Exploring and Analyzing Machine Commonsense Benchmarks

Henrique Santos, Minor Gordon, Zhicheng Liang, Gretchen Forbush, Deborah L. McGuinness
Rensselaer Polytechnic Institute
110 8th St, Troy, New York 12180
{oliveh,gordom6,liangz4,forbug}@rpi.edu, dlm@cs.rpi.edu

Abstract

Commonsense question-answering (QA) tasks, in the form of benchmarks, are constantly being introduced for challenging and comparing commonsense QA systems. The benchmarks provide question sets that systems' developers can use to train and test new models before submitting their implementations to official leaderboards. Although these tasks are created to evaluate systems in identified dimensions (e.g. topic, reasoning type), this metadata is limited and largely presented in an unstructured format or completely not present. Because machine common sense is a fast-paced field, the problem of fully assessing current benchmarks and systems with regards to these evaluation dimensions is aggravated. We argue that the lack of a common vocabulary for aligning these approaches' metadata limits researchers in their efforts to understand systems' deficiencies and in making effective choices for future tasks. In this paper, we first discuss this MCS ecosystem in terms of its elements and their metadata. Then, we present how we are supporting the assessment of approaches by initially focusing on commonsense benchmarks. We describe our initial MCS Benchmark Ontology, an extensible common vocabulary that formalizes benchmark metadata, and showcase how it is supporting the development of a Benchmark tool that enables benchmark exploration and analysis.

1 Introduction

Machine commonsense (MCS) benchmarks have arisen as a way to challenge AI question answering systems by presenting a set of natural language questions that require these systems to solve tasks that involve what some perceive as commonsense knowledge. The MCS community is constantly introducing diversified tasks and allowing question-answering (QA) systems' developers to submit their systems to official leaderboards. These leaderboards have emerged to act as hubs for hosting benchmarks and supporting infrastructure that accepts submissions of systems that then get scored against these tasks. This MCS ecosystem consists of benchmarks and tasks, QA systems and models, structured commonsense knowledge, leaderboards, scientific publications, and the people and institutions behind these efforts. Although surveys exist (Storks, Gao, and Chai 2020), they quickly become outdated due to the field's quick pace.

The DARPA's Machine Common Sense Program[1] is an ongoing effort that has been exploring the boundaries of the commonsense question-answering state-of-the-art, with the ultimate goal of creating a commonsense computational service that can solve diverse challenges. In support of this, the program is promoting the creation of new, improved, or specialized commonsense tasks from within the project as well as adopting challenging benchmarks from outside the project. In addition, it is supporting the augmentation and incorporation of structured commonsense knowledge in QA systems. Under the program, commonsense approaches are being categorized in terms of *evaluation dimensions* that span across tasks and systems. They aim to capture relevant aspects from these elements, allowing approaches to be classified into common categories, supporting insights into the available tasks, and informing stakeholders about the commonsense aspects they explore. In addition, they support the program in making decisions for future tasks. The currently identified dimensions include *question type* (e.g, multiple-choice, open-ended), *topic* (e.g. social, events), and *reasoning type* (e.g. temporal, spatial).

In this context, it is currently a challenge to efficiently perform analysis or extract insights into the dimensions of these composing elements that can support potential use-cases from the community. As an example, a principal investigator could ask, *"what are the tasks that challenge systems in temporal reasoning?"* As benchmarks currently convey limited metadata, even more challenging is how to scale these dimensions to assess not only tasks as a whole, but as well as to assess each question. Benchmarks can contain questions that are varied, each of which focusing on specific dimensions.

In this paper, we demonstrate how we are tackling the problem of assessing MCS approaches by initially focusing on commonsense benchmarks. We introduce our MCS Benchmark ontology and describe how it is being applied in support of a Benchmark tool to enable the exploration and analysis of multiple commonsense tasks. The ontology specifies concepts that describe the datasets, making use of the Schema.org simplified taxonomy. Refined classes allow the description of several constructs that compose benchmarks. The ontology acts as a formalization of a subset of the

[1] https://www.darpa.mil/program/machine-common-sense

currently identified evaluation dimensions within DARPA's MCS program.

2 MCS Benchmark ontology

The MCS Benchmark ontology aims to formalize the diverse benchmark metadata in a common vocabulary. In support of the ontology development, we have surveyed several state-of-the-art benchmarks to understand their constructs, and a summary is shown on Table 1. They were selected based on DARPA's adoption of these tasks for Year 1 of the MCS program.

Benchmark	Constructs	Question type
aNLI (Bhagavatula et al. 2019)	- Observations - Hypothesis	- Multiple choice
CommonsenseQA (Talmor et al. 2019)	- Questions - Answer choices	- Multiple choice
CosmosQA (Huang et al. 2019)	- Context - Questions - Answer choices	- Multiple choice
CycIC[2]	- Questions - Answer choices - Fill in the blank - Categories	- Multiple choice - True/false
HellaSwag (Zellers et al. 2019)	- Context - Ending choices	- Multiple choice
Physical IQa (Bisk et al. 2020)	- Goals - Solution choices	- Multiple choice
Social IQa (Sap et al. 2019)	- Context - Questions - Answer choices	- Multiple choice

Table 1: Benchmarks and their identified constructs.

While all of the benchmarks provide "multiple-choice" questions (with CycIC providing a binary kind of multiple-choice: true/false), the benchmarks diverge on the kinds of constructs they include. CommonsenseQA provides a standard question/answer format. CosmosQA and Social IQa include the context on top of that. HellaSwag also provides context, but it requires systems to choose the more appropriate ending for the context, instead of answering a question. aNLI provides observations (usually a scene or a setting) and asks systems to explain the reasons for that observation to happen in the form of a hypothesis. Physical IQa provides goals (or objectives), alongside possible solutions to achieve them. CycIC is an interesting case as it provides an increased level of metadata for each sample. These include categories, which contain information about what kind of reasoning is needed, and/or question classification into common types.

The MCS Benchmark ontology provides a common modeling for these diverse constructs, as seen in Figure 1. Each entry in a benchmark dataset is defined as a `BenchmarkSample`. Each sample is then composed of one or more instances of `BenchmarkInput`. Input is considered anything that a benchmark provides that systems can use. These are the constructs identified in Table 1. Each sample also contains the possible choices, represented by the

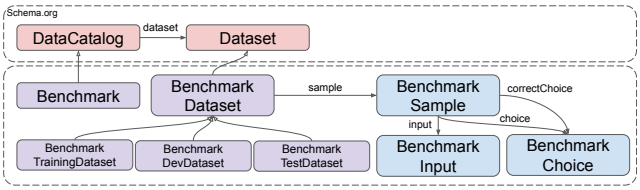

Figure 1: Part of the MCS Benchmark Ontology

`BenchmarkChoice` class. The correct choice, which is used to train/verify proposed models, is linked to the sample by the `correctChoice` property.

Listing 1: Benchmark sample in JSON-LD

```
{
  "@context": "https://.../context.jsonld",
  "@id": "SocialIQa-37",
  "@type": "BenchmarkSample",
  "includedInDataset": "SocialIQa/train",
  "input": [
    {
      "@id": "SocialIQa-37-input-0",
      "@type": "BenchmarkContext",
      "text": "Skylar returned early in the
          evening after a night and day of
          partying."
    },
    {
      "@id": "SocialIQa-37-input-1",
      "@type": "BenchmarkQuestion",
      "text": "How would you describe Skylar?"
    }
  ],
  "choice": [
    {
      "@id": "SocialIQa-37-choice-1",
      "@type": "BenchmarkAnswer",
      "text": "a party girl"
    },
    {
      "@id": "SocialIQa-37-choice-2",
      "@type": "BenchmarkAnswer",
      "text": "very shy"
    },
    {
      "@id": "SocialIQa-37-choice-3",
      "@type": "BenchmarkAnswer",
      "text": "exhausted"
    }
  ],
  "correctChoice": {
    "@id": "SocialIQa-37-choice-1"
  }
}
```

In Listing 1, we show an entry from the Social IQa benchmark represented using the MCS Benchmark ontology, in JSON-LD. The sample is composed of a list of inputs (a context and a question) and a list of choices (possible answers for the question). To assert the sample as part of either the training, development, or test dataset, the ontology defines a set of classes that are used to represent the benchmark datasets. The `includedInDataset` property links samples to instances of `BenchmarkTrainDataset`, `BenchmarkDevDataset`, and

BenchmarkTestDataset.

3 Evaluation: Supporting exploration and analysis of Benchmarks

The MCS Benchmark ontology is used in support of a prototype Benchmark tool that provides several features for interacting with benchmarks from multiple sources. The Benchmark tool allows users to explore and analyze benchmarks by leveraging the common modeling provided by the ontology. To enable this, we have implemented a converter that can receive the datasets that compose benchmarks as input and, using the terminology in the ontology, it outputs them in the common JSON-LD format, as seen in Listing 1. To further simplify the JSON-LD serialization, we have encapsulated many linked data constructs (including namespaces and property types) in a JSON-LD context file (i.e. `@context` key). This allows us to suppress verbose information in the JSON, while keeping the correct and concise expression of the model.

Figure 2 contains a screen of the tool displaying some of the CycIC questions alongside the available metadata. We provide a sample Benchmark tool usage of the ontology in Listing 2. The SPARQL query retrieves training samples of a specific benchmark by constraining the dataset to be of type `BenchmarkTrainDataset`. For each sample, their input texts and types are retrieved.

Listing 2: Querying training samples of a benchmark

```
SELECT ?sample ?input ?inputType WHERE {
 <task_uri> schema:dataset ?train .
 ?train rdf:type mcs:BenchmarkTrainDataset .
 ?train mcs:sample ?sample .
 ?sample mcs:input/schema:text ?input .
 ?sample mcs:input/rdf:type/rdfs:label ?inputType .
}
```

Listing 3 shows a SPARQL query that retrieves samples containing logical reasoning questions across different benchmarks. The ontology represents each of the identified input types, therefore it enables querying by a specific type, in this case, `BenchmarkQuestion`.

Listing 3: Querying question samples across benchmarks

```
SELECT ?sample ?question WHERE {
 ?sample rdf:type mcs:BenchmarkSample .
 ?sample mcs:input/rdf:type mcs:LogicalReasoning .
 ?sample mcs:input ?input .
 ?input rdf:type mcs:BenchmarkQuestion .
 ?input schema:text ?question .
}
```

4 Related work

Throughout the past decade, the challenge for Question Answering over Linked Data[3] (QALD) has been used as a way of promoting the development of question-answering systems capable of solving benchmarks, with an increased difficulty based on the growing availability of linked data on the web. In its latest edition, participants were required to integrate their entrant systems with the GERBIL QA (Usbeck

Text	Type	Categories
Question: A couple saw something while out for a walk. The thing they saw was either a point arena mountain beaver or a ring. It was not a mammal. True or False: The thing was a ring.	True/False	logical reasoning
Question: Rosalyn is coming to your house for dinner. She is an adherent of the vegetarian diet program. Pick the food or drink you couldn't serve to Rosalyn.	Multiple Choice	norms
Question: In U.S. summer, Rob visits the graveyard every day. In U.S. autumn, Rob visits the park every day. Where will Rob go on July 27?	Multiple Choice	temporal reasoning temporal sequences

Figure 2: Benchmark tool displaying questions and associated metadata

et al. 2019) benchmark platform. GERBIL QA serves as an integration service for QA systems, supporting the fair comparison of systems through the use of unified metrics and integrated benchmark datasets. GERBIL QA represents question sets in a JSON-based format that serves as an interface that characterizes each question, including data types, entities, and keywords. This format is highly tailored to question answering over linked data, as it assists systems in building responses that comply with the expected format.

AI Collaboratory (Martínez-Plumed, Hernandez-Orallo, and Gomez 2020) has the objective of being a platform for the analysis and comparison of AI systems. Its scope goes beyond question answering, allowing submissions of systems that solve diverse and specialized AI tasks (e.g. link prediction, speech recognition, and more). Tasks are represented in an Entity-Relationship model without an in-depth formalization of tasks' metadata. AI2 Leaderboard,[4] maintained by the Allen Institute for AI, hosts many commonsense benchmarks and accepts submissions of systems. In this approach, benchmarks are stored as originally released alongside documentation (in natural language) that includes a description of the tasks, and the format of the datasets.

The AIDA Dashboard (Angioni et al. 2020) is an implementation that assists editors of scientific publishers in assessing conferences in Computer Science, with regards to some dimensions, including citations, topic, and similar conferences. It uses the Computer Science Ontology (CSO) to annotate research papers with a common vocabulary and leverages this annotation to provide visualizations that extract insights based on these dimensions. Although it focuses on a different domain, this work is closely related to ours as it relies on a metadata formalization as an ontology that aligns dimensions in support of the creation of dashboards.

To the best of our knowledge, none of the previous attempted to formalize and integrate metadata across the variety of commonsense tasks. Our Benchmark tool, supported by the MCS Benchmark ontology, aims to bridge this gap in machine commonsense, where the lack of metadata formalization is constraining the ability of further assessing tasks in terms of identified dimensions.

5 Conclusion

In a fast-paced field, such as machine common sense, we argue that it is essential to promote the formalization of

[3] http://qald.aksw.org

[4] https://leaderboard.allenai.org

metadata to support the development and analysis of evaluation metrics. Further, a common representation metadata schema can support sharing and communication of results that can be compared and contrasted more easily, and can thus help the MCS community understand more about what the benchmarks well suited to test and how different approaches and methods may compare in different contexts. We presented our evolving MCS Benchmark ontology that is aimed to support our Benchmark exploration and analysis tool. Existing platforms for benchmarks are largely task-specific and limited in terms of metadata formalization. In commonsense reasoning, where the ultimate goal is to have a commonsense service that can actively learn from new and specialized tasks, there is a need to support the incorporation of these diverse benchmarks, while allowing implementations to access them in a standardized way. We believe a centralized platform, where the MCS community can obtain trends and comparisons across tasks, will greatly support this objective.

We are expanding our review of commonsense benchmarks, including open-ended benchmarks (benchmarks that are not multiple-choice, requiring systems to elaborate their answer in the form of a natural language sentence, e.g. CommonGen (Lin et al. 2020)), and we are expanding the ontology to support these. In parallel, we started to work towards the formalization of the evaluation dimensions that are related to QA systems, with a focus on *interpretability*. To this extent, we are analyzing QA systems and identifying information that can be extracted during their execution and represented in the ontology, as a way of assessing a system. The current efforts involve exposing knowledge graph paths that are calculated by systems, such as KagNet (Lin et al. 2019), that leverage such a structure in their pipelines. We expect that this kind of information will help the systems' developers to further analyze their implementations across benchmarks.

We want to support additional use-cases. As an example, to help systems' developers with insights into the tasks that can be used to decide with which datasets may exist that may make sense for them to use for training. Similarly for tasks' creators, we want to allow them to understand their own tasks and how to augment them in order to increase coverage along certain dimensions. In addition, we want to provide them with information about dimensions not currently sufficiently evaluated, potentially indicating the need for new tasks.

We are making the MCS Benchmark ontology open and we hope that it will be adopted and potentially extended by the MCS community. It can be used to describe, compare, and explore diverse aspects of benchmarks. The MCS Benchmark ontology powers the Benchmark tool, which is a step towards a working platform for integrating diverse commonsense benchmarks, supporting a streamlined process of incorporating new or specialized tasks that challenge question-answering systems and models. It acts as a foundation for enabling implementations to access the tasks' data in a standardized way. It enables the description of benchmarks in a common vocabulary, allowing users to analyze their content, comparing these with consistent metadata across benchmarks. This ontology is in active development and available at https://github.com/tetherless-world/mcs-ontology.

Acknowledgments

This work is funded through the DARPA MCS program award number N660011924033 to RPI under USC-ISI West.

References

Angioni, S.; Salatino, A.; Osborne, F.; Reforgiato Recupero, D.; and Motta, E. 2020. The AIDA Dashboard: Analysing Conferences with Semantic Technologies. In *ISWC 2020 Demos and Industry Tracks: From Novel Ideas to Industrial Practice*.

Bhagavatula, C.; Bras, R. L.; Malaviya, C.; Sakaguchi, K.; Holtzman, A.; Rashkin, H.; Downey, D.; Yih, W.-t.; and Choi, Y. 2019. Abductive Commonsense Reasoning. In *Eighth International Conference on Learning Representations*.

Bisk, Y.; Zellers, R.; Bras, R. L.; Gao, J.; and Choi, Y. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. *AAAI Conference on Artificial Intelligence* 34(05).

Huang, L.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Lin, B. Y.; Chen, X.; Chen, J.; and Ren, X. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Lin, B. Y.; Zhou, W.; Shen, M.; Zhou, P.; Bhagavatula, C.; Choi, Y.; and Ren, X. 2020. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. *Findings of EMNLP* .

Martínez-Plumed, F.; Hernandez-Orallo, J.; and Gomez, E. 2020. Tracking AI: The Capability is (Not) Near. In *24th European Conference on Artificial Intelligence*.

Sap, M.; Rashkin, H.; Chen, D.; Le Bras, R.; and Choi, Y. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Storks, S.; Gao, Q.; and Chai, J. Y. 2020. Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. *arXiv:1904.01172 [cs]* .

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Usbeck, R.; Röder, M.; Hoffmann, M.; Conrads, F.; Huthmann, J.; Ngonga-Ngomo, A.-C.; Demmler, C.; and Unger, C. 2019. Benchmarking question answering systems. *Semantic Web* 10(2).

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *57th Annual Meeting of the Association for Computational Linguistics*.