Rensselaer

why not change the world?®

Exploring and Analyzing Machine Commonsense Benchmarks

Henrique Santos, Minor Gordon, Zhicheng Liang, Gretchen Forbush, Deborah L. McGuinness

February 8th 2021

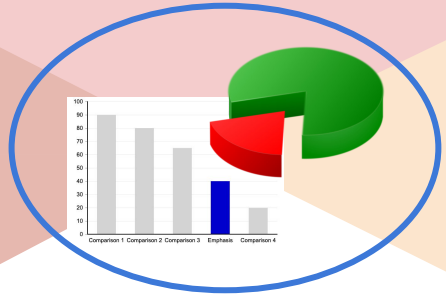
Commonsense Knowledge Graphs Workshop @AAAI-21

Task creators

- What are the dimensions currently not being sufficiently evaluated, potentially indicating the need of new tasks?
- What metadata is relevant to include in my task?

PIs, PMs, grad students

- What dimensions is a benchmark (or a benchmark subset) evaluating?



System developers

- How can I assess my system/model with regards to evaluation dimensions? In what it is doing well or poorly?
- What datasets can I use to train my system in specific dimensions

The Machine Common Sense Ecosystem

How to support insights into tasks?

- A common vocabulary for data and metadata about elements of the MCS ecosystem
- Data model based on dimensions
- Focus on what is already publicly available - a *living literature review*
- Display as a unified dashboard with browsing capabilities

Evaluation dimensions

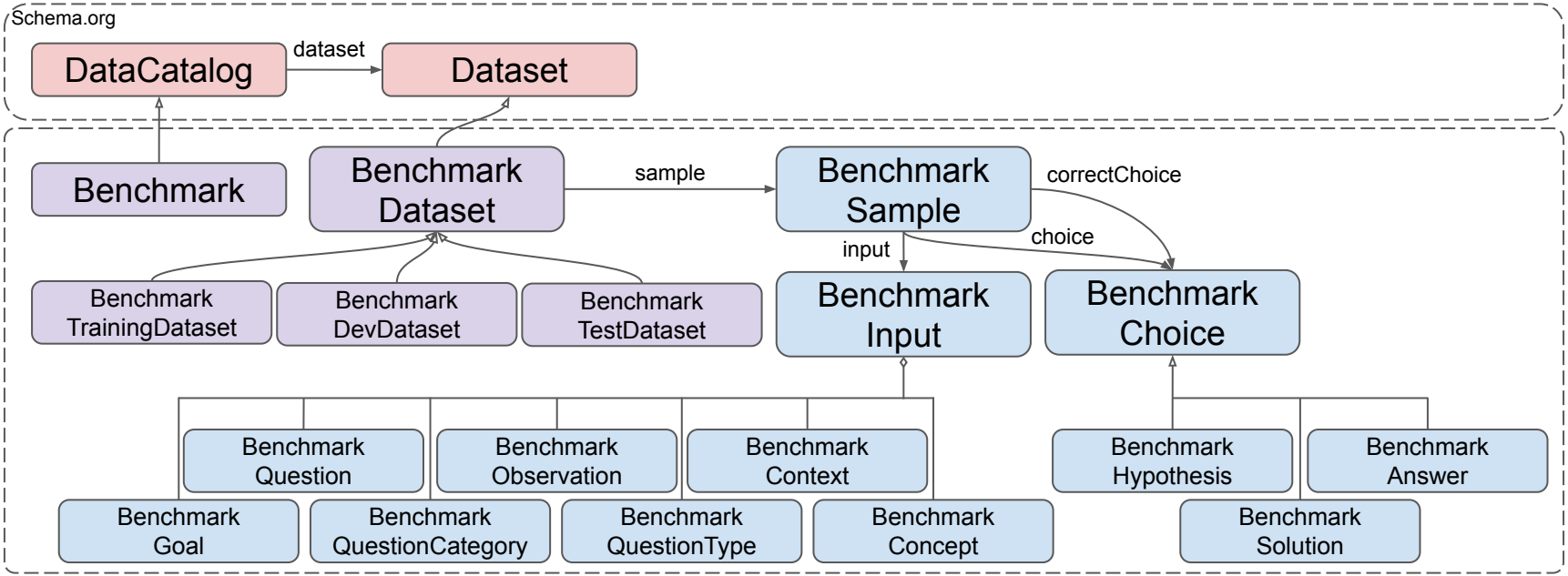
Dimension	Year-1
question type	multiple-choice
topic	various (social, physical, events)
reasoning type	mixed
training data size	full, or slightly reduced
interpretability	black-box, paths
inference level	mixed
model universality	different models

* from USC/ISI slides - September 2020 DARPA meeting (with modifications)

DARPA MCS Year 1 Benchmarks

Benchmark	Constructs	Choice type
aNLI	- Observations - Hypothesis	- Multiple choice
CommonsenseQA	- Questions - Answer choices	- Multiple choice
Cosmos QA	- Context - Questions - Answer choices	- Multiple choice
CycIC	- Questions - Answer choices - Fill in the blank - Categories	- Multiple choice - True/false
HellaSwag	- Context - Ending choices	- Multiple choice
Physical IQa	- Goals - Solution choices	- Multiple choice
Social IQa	- Context - Questions - Answer choices	- Multiple choice

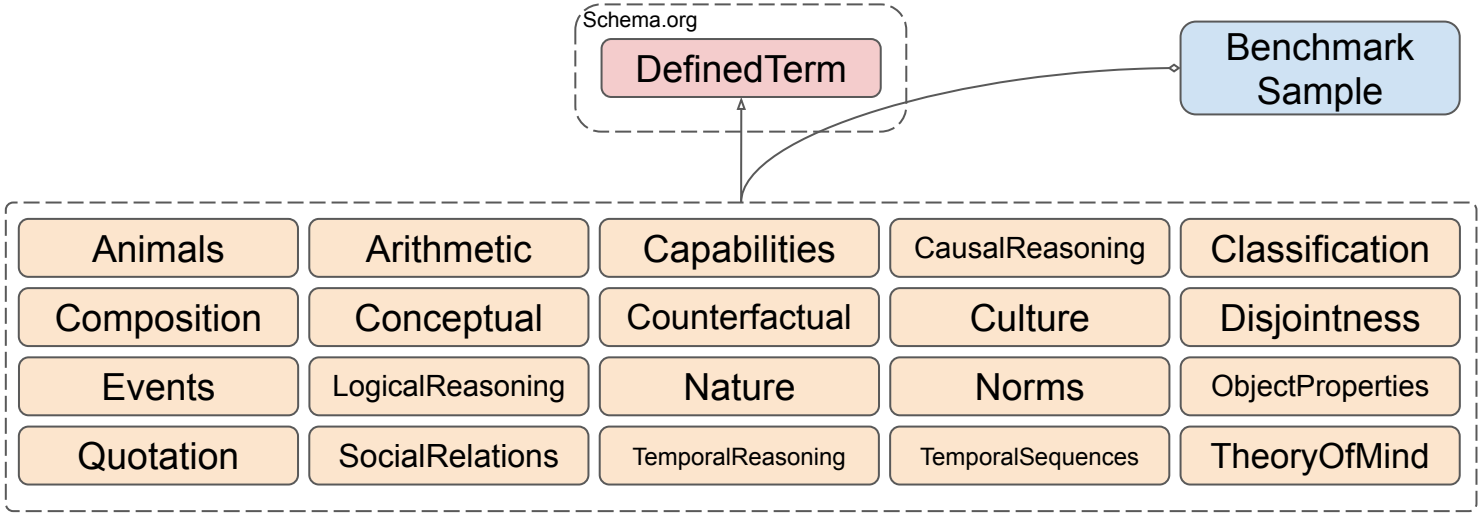
Machine Commonsense Benchmark Ontology



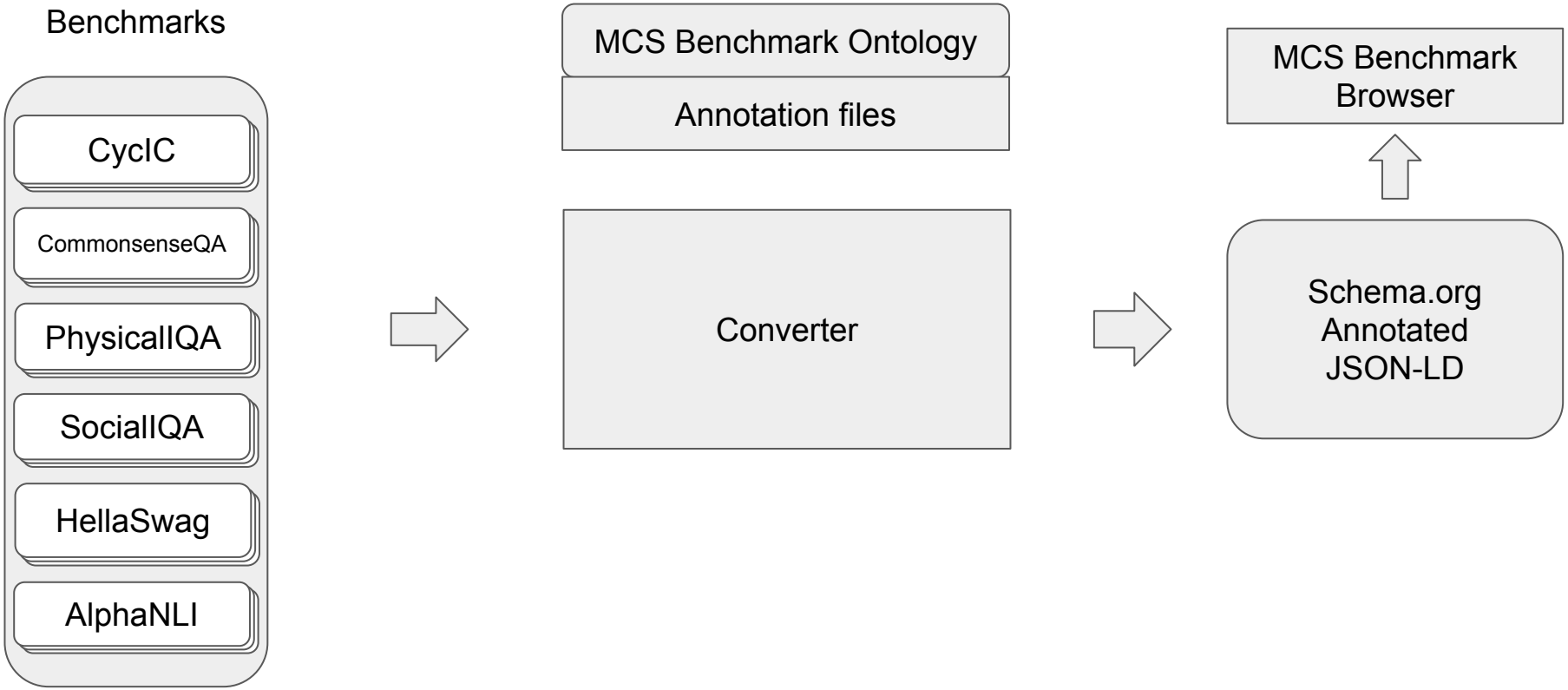
https://github.com/tetherless-world/mcs-ontology

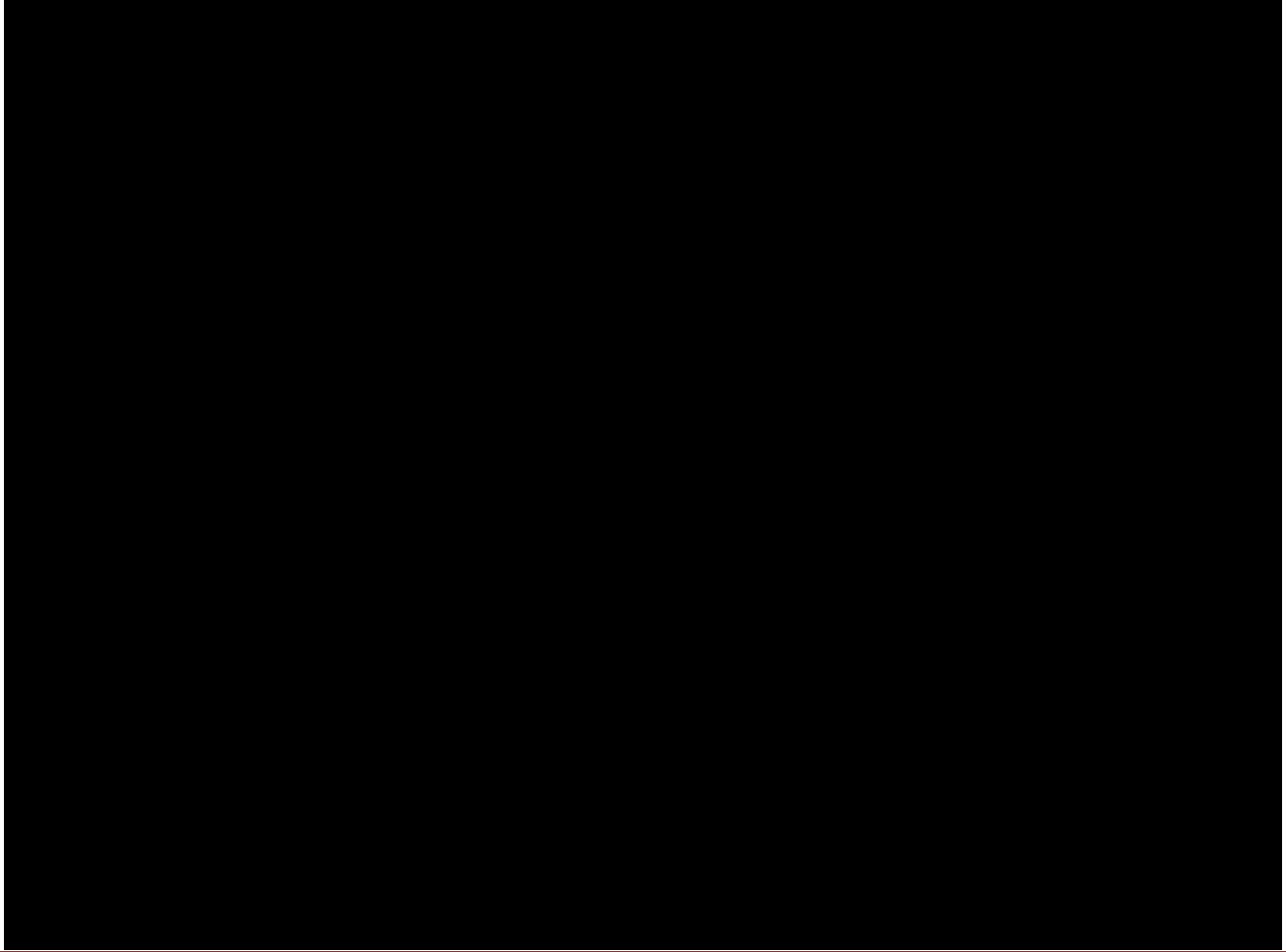
Machine Commonsense Benchmark Ontology

Benchmark categorization



https://github.com/tetherless-world/mcs-ontology


Benchmarks
CycIC
CommonsenseQA
PhysicalIQA
SocialIQA
HellaSwag
AlphaNLI
MCS Benchmark Ontology
Annotation files
Converter
MCS Benchmark Browser
Schema.org Annotated JSON-LD



Next steps

- Incorporation of open-ended benchmarks
- Refinement and deeper representation of commonsense theories
  - Currently researching commonsense principles (or theories - Time, Space, Causality ...) to further characterize tasks

Takeaways

- Benchmark metadata is crucial to allow precise description of tasks
- There is a lack of consensus of what would this metadata be

Our vision

A formalized metadata layer for machine common sense that can comprehensively describe tasks/benchmarks in terms of commonsense psychology and its theories, and capable of effectively characterizing the human thinking process.

Questions?

Exploring and Analyzing Machine Commonsense Benchmarks

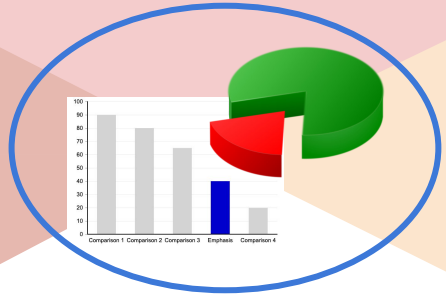
Henrique Santos, Minor Gordon, Zhicheng Liang, Gretchen Forbush, Deborah L. McGuinness
oliveh@rpi.edu
February 8th 2021

Task creators

- What are the dimensions currently not being sufficiently evaluated, potentially indicating the need of new tasks?
- What am I (not) aiming to challenge systems on with this task?
- What metadata is relevant to include in my task?

PIs, PMs, grad students

- What dimensions is a benchmark (or a benchmark subset) evaluating?
- What questions (or question types) are used to challenge systems in a specific dimension (e.g. temporal reasoning)



System developers

- How can I assess my system/model with regards to evaluation dimensions? In what it is doing well or poorly?
- What are the new or current tasks on a specific dimension I could apply my system to?
- What datasets can I use to train my system in specific dimensions

The Machine Common Sense Ecosystem

What is currently possible

- Integration of multi-choice benchmarks
- Categorization of benchmarks and samples (questions) in standardized terms
- Browsing and visualization of benchmarks contents

```
{
 "@context": "https://tetherless-world.github.io/mcs-ontology/utils/context.jsonld",
 "@id": "SocialIQA-[line_number]",
 "@type": "BenchmarkSample",
 "includedInDataset": "SocialIQA/train",
 "input": [
   {
     "@type": "BenchmarkContext",
     "name": "Context",
     "text": "Aubrey offered tribute to the gods. They did this out of reverence.",
     "position": 0
   },
   {
     "@type": "BenchmarkQuestion",
     "name": "Question",
     "text": "How would Others feel as a result?",
     "position": 1
   }
 ],
 "choice": [
   {
     "@type": "BenchmarkAnswer",
     "text": "they were different",
     "position": 1
   },
   {
     "@type": "BenchmarkAnswer",
     "text": "religious and spiritual",
     "position": 2
   },
   {
     "@type": "BenchmarkAnswer",
     "text": "they were unique",
     "position": 3
   }
 ],
 "correctChoice": 3
}
```

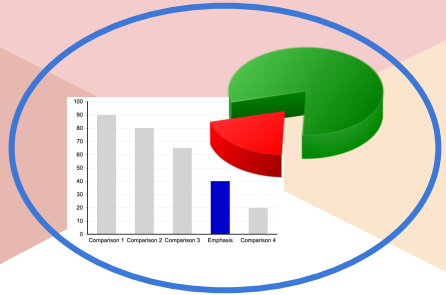

Example question in the JSON-LD format

Task creators

- What are the dimensions currently not being sufficiently evaluated, potentially indicating the need of new tasks?
- What am I (not) aiming to challenge systems on with this task?
- **What metadata is relevant to include in my task?**

PIs, PMs, grad students

- **What dimensions is a benchmark (or a benchmark subset) evaluating?**
- **What questions (or question types) are used to challenge systems in a specific dimension (e.g. temporal reasoning)**



MCS Ecosystem

System developers

- How can I assess my system/model with regards to evaluation dimensions? In what it is doing well or poorly?
- What are the new or current tasks on a specific dimension I could apply my system to?

Selected Motivating Questions

- Can we understand weaknesses of QA systems by analyzing failed cases?
- Is it possible to classify diverse commonsense benchmark samples - questions - in common categories following an agreed criteria?
- How can tools provide stakeholders with insights into the current state-of-the-art of commonsense ecosystems and their sub-tasks?

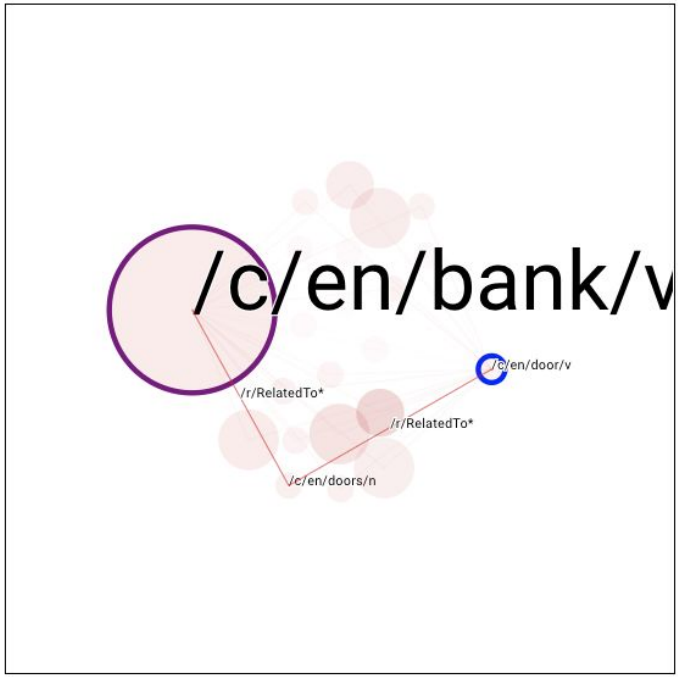
Benchmarks

- Composed by training, development, and test datasets
- Test datasets are usually not available for systems' developers
- May have specific thematics (e.g. social interactions, "how-to's")
  - Or just be general purpose
- They vary in structure
  - Question-answers
  - Observation-hypothesis
  - Goals-solutions
  - …
- They provide different levels of information a system can use
  - Question type, required type of reasoning, source concept ...

kagnet submission answer choice analyses

Answer Paths for **bank** More edges Fewer edges High score Low score

Score: 0.818 20 paths



Start node: /c/en/door/v End node: /c/en/bank/v

Score	Number of nodes	Path
0.07347899675369263	3	/c/en/door/v /r/RelatedTo* /c/en/doors/n /r/RelatedTo* /c/en/bank/v
0.0719200000166893	3	/c/en/door/v /r/RelatedTo /c/en/houses/v /r/RelatedTo* /c/en/bank/v
0.07078800350427628	3	/c/en/door/v /r/RelatedTo /c/en/home/v /r/RelatedTo* /c/en/bank/v
0.06894899904727936	3	/c/en/door/v /r/RelatedTo /c/en/getting/v /r/RelatedTo* /c/en/bank/v
0.06700500100851059	3	/c/en/door/v /r/RelatedTo* /c/en/time_lock/n /r/RelatedTo /c/en/bank/v

DARPA's vision

- Machine common sense as a computational service
- ...that learns from experience, like a child
- ...that learns from reading the Web, like a research librarian

Supporting DARPA's vision

- Ability to actively expose the commonsense service (or pool of services) to diverse tasks, including new and specialized tasks
  - To learn, to probe
- Ability to score systems across benchmarks, not only by task
- Ability to analyze systems' at sample level (question level), not only by the overall score
- Ability to selectively probe systems with specific types of samples (questions, observations, etc.)