓ Closed under intersection
Cone Representation

✅ Region  ✅ Asymmetry  ❌ Disjointness  ✅ Closed under intersection

[Vendrov, Kiros, Fidler and Urtasun 2015]
[Lai and Hockenmaier 2017]
Cone Representation

✓ Region   ✓ Asymmetry   ❌ Disjointness   ✓ Closed under intersection

[Vendrov, Kiros, Fidler and Urtasun 2015]
[Lai and Hockenmaier 2017]
Cone Representation

✓ Region  ✓ Asymmetry  ❌ Disjointness  ✓ Closed under intersection

\[ p(\text{rabbit}) = 0.12 \]

Unit Space

[p(\text{herbivore})]

[p(\text{mammal})]

[p(\text{deer})]

[p(\text{rabbit} \mid \text{deer})]
Cone Representation

✓ Region ✓ Asymmetry ❌ Disjointness ✓ Closed under intersection

Universe

p(herbivore)
p(mammal)
p(rabbit) = 0.12

p(rabbit | deer) = 0.4

p(rabbit)
p(deer)

p(mammal)

[Vendrov, Kiros, Fidler and Urtasun 2015]
[Lai and Hockenmaier 2017]
Cone Representation

✓ Region  ✓ Asymmetry  ❌ Disjointness  ✓ Closed under intersection

\[ p(\text{rabbit}) = 0.12 \]

\[ p(\text{deer}) \]

\[ p(\text{herbivore}) \]

\[ p(\text{mammal}) \]

\[ p(\text{rabbit} | \text{deer}) = 0.4 \]
Box Representation

✓ Region  ✓ Asymmetry  ✓ Disjointness  ✓ Closed under intersection

Unit Box
Box Representation

✓ Region   ✓ Asymmetry   ✓ Disjointness   ✓ Closed under intersection

[Vilnis, Li, Murty and McCallum 2018]
Box Representation

Common Sense

man relax

be on bench

Sit down

Unit Box
Box Training Loss

Hard Box (ACL 2018)  Smoothed Box (ICLR 2019)  Gumbel Box (NeurIPS 2021)


Sentence examples

- Flickr dataset is an entailment dataset containing 45 million image captions.

- Examples

| x              | p(x)   | y              | p(y)    | p(x|y)  |
|----------------|--------|----------------|---------|---------|
| person walk    | 0.11516| blond woman walk down sidewalk | 1.6E-04 | 1.0     |
| person wear clothing | 0.43036| adult dance on floor        | 3.9E-04 | 0.9     |
| man play percussion instrument | 0.00347| drummer         | 3.4E-03 | 0.51    |
| man wear jacket | 0.03077| snow on ground   | 5.1E-04 | 0.31    |
| in basement    | 4.3E-04| hold instrument  | 5.9E-03 | 0.0067  |

Outline

- Commonsense Knowledge.

- Learn the Right Representation.

- Commonsense Knowledge in Pre-trained LMs.

- Benchmark Datasets for Evaluation.
Do Pre-trained LMs **Already** Capture Commonsense Knowledge?

OpenAI’s new language generator GPT-3 is shockingly good—and completely mindless

A robot wrote this entire article. Are you scared yet, human?

We asked GPT-3, OpenAI’s powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace.

For more about GPT-3 and how this essay was written and edited, please read our editor’s note below.
Do Pre-trained LMs **Already** Capture Commonsense Knowledge?

Commonsense KB relations $\Rightarrow$ Natural language template $\Rightarrow$ Using LMs to query / score

- **LAMA**: Petroni et al. (EMNLP 2019)

  - ConceptNet, mining from Wikipedia
  - Hand-crafted templates scored by GPT2
  - BERT

  - Performs worse than supervised methods on ConceptNet but is more likely to generalize to different domains.

- **Feldman et al.** (EMNLP 2019)

  - ConceptNet and Wikidata
  - Hand-crafted templates
  - ELMo / BERT

  - BERT performs well but all models perform poorly on many-to-many relations.

---

**Figure 1:** Querying knowledge bases (KB) and language models (LM) for factual knowledge.

**Table 1:** Example of generating candidate sentences. Several enumerated sentences for the triple (musician, CapableOf, play musical instrument). The sentence with the highest log-likelihood according to a pretrained language model is selected.
Does the prompt matter?

- Yes! It matters! AutoPrompt (Shin et al., EMNLP 2020)
- Generating gradient guided prompt.

<table>
<thead>
<tr>
<th>Prompt Type</th>
<th>Original MRR</th>
<th>Original P@10</th>
<th>Original P@1</th>
<th>T-REx MRR</th>
<th>T-REx P@10</th>
<th>T-REx P@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAMA</td>
<td>40.27</td>
<td>59.49</td>
<td>31.10</td>
<td>35.79</td>
<td>54.29</td>
<td>26.38</td>
</tr>
<tr>
<td>LPAQA (Top1)</td>
<td>43.57</td>
<td>62.03</td>
<td>34.10</td>
<td>39.86</td>
<td>57.27</td>
<td>31.16</td>
</tr>
<tr>
<td>AutoPrompt 5 Tokens</td>
<td>53.06</td>
<td>72.17</td>
<td>42.94</td>
<td>54.42</td>
<td>70.80</td>
<td>45.40</td>
</tr>
<tr>
<td>AutoPrompt 7 Tokens</td>
<td>53.89</td>
<td>73.93</td>
<td>43.34</td>
<td>54.89</td>
<td>72.02</td>
<td>45.57</td>
</tr>
</tbody>
</table>
Properties of Concepts (Weir et al., 2020)

1. Do pre-trained LM correctly **distinguish concepts associated with a given set of properties**?

2. Can pre-trained LMs be used to **list the properties associated with given concepts**?

Adopted from ACL Commonsense tutorial: Knowledge in LMs.
Do Pre-trained LMs **Already** Capture Commonsense Knowledge?

Properties of Concepts (Weir et al., 2020)

1. Do pre-trained LM correctly **distinguish concepts** associated with a given set of properties?

   • A ___ has fur.

   • A ___ has fur, is big, and has claws.

   • A ___ has fur, is big, and has claws, has teeth, is an animal, eats, is brown…

Adopted from ACL Commonsense tutorial: Knowledge in LMs.
Do Pre-trained LMs **Already** Capture Commonsense Knowledge?

Properties of Concepts (Weir et al., 2020)

1. Do pre-trained LM correctly **distinguish concepts** associated with a given set of properties?

- Good performance, RoBERTa > BERT
- Perceptual (e.g. visual) < non-perceptual (e.g. encyclopaedic or functional).
- Highly-ranked incorrect answers typically apply to a subset of properties.

Adopted from ACL Commonsense tutorial: Knowledge in LMs.
Do Pre-trained LMs *Already* Capture Commonsense Knowledge?

Properties of Concepts (Weir et al., 2020)

1. Can pre-trained LMs be used to **list the properties associated with given concepts**?

   - Low correlation with human elicited properties, but coherent and mostly “verifiable by humans”

<table>
<thead>
<tr>
<th>Context</th>
<th>Human Response</th>
<th>Human PF</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Everyone knows that) a bear has ____ .</td>
<td>fur 27</td>
<td></td>
</tr>
<tr>
<td></td>
<td>claws 15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>teeth 11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cubs 7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>paws 7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>teeth</td>
<td>.36</td>
</tr>
<tr>
<td></td>
<td>claws</td>
<td>.18</td>
</tr>
<tr>
<td></td>
<td>eyes</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>ears</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>horns</td>
<td>.02</td>
</tr>
</tbody>
</table>

Adopted from ACL Commonsense tutorial: Knowledge in LMs.
Can we trust knowledge from LMs?

- LMs also generate fictitious facts!

Distributionally-related:

Barack’s Wife Hillary:
Using Knowledge Graphs for Fact-Aware Language Modeling

Robert L. Logan IV*  Nelson F. Liu†  Matthew E. Peters‡
Matt Gardner§  Sameer Singh*

* University of California, Irvine, CA, USA
† University of Washington, Seattle, WA, USA
§ Allen Institute for Artificial Intelligence, Seattle, WA, USA

{nlogon, sameer}@uci.edu, {mattg, matthewp}@allenai.org, nflu@cs.washington.edu

Syntactically-similar:

Negated and Misprimed Probes for Pretrained Language Models:
Birds Can Talk, But Cannot Fly

Nora Kassner, Hinrich Schütze
Center for Information and Language Processing (CIS)
LMU Munich, Germany
kassner@cis.lmu.de

Adopted from ACL Commonsense tutorial: Knowledge in LMs.
Outline

• Commonsense Knowledge.

• Learn the Right Representation.

• Commonsense Knowledge in Pre-trained LMs.

• Benchmark Datasets for Evaluation.
Besides probing the model using commonsense knowledge bases, are there standard commonsense benchmark datasets to evaluate the model?

- Question Answering.
- Natural Language Inference.
- Coreference Resolution.
- ...
**Benchmark Evaluation Dataset**

<table>
<thead>
<tr>
<th>Task Type</th>
<th>Domain</th>
<th>Example</th>
<th>Gap to Human Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi Choice QA</td>
<td>Grounded commonsense.</td>
<td>A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She a) rinses the bucket off with soap and blow dries the dog's head. b) uses a hose to keep it from getting soapy. c) gets the dog wet, then it runs away again. d) gets into the bath tub with the dog.</td>
<td>95.6% - 93.85% = 1.75%</td>
</tr>
<tr>
<td>Multi Choice Selection</td>
<td>Abductive Reasoning</td>
<td>Example HellaSwag Question</td>
<td>92.90% - 89.70% = 3.2%</td>
</tr>
<tr>
<td>Multi Choice QA</td>
<td>Reading Comprehension</td>
<td></td>
<td>94% - 91.79% = 2.3%</td>
</tr>
</tbody>
</table>

**Example HellaSwag Question**

A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She a) rinses the bucket off with soap and blow dries the dog's head. b) uses a hose to keep it from getting soapy. c) gets the dog wet, then it runs away again. d) gets into the bath tub with the dog. = 4.77%

https://leaderboard.allenai.org/
**Benchmark Evaluation Dataset**

<table>
<thead>
<tr>
<th>Task Type</th>
<th>Domain</th>
<th>Example</th>
<th>Gap to Human Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi Choice QA</td>
<td>Grounded commonsense.</td>
<td></td>
<td>95.6% - 93.85% = 1.75%</td>
</tr>
<tr>
<td>Multi Choice Selection</td>
<td>Abductive Reasoning</td>
<td></td>
<td>92.90% - 89.70% = 3.2%</td>
</tr>
<tr>
<td>Multi Choice QA</td>
<td>Reading Comprehension</td>
<td></td>
<td>94% - 91.79% = 2.3%</td>
</tr>
<tr>
<td>Multi Choice QA</td>
<td>Naive physical reasoning</td>
<td></td>
<td>94.9% - 90.13% = 4.77%</td>
</tr>
<tr>
<td>Multi Choice QA</td>
<td>Social commonsense</td>
<td></td>
<td>88.1% - 83.15% = 4.95%</td>
</tr>
<tr>
<td>Multi Choice Selection</td>
<td>Coreference resolution</td>
<td></td>
<td>94% - 91.28% = 2.72%</td>
</tr>
<tr>
<td>Multi Choice QA</td>
<td>Vision &amp; Language</td>
<td></td>
<td>85% - 77.79% = 2.3%</td>
</tr>
</tbody>
</table>

**Generative Evaluation**

[https://leaderboard.allenai.org/](https://leaderboard.allenai.org/)
Benchmark Evaluation Dataset

 kad ProtoQA (EMNLP 2020): dataset that captures prototypical situation.
 ✓ Multiple correct answers.
 ✓ Scores for each answer.

Name something that people usually do before they leave the house for work?
ProtoQA (EMNLP 2020)

✨ Generative Evaluation

✓ Evaluate **multiple** correct answers generative by the model.
✓ Reward models with **correct ranking of answer list**.
✓ Reward models with **higher coverage of answer list**.

Name something that people usually do before they leave for work.

![Diagram showing the evaluation process and reward calculation for a question with answer clusters and reward matrix.]
Results

Numbers reported are percentage of perfect score, i.e. answering with a list with an element from each answer cluster in decreasing order would yield 100.
Summary

• Commonsense Knowledge.

• Learn the Right Representation.

• Commonsense Knowledge in Pre-trained LMs.

• Benchmark Datasets for Evaluation.

Table 1: Example of generating candidate sentences. Several enumerated sentences for the triple (musician, CapableOf, play musical instrument). The sentence with the highest log-likelihood according to a pretrained language model is selected.

<table>
<thead>
<tr>
<th>Candidate Sentence $S_i$</th>
<th>$\log p(S_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;musician can playing musical instrument&quot;</td>
<td>-5.7</td>
</tr>
<tr>
<td>&quot;musician can be play musical instrument&quot;</td>
<td>-4.9</td>
</tr>
<tr>
<td>&quot;musician often play musical instrument&quot;</td>
<td>-5.5</td>
</tr>
<tr>
<td>&quot;a musician can play a musical instrument&quot;</td>
<td>-2.9</td>
</tr>
</tbody>
</table>
Thanks to all the collaborators!

Dan Le  Shikhar Murty  Shib Sankar Dasgupta  Dongxu Zhang  Rajarshi Das

Luke Vilnis  Michael Boratko  Tim O’Gorman  Andrew McCallum