What's in an Explanation? Characterizing Knowledge and Inference Requirements for Elementary Science Exams

Peter Jansen[1*], **Niranjan Balasubramanian**[2], **Mihai Surdeanu**[3], **and Peter Clark**[4]
[1]School of Information, University of Arizona, Tucson, AZ
[2]Dept. of Computer Science, Stony Brook University, Stony Brook, NY
[3]Dept. of Computer Science, University of Arizona, Tucson, AZ
[4]Allen Institute for Artificial Intelligence, Seattle, WA
[*]Corresponding author: pajansen@email.arizona.edu

Abstract

QA systems have been making steady advances in the challenging elementary science exam domain. In this work, we develop an explanation-based analysis of knowledge and inference requirements, which supports a fine-grained characterization of the challenges. In particular, we model the requirements based on appropriate sources of evidence to be used for the QA task. We create requirements by first identifying suitable sentences in a knowledge base that support the correct answer, then use these to build explanations, filling in any necessary missing information. These explanations are used to create a fine-grained categorization of the requirements. Using these requirements, we compare a retrieval and an inference solver on 212 questions. The analysis validates the gains of the inference solver, demonstrating that it answers more questions requiring complex inference, while also providing insights into the relative strengths of the solvers and knowledge sources. We release the annotated questions and explanations as a resource with broad utility for science exam QA, including determining knowledge base construction targets, as well as supporting information aggregation in automated inference.

1 Introduction

Elementary science exams have recently become a common test of question answering (QA) models. Clark and Etzioni (2016) argue that these exams are an excellent benchmark for natural language processing (NLP) systems in many respects, both testing students for many different kinds of knowledge and inference abilities at varying levels of difficulty, while also allowing for a direct comparison of machine to human performance in the science domain on a standardized evaluation. Many different QA approaches have been developed and evaluated on these and similar exams, with methods using a range of representations from unstructured (BOW) lexical semantic models (Fried et al., 2015), structured relation-based representations (Clark et al., 2016; Khot et al., 2015), more complex first-order formalisms (Khot et al., 2015), and other inference methods (Khashabi et al., 2016). Together in concert, these methods can achieve substantial improvements in overall performance, with a 71% accuracy (i.e. passing performance) on one test set (Khashabi et al., 2016).

In this work, we focus on developing a deeper understanding of this problem domain by implementing a fine-grained characterization of the knowledge and inference requirements for science exam QA, driven by generating and annotating gold explanations that justify the correct answer. We believe that this can provide many tangible benefits. First, we can obtain a fine-grained assessment of the abilities of different QA systems to identify areas of competency, and those that need improvement. Second, the detailed knowledge requirements can serve as a specification for knowledge extraction. Third, it can support QA methods that can use problem solving strategies and knowledge tailored to the specific requirements of a given question. Finally, it can support design of QA systems that can provide explanations for why they choose an answer. In this last respect, multiple-choice elementary science questions currently lack a direct way to quantitatively assess systems on this aspect.

Specifying broadly applicable knowledge requirements and explanations poses two main challenges. First, questions can be answered in many ways, and depending on the knowledge source used the type of knowledge ascribed to the question can differ. We follow a pragmatic approach, building on prior work in knowledge categorization, and use knowledge types that correspond to commonly used semantic structures relating to the automatic construction of knowledge bases (KB). Clark et al. (2013) compiled an initial analysis of the questions in these datasets, and identified 7 broad categories of knowledge and inference requirements. However, this analysis forced a single knowledge type for each question, for example *causality*, and from our detailed analysis we find that many types of knowledge are necessary to arrive at the correct answer, e.g., *causality, actions, and purposes*.

A second challenge relates to grounding the requirements and the explanations in appropriate resources such that they can facilitate automated analysis and provide compact, reusable, and linked knowledge for inference. To this end, we use grade appropriate texts, and first identify relevant sentences or nuggets of information that can serve as explanations or supports for the current answers. We then fill in sentences that provide missing links connecting knowledge and terminology in the sentences, while taking care to ensure as much reuse as possible.

We apply this methodology to obtain requirements on a set of 212 questions from an open standardized elementary science exam dataset, and present an analysis of these requirements. This work makes the following contributions:

- We construct a detailed characterization of the knowledge and inference requirements of elementary science exams, highlighting the prevalence of complex inference questions, which require inference methods that combine many facts across multiple types of knowledge.
- We provide an empirical analysis of the performance of different QA methods on questions with specific knowledge and inference requirements, demonstrating that while existing QA systems considerably outperform information retrieval (IR) methods on difficult questions, many of the more complex forms of inference remain to be addressed.
- We provide a knowledge resource in the form of gold explanations for hundreds of science exam questions, as well as annotation describing question-centered and explanation-centered knowledge and inference requirements. We believe this resource will be broadly useful for characterizing performance on current and future models, as well as developing automated methods supporting knowledge type identification, inference, and explanation construction.

2 Related Work

Analyzing knowledge and inference requirements is a first necessary step in designing QA systems. For factoid QA tasks, these requirements are often stated in terms of broad question categories (e.g., What, When, How) and finer-grained types for expected answers (e.g., cities, person, organization). Factoid QA systems use classifiers to identify the types of question and expected answers, which are subsequently used to select specific problem solving routines, and to filter answer candidates (Harabagiu et al., 2000; Li and Roth, 2006; Roberts and Hickl, 2008). For non-factoid QA tasks, requirements are often stated in terms of elements in knowledge representation ontologies. For instance, Chaudhri et al. (2014) study requirements for a QA task defined over AP Biology texts using relations and categories from the CLIB ontology (Barker et al., 2001). Some benchmarks, such as bAbI (Weston et al., 2016), are created to test specific reasoning abilities and come with a grouping of questions into the corresponding categories (e.g., negation reasoning, causal reasoning).

Our work aims to provide similar requirements for the elementary science QA benchmark (Clark and Etzioni, 2016). Prior analyses on this benchmark includes Clark et al. (2013), who identified seven broad kinds of knowledge and inference in three categories: *retrieval questions*, making use of taxonomic, definitional, or property knowledge; *inference questions*, testing a knowledge of causality, processes, or identifing examples of situations; and *domain-specific models*. Crouse and Forbus (2016) further identified questions that involve qualitative reasoning (13% of total), and provide a sub-categorization of these. Here we build upon these prior works and provide both a more fine-grained characterization of the knowledge types required to answer these questions, along with manually curated answer explanations.

This allows us to compare the relative strengths and weaknesses of different QA systems from knowledge and inference requirements identified using both bottom-up (from explanations) and top-down (from questions) approaches.

More broadly and with respect to explanations, there is a recent trend towards emphasizing interpretable models for machine learning (e.g. Ribeiro et al. (2016)) that are able to produce human-readable explanations for their reasoning, both to improve human trust in automated inference, as well as to verify that a given model is accurately capturing the aspects of complex reasoning required for a given task. We view this work as complementary, here characterizing the knowledge and inference requirements that an automated reasoning method for science exams must meet to assemble compelling human-readable explanations as part of the inference process.

3 Knowledge and Inference Analysis

Estimating knowledge and inference requirements is challenging for many reasons. Chief among these is that a question can be answered in many different ways, using different types of knowledge and reasoning depending on the sources of evidence used. At one extreme, with a large knowledge base (KB), many questions can be answered by simply retrieving a fact from the KB that readily provides the correct answer. At the other extreme, with a modest KB, multiple pieces of information have to be aggregated together using some inference method to arrive at the correct answer. A further difficulty in multiple choice exams is that a QA system may select the correct answer, but for the wrong reasons stemming from difficulties in retrieval, inference, or from simply using a backoff strategy (e.g. guessing)[1]. Question answering systems in the science and medical domains should also target providing human-readable explanations for why the selected answer is correct. We examine knowledge requirements for this explainable question answering task, which suggests that, at the very least, requirements must be grounded in explanations drawn from a reasonable collection of target sources of evidence.

Towards this goal, we develop an explanation-centered approach using appropriate grade-level resources, constructing gold natural language explanations that detail why a given answer is correct, and deriving a fine-grained distribution of common inference relations from these explanations. In this section, we first provide a question-centered analysis expanded to a larger set of questions compared to prior work, and demonstrate the challenges with this approach. We then present a fine-grained analysis using the explanation-centered approach on the same set of questions.

Questions: For the following analyses, we make use of the 432 training questions in the AI2 Elementary Science Questions set[2], collected from standardized 3^{rd} to 5^{th} grade science exams in 14 US states.

3.1 Question-centered Analysis

Figure 1 shows the distribution of knowledge and inference requirements when extending the question-centered analysis of Clark et al. (2013) to the larger AI2 elementary questions set. We find two differences when compared to their original analysis on 50 4^{th} grade questions from the New York Regents Science Exam: First, the distribution on this larger question set exhibits a much higher proportion of complex inference (77%) compared to retrieval methods. Second, even though we annotated one knowledge category per question according to the original procedure, we find that many of the complex inference questions naturally require integrating several different kinds of knowledge to arrive at the answer, with more than a third of the questions requiring at least two knowledge types.

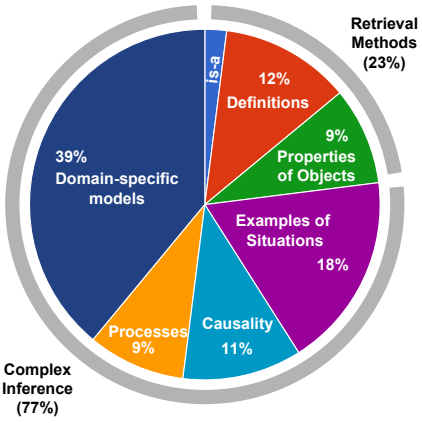

Figure 1: Knowledge types required to correctly answer a given question in the elementary science exam dataset.

[1]Jansen, Sharp, Surdeanu, and Clark (submitted) showed in their error analysis that, for elementary science questions, both retrieval and inference methods produce completely incorrect explanations approximately 20% of the time. A retrieval model produced complete explanations for 45% of questions, while an inference model incorporating intersentence aggregation produced complete explanations for 60% of questions.

[2]The original question set is available at: `http://allenai.org/data.html`

Question	Which of these organisms has cells with cell walls?
Answer Choices	(A) bluebird **(B) A pine tree** (C) A ladybug (D) A fox squirrel
Explanation	A pine tree is a kind of plant. A cell wall is a part of a plant cell.
Question	What form of energy causes an ice cube to melt?
Answer Choices	(A) mechanical (B) magnetic (C) sound **(D) heat**
Explanation	An ice cube is a solid. Changing from a solid to a liquid is called melting. Melting happens when solids are heated. Heated means added heat. Heat is a kind of energy.
Question	Which of the following events involves a consumer and producer in a food chain?
Answer Choices	(A) A cat eats a mouse. **(B) A deer eats a leaf.** (C) A hawk eats a mouse. (D) A snake eats a rat.
Explanation	A leaf is a kind of plant. A deer is a kind of animal. In a food chain, an animal is a consumer. In a food chain, green plants are producers.

Table 1: Explanations for three shorter example questions, including one simpler question about the property of an object (having cell walls), an explicitly causal question (melting), and one question about the role of two entities in a process or model (the food chain). Dashed underlines indicate bridge sentences.

3.2 Explanation-centered Analysis

3.2.1 Gold Explanations

For each question, we create gold explanations that describe the inference needed to arrive at the correct answer. Our goal is to derive an explanation corpus that is grounded in grade-appropriate resources. Accordingly, we use two elementary study guides, a science dictionary for elementary students, and the Simple English Wiktionary as relevant corpora. For each question, we retrieve relevant sentences from these corpora and use them directly, or use small variations when necessary. If relevant sentences were not located, then these were constructed using simple, straightforward, and grade-level appropriate language. Approximately 18% of questions required specialized domain knowledge (e.g. spatial, mathematical, or other abstract forms) that did not easily lend itself to simple verbal description, which we removed from consideration. This resulted in a total of 363 gold explanations.[3]

In addition to using grade-appropriate language, the following considerations were taken in developing the explanation corpus, with the aim to provide broad utility for a variety of tasks from automated knowledge type identification to information aggregation models of inference:

- *Single topic:* To help facilitate automated analysis and reuse, explanations were broken into multiple sentences, with each sentence focusing on a single aspect of the explanation.
- *Reuse:* To assist in identifying overlaps in knowledge between questions, the same explanation sentences were reused as much as possible, where applicable.
- *Sentence Linking:* To support automated inference, the terminology used in different explanation sentences is explicitly linked through "bridge sentences" that include both terms. For example, if one sentence mentions *melting*, and another mentions *heated*, here we include an explicit sentence that links the two, such as *"melting happens when solids are heated"*. Where appropriate, we also include other latent knowledge that may not be explicitly required to answer a question, but would likely be available to a human and link related questions. For example, for a process question about a specific stage of the *life cycle*, we also include a brief overview of where this stage fits in the process as a whole (e.g. *egg to baby to child to adult*). In this way many of the explanations appear overly verbose to a human, but contain enough information to make the inference explicit, link highly related topics, and evaluate the knowledge requirements for automated methods.

Example explanations are shown in Table 1. The 363 gold explanations contain a total of 1272 sentences, or an average of 4 sentences per explanation. With respect to reuse, 943 unique sentences appear across these explanations, with 180 appearing in more than one explanation, and the remaining 763 occurring in only a single explanation.[4]

[3]The gold explanations developed in this work are also available at: `http://allenai.org/data.html`

[4]Frequently-recurring sentences highlight common themes in questions: Sentences such as *"Evaporation is when a liquid changes to a gas", "Sunlight means solar energy", and "Metals conduct electricity"* are 5 of the 42 sentences found in the explanations of 4 or more questions.

Knowledge Type	Prop.	Structure and Examples
Retrieval Types		
Taxonomic	83%	*HYPONYM* is a kind of *HYPERNYM* a <*HYPO: plant*> is a kind of <*HYPER: living thing*>
Definition	64%	*ARG1* means *ARG2 (can be definitions or synonyms)* <*ARG1: cooling*> means <*ARG2: decreasing heat*> *(definition)*
Properties	41%	*PROPERTY* is a property of *ARG1* <*ARG1: iron*> is <*PROP: magnetic*>
PartOf	22%	*MERONYM* is a part of *HOLONYM*. a <*HOLO: bicycle*> has <*MERO: two pedals*>.
Contains	17%	*ARG1* contains *ARG2*. <*ARG1: soil*> contains <*ARG2: nutrients*> that plants absorb through their roots
ExampleOf	9%	*ARG1* is an example of *ARG2* an example of a <*ARG1: seasonal change*> is <*ARG2: growing thick fur*>
MadeOf	8%	*ARG1* is made of *ARG2*. a <*ARG1: rock*> is a hard substance composed of <*ARG2: minerals*>
Inference Supporting Types		
Actions	73%	*SUBJECT ACTION OBJECT* <*SUBJ: bees*> <*ACTION: eat*> <*OBJ: pollen*>
UsedFor	33%	*WHO* uses *WHAT*, and *WHY*. <*WHO: squirrels*> <*WHAT: store food in the autumn*> <*WHY: to eat over the winter*>
Source	23%	*WHO* generates/is a source of *WHAT*, and *HOW* . natural gas is can be burned in power stations to *make electricity* (note sourceof+generate)
IsWhen	22%	*ARG1* is when *ARG2* happens. *(often used for defining events/processes)* <*ARG1: mechanical weathering*> is when <*ARG2: rocks are broken down by mechanical ...*>
VehicleFor	17%	*WHAT* happens by/through some means or *VEHICLE*. ... when <*WHAT: pollen is carried from flower to flower*> <*VEHICLE: by pollinating animals*>
Requires	12%	*WHO* requires *WHAT*, and *WHY*. <*WHO: animals*> need to <*WHAT: eat food*> <*WHY: to get nutrients required for survival*>
Negation	12%	*ARG1* is not *ARG2*. aluminum is not <*NOT: magnetic*>
Duration	10%	*ARG1* has some *DURATION* many birds <*ARG1: migrate to warmer places*> <*DUR: for the winter*>
Complex Inference Types		
Changes	45%	*WHO/LABEL* changes *WHAT*, *FROM* something *INTO* something else. <*LABEL: boiling*> means changing from a <*FROM: solid*> to a <*INTO: liquid*>
Causes	21%	*ARG1* causes *ARG2*. <*ARG1: friction*> causes <*ARG2: the temperature of an object to increase*>
Transfer	21%	*WHAT* gets transfered from a *SOURCE* to *DESTINATION*, and *HOW* this happens. ... breaks down food into <*WHAT: nutrients*> that can be <*HOW: absorbed*> by <*DEST: the body*>
IfThen	14%	*IF* a condition occurs, *THEN* a result happens. if <*IF: something is on fire*>, <*THEN: it burns*>
Relationship	12%	As *EVENT1* happens, *EVENT2* will also happen. <*EVENT1: A decrease in the amount of water*> will cause <*EVENT2: a decrease in plant populations*>
Process	8%	A group of relations, e.g. A *PROCESS STAGE* takes some *ACTION* causing a *RESULT*. as an <*STAGE: adult bird*>, <*ACTION: it will reproduce*>, <*RESULT:starting the life cycle...*>

Table 2: Fine-grained knowledge types, and the proportion of explanations that include at least one instance of a given type. Types are *n-ary* relations, containing between two and five arguments each. Note that a given example sentence often includes more than one relation, as in the case of "cooling means decreasing heat", which includes both a *Definition* relation (i.e. means), and a *Change* relation (i.e. decreasing heat).

3.2.2 Fine-grained Knowledge Types

To characterize the knowledge present in these gold explanations, we annotated the explanation sentences with a fine-grained set of knowledge types which reuses many of the types from Clark et al. (2013) and includes additional types derived from frequently observed semantic structures in the explanation sentences. Each explanation sentence can contain more than one type (e.g. *"boiling means increasing temperature"* contains both a Definition type (*boiling means ...*) and a Change type (*increasing temperature*)). All types were manually annotated using a graphical annotation tool[5]. Due to the time involved in this process, we annotated 212 questions, or approximately 50% of the original set of questions.

Table 2 shows the new fine-grained set of knowledge types, their relative frequencies, and the associated semantic structures. About 21% of the annotated questions had between 1 and 5 instances of types

[5]This simple graphical annotation tool is included with the data distribution.

in their explanations, while 31% had between 6 and 10 instances. The remainder of questions with more than 10 relations across their explanations were largely complex questions that included latent or other background knowledge in their explanations.

The fine-grained types can also be grouped into three broad sets: *Retrieval Types* include binary relations commonly found in taxonomies, dictionaries, and property databases. *Inference Supporting Types* tend to ground the knowledge in the complex inference relations. This includes describing the vehicle that enables something to happen, it's purpose, it's needs, and specific actions that it can take. *Complex Inference Types* describe changes situated in particular contexts, such as causality (e.g. *X causes Y*), transfers (e.g. *X transfers from Y to Z*), and process knowledge (e.g. *Stage A follows Stage B*). Here, while our *Retrieval Types* are binary relations, both the *Complex Inference* and *Inference Supporting Types* can be viewed as *n-ary* relations or light semantic frames, often with two to five "slots" to fill.

4 QA Analysis

Here we conduct an empirical analysis of the performance of two types of QA solvers using the question-centered and explanation-centered views of knowledge and inference types.

4.1 Knowledge Bases

We evaluate performance on two knowledge bases, one free text, the other semi-structured:

Study Guides: A collection of free text from six resources: study guides for two elementary science exams, a teacher's manual, a set of flashcards, and two dictionary resources: a science dictionary for kids, and the open-domain Simple English Wiktionary[6]. A total of 3,832 science-domain sentences and 17,473 open-domain definition sentences were included.

Aristo TableStore: An open collection[7] of approximately 100 semi-formal tables (approximately 10k rows, 30k cells) containing knowledge tailored to elementary science exams, constructed using a mixture of manual and automatic methods (Dalvi et al., 2016). The table knowledge spans across knowledge types, from properties and taxonomic knowledge to causality, processes, and domain models. Each table encodes an aspect of the science domain (e.g., animal adaptations, measuring instruments, energy conversions, etc.), where variations are typically enumerated (e.g. "a <*grill*> converts <*chemical energy*> to <*heat energy*>", "a <*flashlight*> converts <*electrical energy*> into <*light energy*>", etc.).

4.2 Solvers

We characterize QA approaches from two families: a baseline that uses "learning to rank" (L2R) with information retrieval (IR) features, and more recent inference models.

Retrieval Model:

We use an L2R model which finds answers by scoring passage level evidence for each answer choice from the unstructured textual knowledge sources. Our implementation is based on the candidate ranking (CR) model described in Jansen et al. (2014). Short passages are scored based on how similar they are to the words in the question and the corresponding answer choice. The similarity scores are computed using cosine similarity of *tf.idf* representations of the question and passages, and used in a L2R framework to produce the final ranking of the answer choices. We created two versions of the solver: one that uses the study guide collection, and the other with a textual representation of the Aristo TableStore. Apache Lucene[8] is used to index and retrieve passages.

Inference Models:

For inference, we use two models that operate over a structured knowledge base of tables (TableStore). TableILP (Khashabi et al., 2016) is a model that finds answers by building a graph of chained facts, i.e., rows in the knowledge tables, to arrive at the answer. Starting from the question, the model selects rows from a table, and then iteratively uses the selected rows to find rows in other tables, as linkable facts,

[6] http://simple.wiktionary.org
[7] http://allenai.org/data.html
[8] http://lucene.apache.org

AKBC'13 Knowledge Type	N	Proportion Correct L2R (StudyGuides)	L2R (TableStore)	ILP (TableStore)	STITCH (Tablestore)
Retrieval Methods					
Taxonomic	4	75% (0%)	75%	100% (+25%)	100% (+25%)
Definition	27	56% (-7%)	63%	59% (-4%)	63% (0%)
Properties	19	21% (-11%)	32%	53% (+21%)	53% (+21%)
Complex Inference					
Examples	40	35% (-13%)	48%	70% (+22%)	58% (+10%)
Causality	30	30% (-10%)	40%	60% (+20%)	53% (+13%)
Processes	26	52% (+4%)	48%	36% (-12%)	64% (+16%)
Domain-specific Models	66	38% (+6%)	32%	43% (+11%)	53% (+21%)
Overall	212	39% (-4%)	43%	54% (+11%)	56% (+13%)

Table 3: Proportion of questions answered correctly broken down by AKBC'13 knowledge types. Values in parentheses reflect absolute differences with the L2R solver that uses the TableStore knowledge base.

until it arrives at facts that contain or overlap with the answer choices. Rows are selected based on lexical overlap. This graph building problem is modeled using Integer Linear Program (ILP) to find paths that maximize QA performance. STITCH is an alternative algorithm for reasoning over the same tables. It achieves similar overall performance using different heuristics for matching a question to table rows. For both inference models, we made use of the stock models, and did not incorporate any further training. As described below, we make use of a different question corpus and an expanded knowledge base compared to Khashabi et al., evaluating on approximately twice as many questions as were originally reported, including many questions at a higher grade level, and including questions from 13 other state exams in addition to the original New York Regents questions. Similarly, we make use of an expanded knowledge base that is approximately twice the size of that used in Khashabi et al. (2016). As such, our overall inference model performance is slightly lower than they originally reported.

Questions: We compare performance on the 212 elementary science questions from Section 3.2.2 that included a gold explanation annotated with the knowledge and inference types.[9]

4.3 Question-centered Evaluation

We first characterize performance of the two solvers using the seven broad question-centered categories of Clark et al. (2013), with performance shown in Table 3. Overall, the L2R models have lower performance than the inference models. This is in line with our explanation-based analysis of the requirements, which showed that there are more complex inference questions than there are simple retrieval ones. The results also show that the gains in the inference solvers are not completely due to tailored knowledge. Using the highly tailored knowledge base as a retrieval corpus shows a small benefit (+4%), whereas using the knowledge via appropriate inference substantially increases performance (+13%).

In terms of performance on questions with particular knowledge and inference requirements, we find that bulk of the performance benefit for the inference solvers comes from addressing more complex inference questions, rather than simply answering more of the (subjectively easier) retrieval questions. Performance on *Example Identification* and *Causality* questions using the L2R model increases 10-13% when switching from the study guide knowledge base to the Tablestore, and further increases by 10-22% when the inference solvers are used in conjunction with the Tablestore, demonstrating that some complex questions separately benefit from highly tailored knowledge and the capacity to aggregate multiple pieces of that knowledge to form a solution. Conversely, the more challenging *Process* and *Domain Model* categories are not directly benefited by the tailored Tablestore knowledge resource, but show moderate benefits when this knowledge is combined with the inference solvers to form more complex solutions.

On the balance, this high-level analysis shows that inference methods designed to aggregate multiple pieces of information from a knowledge base specifically benefit questions requiring complex inference, more than the contribution of tailoring a similarly-sized retrieval-centered knowledge base alone.

[9]Note that because this set excludes the 18% of questions that did not easily lend themselves to textual explanation, and that 70% of these excluded questions were categorized as requiring model-based reasoning, this evaluation set can be viewed as somewhat easier and containing fewer extremely difficult questions than the broader corpus.

Knowledge Type	N	L2R (Corpus)	Knowledge Advantage	L2R (TableStore)	Inference Advantage	ILP (TableStore)	STITCH (TableStore)
Retrieval Types							
Taxonomic	176	39% (-7%)	→	46%	→	56% (+10%)	55% (+9%)
Definition	135	39% (-2%)	X	41%	→	56% (+15%)	55% (+14%)
Properties	86	38% (+2%)	X	36%	→	49% (+13%)	56% (+20%)
PartOf	47	43% (+11%)	←	32%	→	45% (+13%)	68% (+36%)
Contains	35	29% (-11%)	→	40%	X	43% (+3%)	40% (+0%)
ExampleOf	19	47% (+5%)	X	42%	→	63% (+21%)	63% (+21%)
MadeOf	16	50% (-13%)	→	63%	X	56% (-7%)	63% (+0%)
Inference Supporting Types							
Action	154	40% (-4%)	X	44%	→	54% (+10%)	57% (+13%)
UsedFor	70	44% (0%)	X	44%	→	59% (+15%)	70% (+26%)
SourceOf/Generate	49	43% (+2%)	X	41%	→	53% (+12%)	65% (+24%)
IsWhen/IsCalled	46	28% (-22%)	→	50%	→	70% (+20%)	54% (+4%)
Vehicle	35	40% (-3%)	X	43%	→	54% (+11%)	54% (+14%)
Requires	26	39% (+12%)	←	27%	→	54% (+27%)	50% (+23%)
Negation	26	15% (-7%)	X	22%	→	44% (+22%)	52% (+30%)
Duration	21	57% (0%)	X	57%	→	67% (+10%)	48% (-9%)
Complex Inference Types							
Change	96	34% (-8%)	→	42%	→	53% (+11%)	51% (+9%)
Cause	45	38% (0%)	X	38%	→	56% (+18%)	53% (+15%)
Transfer	45	44% (-9%)	→	53%	→	64% (+11%)	62% (+9%)
Relationship	25	28% (0%)	X	28%	→	44% (+16%)	36% (+8%)
IfThen	29	41% (+6%)	X	35%	→	41% (+6%)	45% (+10%)
Process (Content/Roles)	25	44% (-17%)	→	61%	X	61% (0%)	57% (-4%)
Process (Structural)	12	25% (-50%)	→	75%	←	58% (-17%)	50% (-25%)
Average Performance		39% (-4%)	X	43%	→	54% (+11%)	56% (+13%)

Table 4: Performance on questions whose gold explanations contain *at least one* instance of a given type.Values in parentheses reflect absolute differences with the score of the L2R solver that uses the TableStore knowledge base. Arrows represent where performance on a given relation shows a benefit from either knowledge base, or switching from a retrieval to an inference solver, where an "X" signifies no benefit.

4.4 Explanation-centered Evaluation

We conduct an explanation-centered evaluation to understand the comparative finer-grained competencies of the solvers. Table 4 compares performance relative to whether the *gold explanation* for a given question contains *at least one* instance of the specific type. If a question contains a specific type according to the annotation, then we assert that type of knowledge or inference is required to answer (and produce an explanation for) that science exam question. We note three main observations.

First, the inference solver outperforms L2R solvers across the board, with strong improvements when there are retrieval or inference-supporting types, and smaller improvements for explanations with complex inference types, except for the causal types (+18% gain in P@1). Conversely, despite gains with inference solvers, questions of some complex inference types, such as *If/Then* conditional sequences, or *Coupled Directional Relationships* (i.e. *as X increases, Y decreases*), have low overall absolute performance, pointing to areas for future improvement.

Second, there is a variance in performance across the broader groups when switching over to Tablestore for the L2R solver. For example, *Taxonomic, Containment, and MadeOf* see benefits, whereas *Definition, Properties, and ExampleOf* do not. *PartOf*, and *Requirement* types work better with study guides rather than Tablestore knowledge, suggesting the entirety of the study guide knowledge is not subsumed by the tablestore. Similar variance exist for the complex inference types, as well.

Third, the broad types of the question-based analysis can be inadequate in some cases. The broad *Process* category in Table 3 showed some general improvement with inference methods, but the fine-grained analysis shows the opposite. This is likely because the broad *Process* category is an umbrella for several different types of questions. Some query only a very specific stage of a process (like a *producer's role in the food chain*), and are amenable to being answered by single sentences found using retrieval methods. Others require integrating structural knowledge across many stages of a process (such as *from egg to adult in the life cycle*), and appear to require much more complex inference to explainably answer.

5 Conclusions

In this work we developed an explanation-centered fine-grained characterization of elementary science exams, helping improve our understanding of this problem domain. Rather than existing in easily decoupled categories, these exams show a rich distribution of knowledge and inference requirements, with a majority requiring complex inference. The analyses validate the gains with some inference-based solvers by showing that they specifically address questions requiring complex inference. While a modern inference solver shows steady improvements in complex inference broadly, performance for a number of specific types of complex inference is still quite low, and provides targets for future work.

We release the annotated questions and explanations as a knowledge resource that can be broadly useful for science exam QA. As question variety, difficulty, and domain-specificity increase, any *single* solver is unlikely to work well across the board. This motivates development of solver ensembles and question-specific solver selection, which need the capacity to automatically recognize a given question's knowledge and inference requirements. We believe this resource may have a range of other uses, from providing a specification of knowledge base construction targets, to informing methods of information aggregation in automated inference.

References

[Barker et al.2001] Ken Barker, Bruce Porter, and Peter Clark. 2001. A library of generic concepts for composing knowledge bases. In Proceedings of First International Conference on Knowledge Capture, pages 14–21.

[Chaudhri et al.2014] Vinay K. Chaudhri, Daniel Elenius, Andrew Goldenkranz, Allison Gong, Maryann E. Martone, William Webb, and Neil Yorke-Smith. 2014. Comparative analysis of knowledge representation and reasoning requirements across a range of life sciences textbooks. volume 5:51.

[Clark and Etzioni2016] Peter Clark and Oren Etzioni. 2016. My computer is an honor student but how intelligent is it? standardized tests as a measure of ai. AI Magazine, 37(1):5–12.

[Clark et al.2013] Peter Clark, Philip Harrison, and Niranjan Balasubramanian. 2013. A study of the knowledge base requirements for passing an elementary science test. In Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, AKBC'13, pages 37–42.

[Clark et al.2016] Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D. Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA., pages 2580–2586.

[Crouse and Forbus2016] Maxwell Crouse and Kenneth D. Forbus. 2016. Elementary school science as a cognitive system domain: How much qualitative reasoning is required? In Proceedings of Fourth Annual Conference on Advances in Cognitive Systems.

[Dalvi et al.2016] Bhavana Dalvi, Sumithra Bhakthavatsalam, Chris Clark, Peter Clark, Oren Etzioni, Anthony Fader, and Dirk Groeneveld. 2016. IKE - an interactive tool for knowledge extraction. In Proceedings of the 5th Workshop on Automated Knowledge Base Construction, AKBC@NAACL-HLT 2016, San Diego, CA, USA, June 17, 2016, pages 12–17.

[Fried et al.2015] Daniel Fried, Peter Jansen, Gustave Hahn-Powell, Mihai Surdeanu, and Peter Clark. 2015. Higher-order lexical semantic models for non-factoid answer reranking. Transactions of the Association for Computational Linguistics, 3:197–210.

[Harabagiu et al.2000] Sanda Harabagiu, Dan Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan Bunescu, Roxana Girju, Vasile Rus, and Paul Morarescu. 2000. Falcon: Boosting knowledge for answer engines. In Proceedings of the Text REtrieval Conference (TREC), Gaithersburg, MD, USA.

[Jansen et al.2014] Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL).

[Khashabi et al.2016] Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. 2016. Question answering via integer programming over semi-structured knowledge. In Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI'16, pages 1145–1152.

[Khot et al.2015] Tushar Khot, Niranjan Balasubramanian, Eric Gribkoff, Ashish Sabharwal, Peter Clark, and Oren Etzioni. 2015. Exploring markov logic networks for question answering. In EMNLP.

[Li and Roth2006] Xin Li and Dan Roth. 2006. Learning question classifiers: The role of semantic information. Natural Language Engineering, pages 229–249.

[Ribeiro et al.2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 1135–1144, New York, NY, USA. ACM.

[Roberts and Hickl2008] Kirk Roberts and Andrew Hickl. 2008. Scaling answer type detection to large hierarchies. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[Weston et al.2016] Jason Weston, Antoine Bordes, Sumit Chopra, Sasha Rush, Bart van Merrienboer, Armand Joulin, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In Proceedings of the Fourth International Conference on Learning Representations, volume abs/1502.05698.