How Much Qualitative Reasoning is Required in Elementary School Science Test Questions?

Maxwell Crouse & Kenneth D. Forbus

Qualitative Reasoning Group, Northwestern University
2133 Sheridan Road, Evanston, IL, 60208, USA
MaxwellCrouse2020@u.northwestern.edu; forbus@northwestern.edu

Abstract

Understanding how to build cognitive systems with commonsense is a difficult problem. Since one goal of qualitative reasoning is to explain human mental models of the continuous world, hopefully qualitative representations and reasoning have a role to play. But how much of a role? Standardized tests used in education provide a potentially useful way to measure both how much qualitative knowledge is used in commonsense science, and to assess progress in qualitative representation and reasoning. This paper analyzes a small corpus of science tests from US classrooms and shows that QR techniques are central in answering 13% of them, and play a role in at least an additional 16%. We found that today's QR techniques suffice for standard QR questions, but integrating QR with broader knowledge about the world and automatically understanding the questions as expressed in language and pictures provide new research challenges.

Introduction

When children are learning about science, their initial education is qualitative in nature. It ties scientific concepts to everyday experiences, teaching them how to think about the world around them in terms of more fundamental ideas, including processes (e.g. evaporation, predation) and patterns (e.g. life cycles, food webs). Since these concepts are used in education, there are teaching materials that are accessible to children (and easier for natural language understanding systems to learn from) and standardized tests that measure knowledge in human-normed ways. For example, the New York State Board of Regents makes their exams publically available after they have been given, providing a corpus that supports research. Thus *commonsense science*, as it is sometimes called, provides an excellent frontier for research on qualitative reasoning, since it involves broad-ranging knowledge and multiple kinds of reasoning.

This is not a novel observation. Project Aristo (Clark et al. 2016) identified elementary school science as a productive research area for studying learning by reading and commonsense reasoning. The Science Learning and Teaching working group (which Forbus is part of) adopted such tests as the first phase in a longer research trajectory, with the long-term (2050) goal of AI systems that can help any person learn any area of science, at whatever level they are interested in. This effort is one of multiple efforts that, collectively, are being designed as a replacement for the Turing Test (Forbus, 2016).

That such tests require deeper knowledge can be seen from the recent Allen Institute Science Challenge on Kaggle[1], which used 8th grade science tests. The tag line was "Is your model smarter than an 8th grader?" The answer, for the 738 teams competing, was clearly no. The questions were limited to multiple-choice tests, without diagrams. The rules of the competition were such that no licensed data or software could be used, i.e. no resources from the Linguistic Data Consortium, nothing from Cyc, Watson, or any other system or data that could not be completely open-licensed. Thus the only techniques applied were off-the-shelf machine learning components (including deep learning) and statistical NLP. The best scores achieved on this challenge – which is only a subset of the types of questions on real exams – topped out at 60%[2]. This suggests that deeper knowledge is indeed needed to achieve 8th grade science literacy. Our analysis below argues further that QR is needed as part of that deeper knowledge.

This paper examines how useful qualitative reasoning might be in elementary school science tests. We focus on 4th grade examinations, since that is what Project Aristo has been examining. A prior study of such exams (Clark et al. 2013) provided a useful decomposition of question types,

[1] https://www.kaggle.com/c/the-allen-ai-science-challenge/
[2] Public presentations, Oren Etzioni, Peter Clark, AAAI 2016.

but did not take into account a qualitative reasoning perspective. Hence the questions we ask here are (1) what fraction of exam questions use qualitative representations? (2) How well do today's QR approaches handle the reasoning needed for such questions? After examining the contents of six Regents 4th grade exams, the answers so far are (1) qualitative knowledge is needed for at least 29% percent of exam questions and (2) the standard QR-related questions are naturally handled by existing qualitative reasoning techniques.

An Analysis of Science Tests

Much QR research has focused on specific scientific and engineering domains. By contrast, commonsense science is remarkable for its breadth – such tests cover physics, biology, chemistry, and other areas. Instead of a small vocabulary of structural elements (e.g. circuit components), the entire range of everyday objects is fair game. After all, the purpose of learning science in elementary school and middle school (grades 1-6 and grades 7-8th respectively, in the US) is to ground scientific ideas in a child's experience.

Some questions, such as Figure 1, look exactly like traditional QR scenarios. We call these *standard QR* questions. By viewing the flame as the source, the wire as the destination, and the contact surface with the flame as the path, any reasonable model of heat flow will predict that the temperature of the wire will rise. But translating that insight into heat travelling through the wire involves thinking of the wire itself as a kind of path, which makes the decoding of the language more subtle.

Some problems set up scenarios that are used in multiple questions. Here is an example:

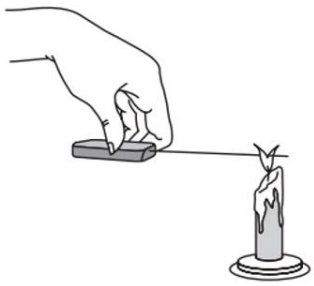
> *One hot, summer day it rained very heavily. After the rain, a plastic pan on a picnic table had 2 cm of rainwater in it. Four hours later, all of the rainwater in the pan was gone.*

One question asked about this scenario was which process caused the disappearance, given condensation, evaporation, precipitation, and erosion as choices. Examining the conditions and influences of these processes enables honing in quickly on the answer. Another question was, if the day was cool instead of hot, would the rainwater have disappeared slower, faster, or in the same amount of time? This is a classic comparative analysis question (Weld, 1986), and again well within the scope of today's QR systems.

Other types of questions require QR, but involve deeper visual reasoning, e.g. comparing which of two inclined planes it would be harder to push a weight up, or choosing among visual configurations as answers to a question posed. Prior research suggests that such problems can be handled via QR, but with additional complexities of visual reasoning, case-based reasoning, or both (e.g. Klenk et al. 2011; Chang et al. 2014). Hence we argue that, to fully capture human capabilities in commonsense science, we should expand our notion of domain theories to include both specific examples and knowledge of patterns of behavior. We call questions that make use of such knowledge *extended QR* questions, because answering them with off-the-shelf purely first-principles QR techniques might be doable, but would be a stretch.

Closely related are questions about patterns found in nature, e.g. food webs, the water cycle, and life cycles of different sorts of living creatures. Such questions are often accompanied by diagrams, showing for example the participants in a food chain or the stages in a life cycle. We refer to these as *pattern questions.* Once a pattern is introduced, some follow-on questions end up being standard QR questions. For example, questions about food webs often require performing comparative analysis on population size changes, taking into account predation. But other pattern questions simply involve placing states in a correct sequence, e.g. the phases of an animal's life cycle.

The drawing below shows a copper wire with a wooden handle being held in a flame.



After a few minutes, what will most likely happen?

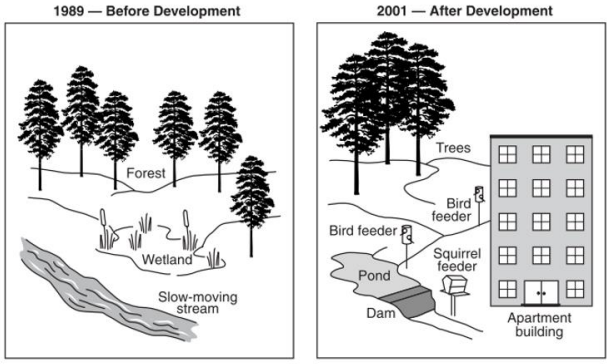
A. The light will change to electricity.
B. The heat will travel through the wire.
C. The flame will get brighter.
D. The flame will go out.

Figure 1: An example test question

While the picture in Figure 1 may help a child understand the problem better, the caption provides, in some sense, all that is needed to solve the problem. But in some problems a deeper understanding of diagrams is necessary to answer the question. Questions often involve decoding information from graphs, tables, and/or drawings of measurement instruments. For example, each exam typically has at least one question about graphs, which requires reading the graph and answering qualitative or quantitative questions about it (e.g. given a population graph, "How many times was there a decrease in the deer population from one year to the next […]?") Problems with pictures often involve recognition, e.g. the different animals in a food web, the different stages in a life cycle, weather icons on a map. Sometimes these

pictures have labels, when recognition would be too demanding, as in Figure 2. We call such questions *visual questions*. This problem is especially interesting because it requires integrating the scenario across two modalities, language and vision, and generating an answer, rather than selecting from multiple-choice answers.

A company bought land in 1989 to build apartments. The diagram labeled 1989 shows the land before the company built the apartments. The diagram labeled 2001 shows the same land after the company built the apartments.



Describe one positive way and one negative way that the organisms living in the area have been affected by the changes shown in the diagrams.

Figure 2: A multimodal scenario problem

Any question that does not fit in one of the above categories we classify as a *world knowledge* question. This is a grab-bag category, involving many different kinds of knowledge. For example, some kinds of questions involve properties of objects, e.g. which object from a list (wax crayon, plastic spoon, rubber eraser, iron nail) is the best conductor of electricity? These involve QR, in that conductivity can be thought of as a parameter – while binary in this case, a harder question would involve iron, tap water, and salt water. But many others involve knowledge about non-continuous aspects of the world. For example, "which characteristic can a human offspring inherit?", where the answers include facial scars, long hair, broken leg, and blue eyes. Another sub-category of questions concern function, e.g. "The functions of a plant's roots are to support the plant and", with "make food", "produce fruit", "take in water and nutrients", and "aid in germination" as alternatives.

While more fine-grained analyses of commonsense science questions are possible, this set of categories suffices to address the first of our two questions. To identify the degree to which QR is needed in commonsense science, we analyzed a corpus of six 4th grade science exams[3]. The results are shown in Table 1.

This analysis suggests that QR knowledge about continuous causality is a necessary part of doing well on the exam: the highest score a student could get would be 71% otherwise. On the other hand, QR is not sufficient to do well on the exam, as indicated by 71% of the questions not involving QR.

Type	# Problems	%
Standard QR	31	13%
Extended QR	38	16%
Patterns	36	15%
Visual	55	22%
World	85	35%

Table 1: Analysis of question types on science exams

Solving QR-based Problems

Now let us turn to the second question: Can current QR techniques solve the QR problems that arise in such science tests? To examine this question, we selected the set of 31 standard QR questions from the corpus of New York Regents exams. To factor out issues in natural language understanding, we hand-coded queries corresponding to each question. We used knowledge base contents from ResearchCyc[4], with our own extensions for qualitative, visual, and analogical reasoning and learning.

While our KB already had a substantial portion of the knowledge needed, some extensions were required. We used qualitative process theory (Forbus, 1984) to express the new domain knowledge. Specifically, we encoded 8 additional physical processes (precipitation, evaporation, erosion, freezing, melting, birth, death, growth) and 5 other model fragments (buoyancy, organism populations, standard gravity, predator/prey, friction, and magnetism), along with 6 new types of quantities (fluid level in a container, fluid displaced, heat produced, friction force applied against an object, magnetic force attracting an object, and roughness) and one ordinal relationship (smooth objects are less rough than rough objects). The rest of the QP domain theory came from previously existing knowledge. It consisted of 2 types of processes (boiling and heat flow) and 7 types of quantities (population size, mass, weight, volume, temperature, size, amount of a substance, and distance). Extending the domain theory required approximately two months of work.

[3] Specifically, the New York Regents science exams for 2004, 2005, 2006, 2009, 2010, and 2011. These and exams for other years and grade levels are available on their web site.

[4] http://www.cyc.com/platform/researchcyc/

```
(isa FreezingProcess QPProcessType)
(mfTypeParticipant FreezingProcess ?thing-freezing LiquidTangibleThing
                   focusOf)
(mfTypeParticipant FreezingProcess
  ?sub ChemicalCompoundTypeByChemicalSpecies substanceOf)
(mfTypeParticipantConstraint FreezingProcess
                             (substanceOfType ?thing-freezing ?sub))
(mfTypeParticipantConstraint FreezingProcess
                  (relationAllInstance freezingPoint ?sub ?f-temp))
(mfTypeCondition FreezingProcess
                 (qLessThan (TemperatureFn ?thing-freezing) ?f-temp))
(mfTypeBiconditionalConsequence FreezingProcess
  (hasQuantity ?self (SolidGenerationRateFn ?self)))
(mfTypeConsequence FreezingProcess
  (qGreaterThan (SolidGenerationRateFn ?self) 0))
(mfTypeConsequence FreezingProcess
  (qprop- (SolidGenerationRateFn ?self)
          ((QPQuantityFn Temperature) ?thing-freezing)))
(mfTypeConsequence FreezingProcess
  (i+ (AmountOfFn ?sub Solid-StateOfMatter ?thing-freezing)
      (SolidGenerationRateFn ?self)))
(mfTypeConsequence FreezingProcess
  (i- (AmountOfFn ?sub Liquid-StateOfMatter ?thing-freezing)
      (SolidGenerationRateFn ?self)))
```

Figure 3: Representation of the process of freezing

Figure 3 shows the description of the axioms for the process of freezing (`FreezingProcess`) as an example. That it is a type of process specified by QP theory is indicated by the isa statement placing it as a member of the collection `QPProcessType`, whose instances are members of QPProcess, e.g. a particular instance of freezing. Type-level predicates are used to define model fragment types. The participants are specified by `mfTypeParticipant`, e.g. here `FreezingProcess` has three participants, whose types are the third argument (e.g. `LiquidTangibleThing`, a pre-existing concept in Cyc), whose template variables are the second argument (e.g. `?thing-freezing`), and whose fourth argument is a role relation that is used to refer to this participant in axioms about instances (e.g. `solidOf`). `mfTypeCondition` expresses the conditions that must hold for an instance to be active. These are interpreted as conjunctions, although this process has only one, i.e. that the temperature of the thing freezing is less than its freezing point. The consequences are expressed via `mfTypeConsequence` and `mfTypeBiconditionalConsequence`, the latter for statements that can only be true when an instance is active. An example of such a constitutive relationship is the existence of a rate at which the process occurs, which does not make sense outside the process acting. The usual causal qualitative mathematics of QP theory appear in the consequences, e.g. `i+` and `i-` for direct influences (i.e. partial specifications of derivatives) and `qprop` and `qprop-`, for indirect influences (i.e. partial specifications of functional dependencies). Wherever possible, we link these descriptions into the Cyc ontology, e.g. `LiquidTangibleThing` comes from the Cyc ontology, so that axioms about them already in the knowledge base can provide leverage. Sometimes the Cyc ontology takes a slightly different perspective on the world. For example, the Cyc concept of Temperature concerns specific values for temperatures, e.g. `Hot` or `(DegreeCelsius 25)`. In QP theory, quantities are fluents, in that they are not values but conceptual entities whose value changes over time. We link the two notions via the logical function `QPQuantityFn`, a second-order function whose domain is Cyc quantities and whose range are functions denoting fluents, here `((QPQuantityFn Temperature) ?thing-freezing)` denotes the fluent representing the temperature of `?thing-freezing`.

Figure 4 provides an example of a model fragment, a description of an object floating in a fluid (`ObjectFloatingInFluid`). It is an instance of `ConceptualModelFragmentType`, that is, instances of this type of model fragment are conceptual knowledge about the situation. (Some types of model fragments indicate the existence of something, such as a contained fluid or population, those are instances of `PhysicalModelFragmentType`.) Note the multiple condition statements, which are interpreted conjunctively. `activeMF` is true when the model fragment instance which is its argument is active.

The queries to solve these problems were relatively straightforward applications of qualitative reasoning. For example, some problems describe a situation and ask what kind of process is involved in the change that is occurring in it. Performing model formulation on the situation and in-

```
(isa ObjectFloatingInFluid ConceptualModelFragmentType)
(mfTypeParticipant ObjectFloatingInFluid ?csolid SolidTangibleThing
                   solidOf)
(mfTypeParticipant ObjectFloatingInFluid ?cfluid FluidTangibleThing
                   fluidOf)
(mfTypeParticipant ObjectFloatingInFluid ?b-mf FluidDisplacement
                   displacementOf)
(mfTypeParticipantConstraint ObjectFloatingInFluid
                             (fluidOf ?b-mf ?cfluid))
(mfTypeParticipantConstraint ObjectFloatingInFluid
                             (contains-Underspecified ?cfluid ?csolid))
(mfTypeCondition ObjectFloatingInFluid (activeMF ?b-mf))
(mfTypeCondition ObjectFloatingInFluid
  (qLessThanOrEqualTo
    ((QPQuantityFn Weight) ?csolid)
    ((QPQuantityFn Weight) (FluidDisplacedFn ?b-mf))))
(mfTypeConsequence ObjectFloatingInFluid
  (qprop (FluidDisplacedFn ?b-mf) ((QPQuantityFn Weight) ?csolid)))
```

Figure 4: Representation of the model fragment describing a floating object

specting the instantiated processes provides a straightforward way to answer such questions. Sometimes chaining is needed, that is, searching through dependencies among model fragments. For example, the question "Which form of energy is needed to change water from a liquid to a gas?" with answers being "heat", "mechanical", "chemical", and "sound" requires a breadth-first search through model fragments, beginning with model fragments whose consequences involve direct influences on amounts of substances of different phases, a negative influence on the liquid version and a positive influence on that substance in the gas phase, and expanding on model fragments that are mentioned as conditions, in this case, heat flow. When questions involve comparisons, differential qualitative analysis (Weld, 1986, 1990) is used to determine the changes to quantities of interest that have occurred.

To provide a sense of how these problems are solved, let us return to the scenario presented earlier:

One hot, summer day it rained very heavily. After the rain, a plastic pan on a picnic table had 2 cm of rainwater in it. Four hours later, all of the rainwater in the pan was gone.

(Q7) Which process caused the rainwater in the pan to disappear as it sat outside in the hot air?

(Q8) If the day were cool instead of hot, the rainwater in the pan would have disappeared _____

Q7 is answered by constructing a qualitative model for the state of the scenario in which water was sitting in a pan and examining the influences on it (see Figure 5), to see which process is responsible for decreasing the amount of water. As Figure 5 illustrates, the model fragments are tied to the Cyc ontology, including the use of Cyc's `ScalarInterval` system for underspecified values (e.g. `Hot`), but which have ordinal relationships tied to other underspecified values in the same dimension (e.g. `Cool`). For Q8, an additional qualitative state is created to represent the cooler day, with everything the same except for that the temperatures of the air and rainwater are `Cool` instead of `Hot`. We use analogy to perform comparative analysis: the analogical mapping[5] provides information about how the two states correspond, in both their values and their causal structure. Figure 6 illustrates the correspondences computed between these two states. The system checks first to see if there is enough

```
(isa LiquidTangibleThing)
(substanceOfType rainwater014 Water)
(isa air435 GaseousTangibleThing)
;;From "the rainwater ... sat outside in the hot air"
(touches-Directly rainwater014 air435)
;; QPQuantityFn converts Cyc's value notion to QP's fluents
(qEqualTo ((QPQuantityFn Temperature)
air435) Hot)
(qEqualTo ((QPQuantityFn Temperature)
rainwater014) Hot)
```

Figure 5: Partial representation of Q7 evaporation scenario

information about the goal quantity to directly determine if it is different. (For example, if in a different question the system were asked about the temperature of the desert during the day (`Hot`) and during the night (`Cool`), the ordinal difference between these values would be sufficient to answer the question.) Otherwise, it looks for causal structure that specifies the goal quantity in terms of others, and recursively seeks their comparative values. Here, the aligned causal influences (`qprop` relations) linking the evaporation rate and temperature of each scenario enable the system to

[5] Mappings are computed using SME, the Structure-Mapping Engine (Falkenhainer et al. 1989; Forbus et al. in press).

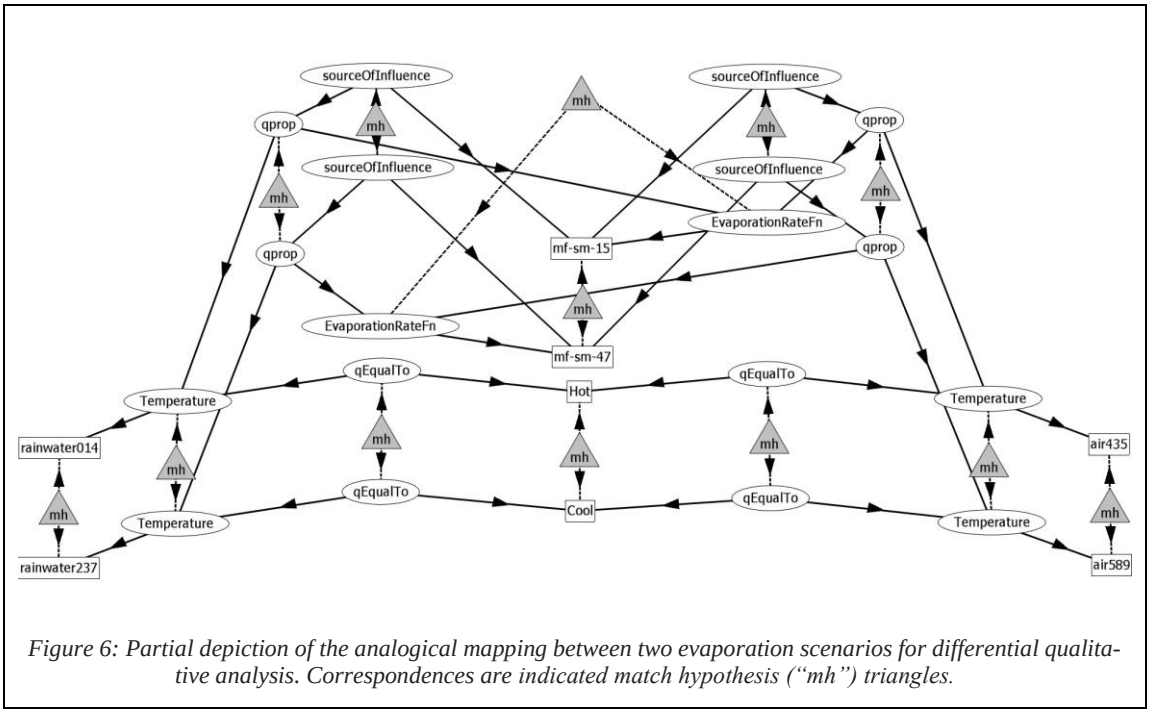


Figure 6: Partial depiction of the analogical mapping between two evaporation scenarios for differential qualitative analysis. Correspondences are indicated match hypothesis ("mh") triangles.

infer that, since the temperature is reduced, and the rate of evaporation depends on temperature, then the rate of water disappearance would be slower in the new scenario.

Using existing qualitative reasoning techniques, the system was able to solve all 31 standard QR problems. Less success was achieved on the extended QR problems. Of the few we tackled, none were solvable with strictly first-principles QR techniques. They require more knowledge of the everyday world. Consider again the problem shown in Figure 2. This problem requires inferring that there are fewer trees after construction than before construction. This is indicated schematically by there being fewer trees on the right, but also by the associated labels, e.g. "forest" versus "trees". Students must know that trees provide habitats for birds and squirrels, which are part of what helps determine the size of their populations, and hence that fewer trees means less habitat and hence a negative effect on population. On the other hand, adding feeders improves their food supply. (Whether or not this benefit outweighs the loss of food supply from habitat loss seems dubious, but nevertheless it is a positive influence, even if dominated by another factor.) Other examples involve richer interactions between dynamics and spatial knowledge (e.g. knowing that liquids take the shape of their container). Again, some forays into representing these ideas have been done before in QR, e.g. Kim's bounded stuff ontology (Kim 1993), but domain theories which tightly integrate qualitative dynamics and spatial representations are few and limited in coverage currently. Accumulating examples to reason from (e.g. Klenk et al 2011), plus more flexible multimodal interaction (e.g. Chang & Forbus, 2015) should be helpful for teaching systems the knowledge that they need to tackle problems like these.

Related Work

The most successful system thus far in answering elementary science exam questions is AI2's Aristo (Clark et al. 2016) which combines techniques from information retrieval, statistical NLP, and rule-based systems. The success of Aristo relied on both the ensemble of techniques and its ability to estimate which technique's answer is most likely to be correct. With its diverse set of techniques, Aristo achieved a score of 71.3% on a corpus of 129 Regents 4th grade non-diagram multiple-choice-only questions. In the analysis of its performance, five types of questions were identified as being challenging for Aristo to solve. The question types were comparison questions, simple arithmetic reasoning, complex inference, structured questions, and story questions. In our application of QR techniques to Regents exam questions, we found that a large proportion of the 31 questions solvable by our techniques were comparison and story questions, indicating that the addition of QR to Aristo may boost its performance.

We further note that most attempts to solve problems such as these focus on information retrieval techniques over text (e.g. Sachan et al. 2016), or lightweight knowledge representation schemes where the tokens in semi-structured representations are still words or phrases (e.g. Khashabi et al. 2016). By contrast, we are using deductive reasoning over conceptual representations. While we agree that there are

roles for maintaining linguistic information in extracting knowledge from text, we also believe that the refinement of such knowledge into conceptual knowledge is a crucial, but underexplored, component of learning by reading. Efforts to date at such refinements include Semantic Construction Grammar (Schneider & Witbrock, 2015) and Companion-based learning by reading (Barbella & Forbus, 2015).

One of the foundational works for qualitative reasoning was Hayes' (1979) naïve physics manifesto, which encouraged the field to look at commonsense physical reasoning. Some research has focused on broad, axiomatic accounts of phenomena, e.g. liquids (Hayes 1985), matter (Davis, 2010), and containers (Davis et al. 2013), but none of these efforts were tied into a large, overarching ontology. We believe that integrating such efforts into the Cyc ontology (which can be used freely, by staying with OpenCyc) would radically improve the ability to create the kind of larger-scale, integrated accounts needed to broadly cover commonsense science. In 4th grade science, qualitative simulation seems unnecessary, but that is unlikely to be true at higher grades, at which point qualitative simulators like Garp3 (Bredeweg et al. 2009) may prove valuable.

Discussion and Future Work

We agree with AI2 that commonsense science is a useful approach to studying the nature of commonsense reasoning more broadly. We are encouraged that over a quarter of the exam questions involve qualitative representations and reasoning, and that standard QR techniques perform well on this portion of 4th grade exams. Prior research by Bruce Sherin[6] indicates that the content of middle-school science remains focused on qualitative knowledge, to provide a firm foundation for integrating with algebra and calculus later on. An analysis of 8th grade exams, in progress, looks likely to provide additional evidence for the centrality of qualitative representations and reasoning for commonsense science.

We note that, like in prior projects, the broad contents of the ResearchCyc knowledge base provide significant leverage for this kind of research. Being able to draw on a wide-ranging ontology is useful to reduce tailorability, but more importantly, it provides leverage on its own (e.g. `ScalarInterval` as a simple form of qualitative value well suited for capturing the ambiguities inherent in natural language). Even when there are design choices that are not optimal from a particular perspective (e.g. formalizing some quantities as values instead of fluents), simple coercions typically suffice to put the knowledge in a more useful form.

Much future work remains, of course. First, we plan to extend the Companion natural language facilities to automatically interpret exam questions to generate the kinds of queries that here were created by hand. Second, we plan to extend our learning by reading work (e.g. Lockwood & Forbus, 2009; Barbella & Forbus, 2015) to provide the broad-scale knowledge needed to handle these kinds of questions. Third, we plan to use a combination of computer vision techniques and sketch understanding (Forbus et al. 2011) to automatically process the visual aspects of questions. Finally, we plan on exploring interactive training of Companions on commonsense science, by posing scenarios and asking questions, including follow-up questions aimed at exposing misconceptions gleaned from learning by reading.

[6] Personal communication

Acknowledgements

This research was supported by the Intelligent and Autonomous Systems Program of the Office of Naval Research.

References

Barbella, D. and Forbus, K. (2015). Exploiting Connectivity for Case Construction in Learning by Reading. *Proceedings of the Third Annual Conference on Advances in Cognitive Systems*.

Bredeweg, B., Linnebank, F., Bouwer, A., and Liem, J., 2009. Garp3 - Workbench for Qualitative Modelling and Simulation. *Ecological Informatics* 4(5-6), 263-281.

Clark, P. (2015) Elementary School Science and Math Tests as a Driver for AI: Take the Aristo Challenge! *Proceedings of IAAI-2015*, pp. 4019-4021.

Clark, P., Harrison, P., & Balasubramanian, N., (2013). A study of the knowledge base requirements for passing an elementary school science test. *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, pp 37-42.

Clark, P., Etzioni, O., Khashabi, D., Khot, T., Sabharwal, A., Tafjord, O., & Turney, P. (2016) Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions. *Proceedings of AAAI 2016.*

Davis, E. (2010) Ontologies and Representations of Matter, *Proceedings of AAAI-2010*.

Davis, E., Marcus, G., & Chen, A. (2013) Reasoning from Radically Incomplete Information: The Case of Containers. *Proceedings of the 2nd Annual Conference on Advances in Cognitive Systems.* Baltimore, MD.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, 41(1), 1-63.

Forbus, K. (1984) Qualitative Process Theory. *Artificial Intelligence*, 24, pp. 85-168.

Forbus, K. (2016) Software Social Organisms: Implications for measuring AI Progress. *AI Magazine*, 37(1), pp 85-90.

Forbus, K., Usher, J., Lovett, A., Lockwood, K., & Wetzel, J. (2011). CogSketch: Sketch understanding for Cognitive Science Research and for Education. *Topics in Cognitive Science*. 3(4), pp 648-666.

Forbus, K., Ferguson, R., Lovett, A., & Gentner, D. (in press). Extending SME to handle large-scale cognitive modeling. *Cognitive Science*.

Hayes, P. (1979). The naive physics manifesto. In D. Michie (Ed.), *Expert Systems in the Micro-Electronic Age*. Edinburgh University Press.

Hayes, P. (1985). Naive Physics I: Ontology for liquids. In Hobbs, R., & Moore, R. (Eds.), *Formal Theories of the Commonsense World*. Norwood, NJ: Ablex Publishing Corporation.

Khashabi, D., Khot, T., Sabharwal, A., Clark, P., Etzioni, O., & Roth, D. (2016) Question Answering via Integer Programming over Semi-Structured Knowledge. *Proceedings of IJCAI 2016*.

Kim, H. (1993). Qualitative Reasoning about Fluids and Mechanics. Ph.D. Dissertation and ILS Technical Report, Northwestern University.

Klenk, M., Forbus, K., Tomai, E., & Kim, H. (2011) Using analogical model formulation with sketches to solve Bennett Mechanical Comprehension Test problems. *Journal of Experimental and Theoretical Artificial Intelligence*, 23(3): 299-327.

Lockwood, K. and Forbus, K. (2009). Multimodal knowledge capture from text and diagrams. *Proceedings of KCAP-2009*.

Sachan, M, Dubey, A., & Xing, E. (2016) Science Question Answering using Instructional Materials. Arxiv:1602.04375v5, April 5th.

Schneider, D. & Witbrock, M. (2015) Semantic Construction Grammar: Bridging the NL/Logic Divide. *Proceedings of WWW-2015*.

Weld, D. (1986) Comparative Analysis. *Artificial Intelligence*, 36, pp. 333-373.

Weld, D. (1990) *Theories of Comparative Analysis*. MIT Press. Cambridge, MA.